

Os centros urbanos, de um modo geral, sofrem com problemas de congestionamento e baixa qualidade do ar. O incentivo ao aumento do uso do transporte coletivo pode auxiliar na resolução desses problemas. O conforto percebido pelo passageiro é um importante aspecto a ser estudado nesse sentido. As soluções encontradas na literatura atualmente não consideram de maneira conjunta os diferentes aspectos que afetam o conforto dos passageiros. Em geral, essas pesquisas se limitam a analisar um único tipo de dado oriundo da telemetria, e essas análises costumam ser de ordem matemática ou observacional. Alternativamente, o trabalho em tela sugere uma abordagem que se beneficia da utilização de técnicas de aprendizagem de máquina. Este trabalho propõe a construção de um modelo de classificação de conforto de passageiros com base na análise de dados de trajetórias de veículos de transporte coletivo. Esse modelo é construído por meio da aplicação de técnicas de aprendizado de máquina sobre dados de sensores instalados no ônibus e dados qualitativos colhidos diretamente dos passageiros. Esses dados são combinados por meio de um modelo de empilhamento de submodelos de classificação. Na maior parte dos testes realizados, o meta modelo tem resultados superiores aos dos modelos que o compõe e o modelo tem robustez para generalização das classificações.

Orientador:

Fabiano Baldo

JOINVILLE, 2019

VANESSA ROLIM
ANO
2019

MODELO DE CLASSIFICAÇÃO DE CONFORTO DE PASSAGEIROS DO
TRANSPORTE PÚBLICO COM BASE EM DADOS DE SENSORES

**UNIVERSIDADE DO ESTADO DE SANTA CATARINA – UDESC
CENTRO DE CIÊNCIAS TECNOLÓGICAS – CCT
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
APLICADA – PPGCA**

**DISSERTAÇÃO DE MESTRADO
MODELO DE CLASSIFICAÇÃO DE CONFORTO DE
PASSAGEIROS DO TRANSPORTE PÚBLICO COM
BASE EM DADOS DE SENSORES**

VANESSA ROLIM

JOINVILLE, 2019

**UNIVERSIDADE DO ESTADO DE SANTA CATARINA - UDESC
CENTRO DE CIÊNCIAS TECNOLÓGICAS - CCT
MESTRADO ACADÊMICO EM COMPUTAÇÃO APLICADA**

VANESSA BARBOSA ROLIM

**MODELO DE CLASSIFICAÇÃO DE CONFORTO DE PASSAGEIROS
DO TRANSPORTE PÚBLICO COM BASE EM DADOS DE
SENSORES**

JOINVILLE

2019

VANESSA BARBOSA ROLIM

**MODELO DE CLASSIFICAÇÃO DE CONFORTO DE PASSAGEIROS
DO TRANSPORTE PÚBLICO COM BASE EM DADOS DE
SENSORES**

Dissertação submetida ao Programa de Pós-Graduação em Computação Aplicada do Centro de Ciências Tecnológicas da Universidade do Estado de Santa Catarina, para a obtenção do grau de Mestre em Computação Aplicada.

Orientador: Dr. Fabiano Baldo

**JOINVILLE
2019**

**Ficha catalográfica elaborada pelo programa de geração automática da
Biblioteca Setorial do CCT/UDESC,
com os dados fornecidos pelo(a) autor(a)**

Rolim, Vanessa

Modelos de classificação de conforto de passageiros do
transporte público com base em dados de sensores /
Vanessa Rolim. -- 2019.
110 p.

Orientador: Fabiano Baldo
Dissertação (mestrado) -- Universidade do Estado de
Santa Catarina, Centro de Ciências Tecnológicas, Programa
de Pós-Graduação em Computação Aplicada, Joinville, 2019.

1. Conforto. 2. Trajetórias. 3. Telemetria. 4. Ônibus. 5.
Aprendizado de máquina. I. Baldo, Fabiano. II. Universidade
do Estado de Santa Catarina, Centro de Ciências
Tecnológicas, Programa de Pós-Graduação em Computação
Aplicada. III. Título.

**Modelo de Classificação de Conforto de Passageiros do Transporte Público
com Base em Dados de Sensores**

por

Vanessa Barbosa Rolim

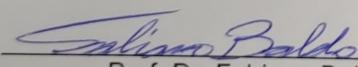
Esta dissertação foi julgada adequada para obtenção do título de

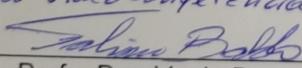
Mestra em Computação Aplicada

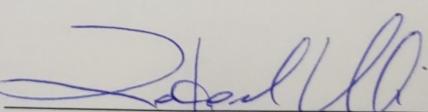
Área de concentração em “Ciência da Computação”,
e aprovada em sua forma final pelo

CURSO DE MESTRADO ACADÊMICO EM COMPUTAÇÃO APLICADA
DO CENTRO DE CIÊNCIAS TECNOLÓGICAS DA
UNIVERSIDADE DO ESTADO DE SANTA CATARINA.

Banca Examinadora:


Prof. Dr. Fabiano Baldo
CCT/UDESC (Orientador/Presidente)

por videoconferência

Profa. Dra. Vania Bogorny
INE/UFSC


Prof. Dr. Rafael Stubs Parpinelli
CCT/UDESC

Joinville, SC, 30 de setembro de 2019.

Ao meu esposo e minha família que nunca mediram esforços e sempre estiveram ao meu lado me dando forças e esperanças para seguir trilhando esse caminho.

AGRADECIMENTOS

Em primeiro lugar agradeço ao meu companheiro, melhor amigo e tremenda inspiração, Rafael, que tentou deixar mais leve cada passo dessa longa e dura jornada, tanto para o desenvolvimento e escrita desse trabalho quanto para todas as batalhas que enfrentamos durante todos esses felizes anos juntos. Sem você eu não teria coragem nem força para a conclusão de mais esse marco.

Agradeço à minha mãe que em todas as oportunidades me lembrou que nada nessa vida veio fácil e nem por isso deixamos de perseguir nossos objetivos. Mesmo à distância sempre me apoiou e foi compreensiva com todas as ausências e ansiedades.

Ao meu pai, meu profundo e sincero agradecimento por ser sempre uma rocha de tranquilidade e serenidade por todas as tribulações que a vida nos colocou. O seu apoio foi tremendamente importante não apenas para a conclusão desse trabalho, como também para todos os desafios que já enfrentei na vida.

Ao professor Fabiano Baldo, meu orientador, mil agradecimentos por todas as broncas muito merecidas, pelas risadas, por toda a paciência e apoio.

Agradeço ao professor e amigo Fernando Braz, que acreditou em mim desde o princípio, muito antes de eu mesma acreditar que poderia sequer iniciar essa jornada e todas as outras que se seguiram.

Agradeço ainda, a todos os colegas de curso com os quais chorei muitas pitangas e compartilhei diferentes perrengues.

Agradeço à Conta Azul pela generosidade com que me tratou durante todo esse período, em particular ao Robson e ao Capelari.

Aproveito, ainda, para agradecer a todos que passaram pela minha vida nos últimos dois anos. Todos e cada um me ajudaram de uma forma diferente, tanto pelas risadas, quanto pelas cervejas e as eventuais lágrimas que pavimentaram esse caminho.

“A noite fria me ensinou a amar mais o meu
dia
E pela dor eu descobri o poder da alegria
E a certeza de que tenho coisas novas
Coisas novas pra dizer”
Belchior

RESUMO

Com base na adoção de tecnologias de acompanhamento de veículos, chamadas de telemetria, por meio de sensores variados, percebe-se a oportunidade para extração de conhecimento na condução e condição de veículos. O meio urbano sofre de muitos problemas do ponto de vista da mobilidade, e diversas pesquisas apontam que o maior uso de veículos de transporte coletivo pode auxiliar na resolução de diversos desses problemas. O conforto percebido pelo passageiro é um importante aspecto a ser estudado para incentivar mais pessoas a utilizarem o transporte coletivo. Existem diversos estudos na área de conforto de passageiros, porém é possível notar lacunas nas soluções propostas. Em particular, essas lacunas estão ligadas à avaliação de diferentes características de uma viagem de ônibus, como essas características afetam umas às outras e como, em conjunto, afetam o conforto do passageiro. Em geral, essas pesquisas se limitam a analisar um único tipo de dado oriundo da telemetria, e essas análises costumam ser de ordem matemática ou observacional. Atualmente, existem diversas técnicas de análise de grandes volumes de dados por meio de algoritmos de aprendizado de máquina que buscam padrões em meio a um conjunto que podem ser generalizados para outros conjuntos de dados. Este trabalho propõe a construção de um modelo de classificação de conforto de passageiros com base na análise de dados de trajetórias de veículos de transporte coletivo. Esse modelo será construído por meio da aplicação de técnicas de aprendizado de máquina sobre dados de sensores instalados no ônibus e dados qualitativos colhidos diretamente dos passageiros. A descoberta dos padrões que são classificados como confortáveis e desconfortáveis pelos passageiros é um problema combinatório que é abordado pela adoção da combinação de um algoritmo de construção de árvores de decisão e um algoritmo de modelo oculto de markov, combinados em um meta modelo que implementa outra árvore de decisão, que é capaz de apresentar resultados aceitáveis em tempo viável. Os experimentos indicam que é possível construir um modelo que classifique o comportamento do motorista com base na percepção de conforto sentida pelos passageiros. Na maior parte dos testes realizados, o meta modelo tem resultados superiores aos dos modelos que o compõe e o modelo tem robustez para generalização das classificações.

Palavras-chaves: conforto, trajetórias, telemetria, ônibus, aprendizado de máquina, veículo de transporte coletivo.

ABSTRACT

Based on the adoption of technologies that are able to collect data from vehicles through different kind of sensors, called telemetry, it is possible to extract knowledge about the driving and the conditions of the vehicle. The urban environment suffers from a range of problems, particularly from a mobility point of view, and several researches suggest that the increase use on collective transport may help on the resolution of several of those problems. The comfort perceived by the passengers is an important aspect to be studied to be applied on the task to encourage a higher number of people to use collective transportation. There are several studies on the passenger comfort area, but it is possible to notice gaps on the proposed solutions. In particular, those gaps are connected to the evaluation of the different characteristics of a bus trip, how those characteristics affect each other and how, together, they affect the perceived passenger comfort. Overall these researches are limited to the analysis of a single feature collected from the telemetry, and this analysis is usually either mathematical or observational. Currently, however, there are all kinds of techniques aimed to analyse large amounts of data through machine learning algorithms, that search for patterns within a data set and try to generalize such pattern to a different set of data. This paper proposes the identification of comfort limits based on the analysis of collective transport vehicle trajectory data. The identification is made by the union of sensor data and qualitative data collected directly from passengers. The discovery of patterns on data that is classified by passengers as comfortable or not comfortable is a combinatory problem, which is approached by the use of a decision tree algorithm combined with a hidden markov model, stacked by an ensemble, which also implements a decision tree, the model is capable of presenting acceptable results on a viable amount of time. The experiments indicate that it is possible to classify the passenger comfort with this approach. In the majority of experiments, the ensemble has superior results compared to the models that constitute it. The ensemble is robust enough to generalize its classifications on the available data.

Key-words: comfort, trajectory, telemetry, bus, collective transport vehicle, machine learning.

LISTA DE ILUSTRAÇÕES

Figura 1 – Acelerações e seus efeitos sobre os seres humanos.	30
Figura 2 – Modelo Oculto de Markov com $N = 3$ representado como um grafo	37
Figura 3 – Modelo de Markov cuja matriz de transição Λ é representada nas arestas do grafo	38
Figura 4 – Exemplo de classificação de dados com HMM.	40
Figura 5 – HMM Univariado, representado por uma matriz de estados.	42
Figura 6 – HMM multivariado representado por uma matriz multimensional de estados.	43
Figura 7 – Equilíbrio entre variância e <i>bias</i>	44
Figura 8 – Exemplo genérico da abordagem de bagging.	45
Figura 9 – Exemplo genérico da abordagem de boosting.	45
Figura 10 – Exemplo genérico do refinamento de resultados com os subconjuntos de dados e seus pesos.	46
Figura 11 – Exemplo genérico de criação de um <i>Stacking</i>	46
Figura 12 – Processo implementado pelo trabalho em tela.	53
Figura 13 – Segmentação de trajetória em subtrajetórias de passageiro, com ilustração de limites adicionais para garantia de consideração de dado completo.	57
Figura 14 – Representação de sobreposição de subtrajetórias de passageiros.	58
Figura 15 – Anotação de percepção de passageiros nas trajetórias com base na distribuição de acelerações em cada eixo, classificadas por cores.	60
Figura 16 – Representação de metamodelo de classificação.	63
Figura 17 – Distribuição das respostas dos usuários com relação à percepção de conforto.	67
Figura 18 – Comparação de uma trajetória antes e após a retirada de <i>staypoints</i> . A Figura (a) representa uma trajetória completa, com todos os dados coletados. A Figura (b) representa a mesma trajetória após a retirada dos <i>staypoints</i>	68
Figura 19 – Trajetória de um veículo após a retirada de <i>staypoints</i> , mantendo os pontos de aceleração brusca comumente encontrados em terminais de ônibus.	69
Figura 20 – Curvas ROCs do primeiro cenário de testes.	82
Figura 21 – Curvas ROCs do segundo cenário de testes.	86
Figura 22 – Curvas ROCs do terceiro cenário de testes.	90
Figura 23 – Curvas ROCs do quarto cenário de testes.	94

LISTA DE TABELAS

Tabela 1 – Limites de provável percepção de conforto de passageiros, como indicados pela ISO 26311:1997.	31
Tabela 2 – Matriz de probabilidade Λ	38
Tabela 3 – Resumo dos trabalhos relacionados	52
Tabela 4 – Dados disponibilizados pelo dispositivo instalado no veículo.	55
Tabela 5 – Classes de estratificação das notas de percepção de conforto comunicadas pelos passageiros durante a aplicação do questionário.	66
Tabela 6 – Distribuição da anotação percepção de conforto geral por classes.	71
Tabela 7 – Atributos selecionados para o método de classificação proposto.	73
Tabela 8 – Matriz de confusão da árvore de decisão do eixo x.	74
Tabela 9 – Matriz de confusão da árvore de decisão do eixo y.	74
Tabela 10 – Matriz de confusão da árvore de decisão do eixo z.	75
Tabela 11 – Matriz de confusão do HMM do eixo x.	76
Tabela 12 – Matriz de confusão do HMM do eixo y.	77
Tabela 13 – Matriz de confusão do HMM do eixo z.	78
Tabela 14 – Matriz de confusão do meta modelo.	78
Tabela 15 – Resumo dos resultados de treinamento.	79
Tabela 16 – Matriz de confusão da árvore de decisão do eixo x do primeiro experimento	82
Tabela 17 – Matriz de confusão do HMM do eixo x do primeiro experimento	83
Tabela 18 – Matriz de confusão da árvore de decisão do eixo y do primeiro experimento	83
Tabela 19 – Matriz de confusão do HMM do eixo y do primeiro experimento	84
Tabela 20 – Matriz de confusão da árvore de decisão do eixo z do primeiro experimento.	84
Tabela 21 – Matriz de confusão do HMM do eixo z do primeiro experimento.	84
Tabela 22 – Matriz de confusão do metamodelo do primeiro experimento.	85
Tabela 23 – Anotações das trajetórias do cenário dois relação aos eixos de aceleração e conforto geral.	86
Tabela 24 – Matriz de confusão da árvore de decisão do eixo x do segundo experimento.	86
Tabela 25 – Matriz de confusão do HMM do eixo x do segundo experimento.	87
Tabela 26 – Matriz de confusão da árvore de decisão do eixo y do segundo experimento.	87
Tabela 27 – Matriz de confusão do HMM do eixo y do segundo experimento.	87

Tabela 28 – Matriz de confusão da árvore de decisão do eixo z do segundo experimento.	88
Tabela 29 – Matriz de confusão do HMM do eixo z do segundo experimento.	88
Tabela 30 – Matriz de confusão do meta modelo do segundo experimento.	88
Tabela 31 – Anotações das trajetórias do cenário três com relação aos eixos de aceleração e conforto geral.	90
Tabela 32 – Matriz de confusão da árvore de decisão o eixo x do terceiro cenário.	90
Tabela 33 – Matriz de confusão do HMM do eixo x do terceiro experimento.	91
Tabela 34 – Matriz de confusão da árvore de decisão o eixo y do terceiro cenário.	91
Tabela 35 – Matriz de confusão do HMM do eixo y do terceiro experimento.	91
Tabela 36 – Matriz de confusão da árvore de decisão o eixo z do terceiro cenário.	92
Tabela 37 – Matriz de confusão do HMM do eixo z do terceiro experimento.	92
Tabela 38 – Matriz de confusão do meta modelo do terceiro cenário.	93
Tabela 39 – Anotações das trajetórias do cenário quatro com relação aos eixos de aceleração e conforto geral.	94
Tabela 40 – Matriz de confusão da árvore de decisão o eixo x do quarto cenário.	95
Tabela 41 – Matriz de confusão do HMM do eixo x do quarto experimento.	95
Tabela 42 – Matriz de confusão da árvore de decisão o eixo y do quarto cenário.	95
Tabela 43 – Matriz de confusão do HMM do eixo y do quarto experimento.	96
Tabela 44 – Matriz de confusão da árvore de decisão o eixo z do quarto cenário.	96
Tabela 45 – Matriz de confusão do HMM do eixo z do quarto experimento.	96
Tabela 46 – Matriz de confusão do meta modelo do quarto cenário.	97

LISTA DE ABREVIATURAS E SIGLAS

ISO	<i>International Organization for Standardization</i>
GPS	<i>Global Positioning System</i>
OBD	<i>On-Board Diagnostic</i>
IMU	<i>Inertial Measurement Unit</i>

SUMÁRIO

1	INTRODUÇÃO	16
1.1	OBJETIVO	19
1.1.1	Objetivos Específicos	19
1.2	METODOLOGIA DE PESQUISA	19
1.2.1	Caracterização Metodológica	20
1.2.2	Procedimento Metodológico	20
1.3	Estrutura do Trabalho	21
2	REFERENCIAL TEÓRICO	22
2.1	Trajetórias de Objetos Móveis	22
2.1.1	Definições Gerais	22
2.1.2	Definições Específicas	25
2.2	Conforto e Conceitos Correlatos	26
2.2.1	Percepção de Conforto	28
2.2.2	Padrões de Conforto	29
2.3	Método de Classificação	31
2.3.1	Tipos de classificação de dados	31
2.3.2	Árvores de Decisão	32
2.3.2.1	Poda de Árvores de Decisão	34
2.3.2.2	Algoritmo C4.5	35
2.3.3	Modelos Ocultos de Markov	35
2.3.3.1	Classificação de dados	39
2.3.3.1.1	<i>Problema da Avaliação</i>	40
2.3.3.1.2	<i>Problema da Decodificação</i>	41
2.3.3.1.3	<i>Problema do Aprendizado</i>	41
2.3.3.2	HMM univariado e HMM multivariado	42
2.3.4	Meta Modelo de Classificação	43
2.3.4.1	Bagging	44
2.3.4.2	Boosting	44
2.3.4.3	Stacking	46
2.4	Trabalhos Relacionados	46
2.4.1	Considerações sobre os Trabalhos Relacionados	50
3	MODELO PROPOSTO	53
3.1	Dados de trajetórias	54

3.2	Questionário	54
3.3	Pré-processamento	54
3.4	Criação do Conjunto de Treinamento	56
3.4.1	Segmentação da Trajetória do Ônibus com Base na Viagem do Passageiro	57
3.4.2	Consenso entre Percepções de Conforto em cada Período da Viagem	58
3.4.3	Anotação da Base com a Percepção de Conforto do Passageiro	58
3.5	Treinamento dos modelos	60
3.5.1	Treinamento das árvores de decisão	60
3.5.2	Treinamento dos Modelos Ocultos de Markov - HMM	61
3.5.3	Treinamento do metamodelo	62
4	IMPLEMENTAÇÃO	64
4.1	Coleta de Dados de Sensores	64
4.2	Questionário	65
4.3	Pré-processamento	66
4.3.1	Retirada de <i>staypoints</i>	66
4.3.2	Consenso de Percepções	69
4.4	Criação de base de conhecimento	70
4.5	Seleção de subconjuntos	71
4.6	Treinamento das Árvores de Decisão	73
4.6.1	Árvore de decisão eixo x - acelerações e frenagens	74
4.6.2	Árvore de decisão eixo y - curvas	74
4.6.3	Árvore de decisão eixo z - trepidação	75
4.7	Treinamento dos Modelos Ocultos de Markov	75
4.7.1	HMM eixo x - acelerações e frenagens	76
4.7.2	HMM eixo y - curvas	77
4.7.3	HMM eixo z - trepidação	77
4.8	Meta modelo de classificação	78
4.9	Resumo dos resultados de treinamento	79
5	ANÁLISE DOS RESULTADOS	80
5.1	Protocolo de experimentação	80
5.2	Cenários	81
5.2.1	Cenário 1 - Conforto relacionado com o eixo x	81
5.2.1.1	Eixo x	82
5.2.1.2	Eixo y	83
5.2.1.3	Eixo z	83

5.2.1.4	Metamodelo	84
5.2.1.5	Considerações sobre o primeiro cenário	85
5.2.2	Cenário 2 - Conforto relacionado com o eixo y	85
5.2.2.1	Eixo x	86
5.2.2.2	Eixo y	87
5.2.2.3	Eixo z	87
5.2.2.4	Metamodelo	88
5.2.2.5	Considerações sobre o segundo cenário	89
5.2.3	Cenário 3 - Conforto relacionado com o eixo z	89
5.2.3.1	Eixo x	90
5.2.3.2	Eixo y	91
5.2.3.3	Eixo z	92
5.2.3.4	Metamodelo	92
5.2.3.5	Considerações sobre o terceiro cenário	93
5.2.4	Cenário 4 - Conforto geral	94
5.2.4.1	Eixo x	94
5.2.4.2	Eixo y	95
5.2.4.3	Eixo z	95
5.2.4.4	Metamodelo	96
5.2.4.5	Considerações sobre o quarto cenário	97
5.3	Considerações sobre os resultados	97
6	CONCLUSÕES	99
6.1	Trabalhos Futuros	101
	REFERÊNCIAS	103

1 INTRODUÇÃO

Um dos desafios atuais nas grandes cidades é a mobilidade urbana. A mobilidade urbana se refere à capacidade de deslocamento de pessoas e cargas em áreas com grande adensamento populacional, comum em cidades de médio e grande porte. Os desafios enfrentados pela população dessas áreas vão desde os relacionados ao tempo gasto em congestionamentos, até a qualidade de vida afetada pela poluição do ar e estresse do trânsito (SCHRANK; EISELE; LOMAX, 2012).

No Brasil, 0,05% da frota de veículos é composta por caminhões, 0,006% por veículos de transporte coletivo e 87% por veículos privados (CNT, 2017). A abundância de veículos privados é a grande responsável pelos congestionamentos diários e pelo declínio na qualidade do ar nas cidades brasileiras (COHN, 2014). Essa situação se agrava pela constatação de que grande parte dos veículos privados são utilizados para o transporte individual de passageiros.

Algumas cidades encontraram paliativos aos problemas de mobilidade com investimentos em políticas de transporte multimodal e alternativo, como é o caso do estímulo do uso da bicicleta e o melhoramento da integração entre trens, metrôs e ônibus (TENNØY, 2010). Além disso, outras soluções focam na melhoria da qualidade do transporte coletivo, como o uso de veículos novos, com assentos confortáveis, ar condicionado e conexão à Internet. Essas iniciativas são importantes e trazem benefícios para o uso do transporte coletivo, entretanto, não são suficientes para estimular grande parte da população a substituir o transporte individual pelo coletivo. Portanto, é necessário que outras ações sejam desenvolvidas de forma a estimular o uso do transporte coletivo e assim contribuir para a melhoria da qualidade de vida nas grandes cidades (LE; TRINH, 2016).

Segundo a Secretaria de Planejamento Urbano e Desenvolvimento Sustentável (SEPUD) de Joinville, 20% dos moradores da cidade utilizam transporte coletivo diariamente. Entretanto, esse índice diminuiu 16% em 15 anos (SEPUD, 2016). Além disso, outra pesquisa, realizada pelo instituto Rio Como Vamos, sugere que o aumento do conforto de passageiros pode ser um fator determinante para que mais pessoas utilizem o transporte coletivo (VAMOS, 2015). Devido ao impacto positivo que o conforto pode causar na utilização do transporte coletivo, vários trabalhos abordaram a avaliação do conforto no transporte coletivo nos últimos anos. Entretanto, apesar das contribuições apresentadas pelos trabalhos de Rotmans, Kemp e Asselt (2001), Simpson e Simpson (2003), Hassan, Hawas e Ahmed (2013), Innamaa e Penttinen (2013), Eboli, Mazzulla e Pungillo (2016) e Park et al. (2017), a análise do conforto do transporte

coletivo ainda é uma questão em aberto, pois envolve a generalização de aspectos que são afetados pelos diferentes pontos de vista dos passageiros. Portanto, essa é considerada uma tarefa não trivial.

Segundo Fellesson e Friman (2012), os principais aspectos a serem avaliados com relação ao conforto são: i) estado dos bancos, limpeza das estações e dos veículos; ii) sensação de segurança e; iii) conforto na condução dos veículos. Aspectos como o estado dos bancos, lotação do ônibus e limpeza em geral são de responsabilidade das empresas concessionárias do serviço e são avaliados pelos órgãos de fiscalização, sendo afetados, normalmente, por escolhas gerenciais associadas à economia e à margem de lucro das empresas.

A sensação de segurança, de um modo geral, está relacionada à percepção por parte do passageiro do risco de um assalto, por exemplo. Essa característica também pode compor o conforto do passageiro, mas está correlacionada em maior grau com o estado de segurança percebido por ele. Ainda, a sensação de segurança pode estar ligada à percepção de integridade física do passageiro, sendo percebida tanto por meio da condição mecânica do veículo, como, por exemplo, os sons de peças batendo, motor se comportando de forma errática, suspensão danificada, quanto pela condução do motorista, em situações como frenagens bruscas ou acelerações acentuadas. Dessa forma, é possível que alguns aspectos da sensação de segurança se sobreponham a condução do veículo, onde essa última está associada a frenagens e acelerações bruscas, curvas desnecessariamente acentuadas ou não frenagem em redutores de velocidade (lombadas).

Ainda que as características de sensação de segurança possam ser colhidas por meio da percepção dos passageiros, essas características físicas e cinéticas podem ser melhor aferidas por meio de sensores instalados no veículo. É possível coletar dados de sensores dos mais variados tipos, desde temperatura no interior e exterior do veículo, vibração do chassis e até pressão que o motorista exerce sobre o pedal de aceleração ou freio. Um dado comumente coletado é a localização espaço-temporal do veículo, por meio de um receptor do Sistema de Posicionamento Global (GPS) (KERR; DUNCAN; SCHIPPERJIN, 2011). Esses dados, quando agrupados de forma cronológica, podem representar a movimentação de um objeto, também chamada de trajetória.

As trajetórias enriquecidas com dados de outros sensores, como acelerômetro, giroscópio e telemetria do barramento de dados do veículo, assim como com a percepção do passageiro, podem ser analisadas em busca de padrões de condução. Esses padrões podem auxiliar na identificação dos comportamentos dos motoristas que afetam o conforto do passageiro. Entretanto, a junção de dados de sensores e percepção de passageiros pode ser uma tarefa desafiadora.

A literatura apresenta diferentes técnicas, sejam elas matemáticas, estatísticas, algorítmicas ou de simples observação, para a classificação de conforto de passageiros em transporte coletivo. No entanto, essas técnicas costumam analisar os dados coletados pelos sensores de forma separada, sem considerar os efeitos que diferentes características de condução possam ter umas nas outras, do ponto de vista de conforto.

As dificuldades na análise de conforto começam pelo fato de que a própria definição de conforto ainda é um tema em discussão no meio acadêmico. Por exemplo, existem definições que caracterizam o conforto com o estado positivo de percepção do passageiro, enquanto outras definem o conforto como a ausência de desconforto. Ainda, em alguns casos, o conforto é medido com base em graus de desconforto e, em outros, o conforto só é atingido após a eliminação completa do desconforto.

Para alguns pesquisadores, o conforto é apenas um aspecto secundário utilizado para compor uma métrica de qualidade de serviço (PALMA; KILANI; PROOST, 2015)(SHAM; SAMSUDIN; RAHMAN, 2013)(PIMENTA; ASSUNÇÃO, 2015). Nesses casos, o conforto não é percebido como uma necessidade. Para outros pesquisadores, o conforto é métrica preponderante para o sucesso ou fracasso de um projeto (CASCIOLI; HEUSCH; MCCARTHY, 2011)(ERIKSSON; FRIBERG, 2000)(ZEEMAN; BOOYSEN, 2013). No entanto, todos esses pesquisadores concordam que o conforto por si só é imensurável, pois é uma percepção pessoal de cada ser humano, e essa percepção é afetada pelos mais diversos aspectos físicos e culturais.

Posto isso, este trabalho pretende tratar conforto de passageiros de transporte coletivo por meio da mensuração do grau de desconforto causados pelos diversos aspectos cinéticos envolvidos em uma viagem de ônibus. Esses aspectos cinéticos são obtidos de sensores de GPS, aceleração e de telemetria do barramento do veículo. O trabalho em tela admite que existam outros prismas para esse mesmo problema, mas entende que essa visão é consistente e suficiente para quantificar o conforto dos passageiros do transporte coletivo.

Portanto, a pergunta de pesquisa que se pretende responder é: Como mensurar o grau de conforto dos passageiros do transporte coletivo por meio da análise de trajetórias do veículo enriquecidas com dados de sensores coletados durante a viagem?

É possível realizar essa mensuração com a opinião de especialistas, por exemplo, ou, ainda, através da análise de elementos cinéticos particulares. Para esse trabalho, utiliza-se um conjunto de técnicas de aprendizado de máquina, de forma que o viés proveniente de análises esparsas seja diminuído e que diversos elementos possam ser considerados em cada instante.

Para tanto, o trabalho em tela se utiliza de árvores de decisão, que são modelos de aprendizado de máquina de uso geral consolidados no contexto da computação. São utilizados, ainda modelos ocultos de Markov que são modelos estatísticos que, com base em observações do comportamento dos dados, busca a fonte geradora oculta ao observador.

O uso combinado de diferentes modelos de classificação é importante para que a classificação dos dados seja, de maneira geral, mais acurada e tenha resultados mais robustos. Além disso, os dados são classificados com base nos eixos de aceleração e é necessário combiná-los para a classificação com base no conforto geral. Finalmente, um modelo oculto de Marko tem a característica de descoberta de influências não óbvias entre as diferentes variáveis de um mesmo modelo, dessa maneira, a associação de um modelo oculto de Markov com árvores de decisão é útil para a descoberta de características desconsideradas pelas árvores de decisão.

1.1 OBJETIVO

Este trabalho tem por objetivo construir um modelo heterogêneo de classificação de conforto de passageiros de transporte coletivo utilizando dados de trajetórias dos ônibus combinados com dados de sensores de aceleração e de telemetria do barramento do veículo, classificados pela própria percepção de conforto de uma amostra de passageiros.

1.1.1 Objetivos Específicos

Os objetivos específicos definidos para alcançar o objetivo geral deste trabalho são:

- Definir uma estratégia para filtrar os dados das trajetórias de forma a manter apenas os dados relevantes ao contexto em estudo;
- Realizar análises exploratórias dos dados de forma a apoiar a elaboração da estratégia de classificação de conforto;
- Propor um modelo de classificação de conforto de passageiros de transporte coletivo com base na aplicação de técnicas de aprendizado de máquina.

1.2 METODOLOGIA DE PESQUISA

Esta seção apresenta a caracterização metodológica da pesquisa, assim como o detalhamento do procedimento metodológico utilizado no trabalho em tela.

1.2.1 Caracterização Metodológica

De acordo com a classificação metodológica encontrada na literatura, este trabalho pode ser categorizado com relação ao tipo de pesquisa como uma apresentação de algo presumidamente melhor, pois propõe uma abordagem diferente para solucionar um problema já explorado em outros trabalhos, visando ganhos de desempenho e precisão (MARCONI; LAKATOS, 2003). Do ponto de vista do paradigma da pesquisa científica, pode ser enquadrado como tecnocatra, pois utiliza a avaliação de experimentos empíricos para compreender as contribuições da solução (EDEN, 2007), ao contrário das abordagens racionalistas e naturalistas, que buscam conhecimento *a priori* por meio de deduções de raciocínio ou por dedução formal e experimentação científica rigorosa, respectivamente. Quanto ao procedimento técnico adotado, este trabalho se enquadra como uma Pesquisa-Ação, ou seja, é realizado em ciclos de investigação-ação que utilizam técnicas científicas para encontrar resultados práticos que, por sua vez, são utilizados para aprimorar os meios de pesquisa a fim de encontrar melhores resultados práticos (TRIPP, 2005). Além disso, utiliza a avaliação indutiva, ou seja, que busca por padrões em casos particulares que explicam ou se aplicam a todos os casos análogos aos analisados (MORIN; MOIGNE, 2000). Por fim, a pesquisa é considerada de nível de maturidade 2 (WAZLAWICK, 2017), pois propõe um método diferente e busca conhecimento empírico, *a posteriori*, para validação de proposta.

1.2.2 Procedimento Metodológico

O procedimento de pesquisa adotado inicia com a revisão bibliográfica dos conceitos básicos sobre trajetórias, e se estende para as definições de conceitos específicos relacionados a trajetórias. A seguir, os conceitos relacionados ao conforto e percepção são revisados. Ainda durante a etapa de revisão bibliográfica, são revisados os conceitos de algoritmos de aprendizado de máquina, em particular os de classificação de dados baseados em Árvores de Decisão e Modelos Ocultos de Markov. Como última etapa da revisão, são revisados e comparados os trabalhos relacionados encontrados na literatura.

Em paralelo à etapa de revisão, são realizadas a aplicação de um questionário aos passageiros do ônibus que contém o dispositivo de coleta de dados e também a análise exploratória sobre a massa de dados coletada desse veículo a fim de identificar dados relevantes para a caracterização do problema e, consequentemente, aplicação da solução proposta. A base de dados consiste em arquivos de trajetórias de um veículo de transporte coletivo monitorado, enriquecidas com dados de sensores de aceleração e de telemetria do barramento de dados do veículo. Com base na análise exploratória, são propostas etapas de limpeza, filtragem e seleção dos dados

relevantes para compor o conjunto de treinamento para a criação do modelo de classificação por meio da utilização de técnicas de aprendizado de máquina supervisionada, a saber, árvores de decisão e modelos ocultos de Markov.

Na sequência, usando como base uma abordagem evolucionária baseada no procedimento pesquisa-ação, o modelo de classificação concebido é devidamente adaptado para aprimorar a solução do problema em questão. De acordo com o procedimento de Pesquisa-Ação, os resultados dos experimentos guiarão a evolução iterativa e incremental do modelo, resultando, ao final de um conjunto finito de iterações, no modelo suficientemente adequado para a solução do problema abordado. Por fim, os resultados são avaliados por meio de testes em conjuntos de dados previamente classificados a fim de mensurar a acurácia do modelo criado.

1.3 ESTRUTURA DO TRABALHO

Este trabalho está estruturado da seguinte forma. O capítulo 2 apresenta a fundamentação teórica desta pesquisa. Serão apresentadas as definições básicas e específicas sobre trajetórias. Além disso, são revisados os conceitos de conforto e os padrões internacionais de conforto. Ainda, apresentada uma revisão sobre métodos de classificação do tipo Árvore de Decisão e Modelo Oculto de Markov, finalizando com a apresentação e comparação dos trabalhos relacionados. O capítulo 3 apresenta o método proposto, iniciando pela coleta dos dados, passando pelo preprocessamento de dados, combinação dos dados de telemetria e percepção dos passageiros, finalizando com o treinamento do modelo de classificação. O capítulo 4 apresenta a implementação do modelo proposto, detalhando cada etapa, descrevendo algoritmos e parametrizações. O Capítulo 5 apresenta a análise dos resultados da aplicação do modelo em cenários variados escolhidos *a priori*. Por fim, o Capítulo 6 apresenta as conclusões e trabalhos futuros.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta os principais conceitos relacionados ao tema do trabalho em tela. A seção 2.1 apresenta a definição de trajetória e os conceitos relacionados ao movimento de veículos, particularmente, de transporte coletivo. Já a seção 2.2 detalha os principais desafios na caracterização de conforto, apresentando definições e padrões internacionais para a medição e análise de conforto, bem como suas limitações. Na sequência, a seção 2.3 descreve os métodos de aprendizado de máquina supervisionado para treinamento de Árvores de Decisão e Modelos Ocultos de Markov. Finalmente, a seção 2.4 faz uma abrangente revisão dos trabalhos relacionados presentes na literatura.

2.1 TRAJETÓRIAS DE OBJETOS MÓVEIS

Esta seção apresenta os principais conceitos relacionados às trajetórias de objetos móveis. Esses conceitos são essenciais para a compreensão do método proposto pelo trabalho em tela.

2.1.1 Definições Gerais

O registro de movimento é a representação discreta da movimentação contínua de um objeto. Nessa representação os registros são feitos a cada intervalo de tempo definido pelo mecanismo de coleta. Quanto menor for o intervalo de coleta maior será a percepção de continuidade do movimento do objeto.

Definição 1. *Registro de Movimento* Para fins de manipulação computacional, os movimentos de um objeto móvel são armazenados de forma discreta em uma lista (PARENT et al., 2013). Ou seja, em cada momento i , um ponto P é registrado por meio das informações de posicionamento espacial c e tempo t .

$$RM = \{P_1, P_2, \dots, P_n\} \rightarrow P_i = (c_i, t_i) \quad (2.1)$$

As coordenadas podem ser espaciais ou geográficas. As coordenadas espaciais são tuplas de pontos que identificam o posicionamento do objeto em um espaço qualquer, como o plano cartesiano. No caso do plano cartesiano, uma coordenada composta por dois atributos (x, y) é o suficiente para a identificação da posição do objeto. As coordenadas geográficas são uma especialização das coordenadas espaciais, pois representam a posição do objeto no planeta. Essas coordenadas são compostas

de pelo menos dois atributos, a latitude e a longitude. É possível, ainda, que contenham outro atributo, a altitude do objeto com relação ao nível do mar.

Definição 2. Coordenada *Uma coordenada c é a identificação do posicionamento espacial de um objeto por meio de um conjunto de dimensões representado por $c = (x, y, [z])$, onde x e y representam o posicionamento bidimensional do objeto no espaço e z é um parâmetro opcional que representa a altura do objeto no espaço (BOGORNY; BRAZ, 2012).*

Um ponto P é composto por uma coordenada e o momento em que ela ocorreu. Um conjunto de pontos representa o movimento de um objeto. Cada ponto P pode, ainda, informar atributos adicionais sobre o movimento de um objeto. Esses atributos, no contexto da análise de trajetórias de veículos, podem representar, por exemplo, a velocidade do veículo no momento da coleta, a direção do veículo, a temperatura interna, ou qualquer outro dado relevante e que possa ser coletado por sensores interligados com o sistema que colete as coordenadas espaciais.

Definição 3. Ponto *Um ponto P é a identificação do movimento de um objeto durante um momento t , tal que $P = (c, t, S_VA)$, onde c é a coordenada do objeto, t o instante de tempo, e S_VA se refere aos aspectos que enriquecem o ponto, como cada ponto pode ter mais de um aspecto associado a si, como velocidade e aceleração, eles são considerados pontos multi aspecto (MELLO et al., 2019).*

Definição 4. Trajetória *Uma trajetória $T = (p_1, p_2, \dots, p_n)$ é um conjunto de pontos temporalmente ordenados em que p_1 e p_n representam o início e o fim da trajetória, respectivamente. Uma trajetória é delimitada pelo objetivo de análise da movimentação do objeto pretendida (GIANNOTTI et al., 2007). As trajetórias desse trabalho são compostas por pontos multi aspecto, portanto as trajetórias são também consideradas multi aspecto (MELLO et al., 2019).*

Uma trajetória pode ser segmentada de forma a viabilizar a implementação de processos computacionais e para a aplicação de regras definidas pela análise realizada. No contexto do estudo de trajetórias de veículos de transporte coletivo, a segmentação de uma trajetória pode ocorrer sob diversos critérios como, por exemplo, em paradas em pontos de embarque e desembarque de passageiros, em terminais de linhas ou em semáforos.

Definição 5. Subtrajetória *Uma subtrajetória S é um conjunto de pontos temporalmente ordenados $S = \{p_i, p_{i+1}, \dots, p_j\}$, em que $i \geq 0$ e $j \leq n$, e $S \subset T$. O início e o fim da subtrajetória são definidos de acordo com o estudo realizado.*

A velocidade é uma das principais medidas utilizadas pelas entidades responsáveis pelo trânsito nas cidades, sendo controlada em todo o território nacional. Esse controle ocorre pois o abuso dos limites de velocidade são creditados como principal

fator de mortalidade e lesão grave no trânsito (MARÍN; QUEIROZ, 2000).

Definição 5. Velocidade *Medida física calculada por meio da divisão da distância (d) percorrida no tempo (t) necessário para percorrê-la.*

$$v = \frac{d}{t} \quad (2.2)$$

A aceleração está relacionada com o aumento ou diminuição da velocidade durante um determinado período. O cálculo de aceleração pode resultar em zero, em situações em que durante o período de avaliação não ocorreu variação da velocidade. Além disso, quando a aceleração tem baixa variação, considera-se que o veículo está em velocidade de cruzeiro.

Definição 6. Aceleração *Medida física que calcula a variação da velocidade em função do tempo (MALVERN, 1969). Portanto, o cálculo implica na existência de pelo menos dois registros de velocidade e o tempo entre um registro e o outro.*

$$a = \frac{\Delta v}{\Delta t} \quad (2.3)$$

A aceleração própria pode ser capturada por meio de um acelerômetro. Um acelerômetro em repouso sempre indicará que a aceleração vertical é de 1g. O princípio de medição de aceleração vertical pode ser expandido para qualquer eixo que se pretenda medir. Dessa forma, é possível medir a aceleração que um ser humano sofre quando o veículo em que viaja passa por um desnível na pista, fazendo com o que o acelerômetro indique que a velocidade vertical aumentou ou diminuiu.

Definição 7. Aceleração Própria *Medida calculada em relação a outro sistema em queda livre. Quando em repouso na Terra, um objeto estará constantemente sob o efeito de 1g, que é a aceleração exercida pela gravidade (WALD, 2010). Esse conceito é derivado do Princípio da Equivalência de Einstein, que postula que a aceleração da gravidade é indistinguível de uma aceleração vertical (HARTLE, 2003).*

Em um veículo os eixos medidos pela aceleração triaxial se relacionam com a direção em que ele se movimenta. A aceleração horizontal está relacionada ao percurso percorrido pelo veículo sem realizar curvas, ou seja, quando veículo segue em frente ou em marcha ré. A aceleração lateral está relacionada às curvas que um veículo realiza durante seu percurso. Já a aceleração vertical está relacionada com a trepidação realizada pelo veículo durante seu percurso.

Definição 8. Aceleração Triaxial *A aceleração triaxial é um conjunto composto por três acelerações próprias: horizontal, lateral e vertical. Essas acelerações estão relacionadas a direção em que um objeto se desloca e as modificações da velocidade que ocorrem durante o percurso.*

A aceleração triaxial pode ser obtida por um acelerômetro. O acelerômetro é um sensor que pode ser acoplado a um dispositivo embarcado para armazenar as acelerações coletadas. Assim como outros sensores de aquisição de dados, ele possui algumas características, tais como a sensibilidade, que é a propriedade relacionada com a precisão do resultado apresentado pelo sensor.

Definição 9. Sensibilidade de sensor A sensibilidade é definida pela diferença entre o valor máximo e mínimo mensurável pelo sensor em um intervalo de tempo. Quanto maior for essa diferença e quanto menor for esse intervalo, maior será a sensibilidade do sensor. Sensores pouco sensíveis não conseguem detectar toda a gama de entradas possíveis, enquanto que sensores muito sensíveis podem detectar praticamente todas as possíveis variações da entrada em um instante reduzido de tempo (BOUTEN et al., 1997).

2.1.2 Definições Específicas

Essa seção apresenta a definição dos conceitos específicos de forma a facilitar o entendimento do contexto e dos métodos apresentados no trabalho em tela. A nomenclatura e as definições a seguir têm como finalidade facilitar o entendimento do trabalho e, portanto, não devem ser consideradas definições formais dos termos.

Definição 10. Passageiro Um passageiro de transporte coletivo é um cidadão pagante ou não que utiliza veículos de transporte coletivo. Os passageiros podem viajar nesses veículos sentados ou em pé. Todos as pessoas que estão em um ônibus fornecendo serviços para a empresa que gerencia a linha ônibus, como motoristas e fiscais, não são considerados passageiros.

Definição 11. Linha de ônibus Uma linha de ônibus é um percurso pré-definido que deve ser seguido por todos os veículos que a ela foram atribuídos. Uma linha de ônibus é, ainda, um percurso divulgado e conhecido entre passageiros, de forma que um grupo de passageiros possa se deslocar de coordenadas diferentes para uma mesma, permitindo que todos embarquem em um mesmo veículo que segue para um destino comum a eles.

Definição 12. Ponto de ônibus/Parada de ônibus Um ponto de ônibus ou parada de ônibus é um local pré-definido em que o veículo deve efetuar uma parada completa para embarque e/ou desembarque de passageiros quando sinalizado. Em uma linha pode existir um conjunto de paradas de ônibus, ou, em caso de linhas diretas, não existir nenhuma parada além da primeira e da última. Esses locais são marcados através de sinalização de trânsito apropriada, permitindo que motoristas e passageiros consigam identificá-los facilmente.

Definição 13. Terminal de ônibus Um terminal de ônibus é um local pré-

determinado onde os ônibus iniciam e finalizam suas linhas. Esse local pode oferecer infraestrutura para motoristas e passageiros, como banheiros, lanchonetes, quadros de horários das linhas e venda de passagens. No entanto, também podem ser locais sem qualquer infraestrutura onde a linha apenas termina sua circulação com passageiros.

Definição 14. Arrancada *Uma arrancada é um movimento rápido, brusco e repentino realizado pelo veículo em momentos de aceleração acentuada. É comum que ocorram em situações de início de deslocamento, tais como na saída de um terminal ou ponto de ônibus, ou em horários de pico, onde ocorrem congestionamentos.*

Definição 15. Freada *Uma freada é um movimento rápido, brusco e repentino realizado pelo veículo em momentos de desaceleração exacerbada. É comum que ocorram quando o veículo chega a um terminal ou ponto de ônibus, ou em situações em que o condutor do veículo tenta evitar acidentes de trânsito.*

2.2 CONFORTO E CONCEITOS CORRELATOS

Conforto é um conceito estudado há muitos anos e que tem diversos significados. Esses significados variam entre as áreas da ciência mas, de modo geral, alguns pontos básicos estão acordados entre pesquisadores, tais como: conforto é um estado físico ou psicológico positivo constantemente em avaliação e comparação com estados menos positivos (PINEAU, 1982). É importante ressaltar que o conforto pode ser físico, psicológico ou uma união dos dois. Estudos que consideram apenas um desses aspectos costumam reconhecer que os dois aspectos são inseparáveis e são afetados um pelo outro. Conforto pode significar o alívio do desconforto, ou seja, a eliminação de elementos e fatores que causem dor, tensão, insegurança ou angustia (KOLCABA; KOLCABA, 1991). Ainda, pode significar um estado mental de paz e de contentamento.

Do ponto de vista da medicina, conforto está intrinsecamente conectado à dor e ao sofrimento. Esta linha de argumentação considera conforto como uma graduação de estados, sendo eles: “conforto”, “desconforto não percebido”, “desconforto percebido”, “fadiga” e “dor” (KÖLSCH; BEALL; TURK, 2003). Nessa classificação, o pior estado de conforto é a dor e o conforto só é alcançado quando nenhum elemento de desconforto é identificado, mesmo que esse desconforto não seja percebido de forma consciente.

Para a ergonomia, o conforto está relacionado com a facilidade e contentamento com o ambiente ou produto que facilita o desempenho (ORTIZ; KURVERS; BLUYSEN, 2017). Nesse ramo da ciência, a maior base de pesquisa em conforto está concentrada em conforto de sujeitos sentados.

O estudo de conforto de passageiros do transporte coletivo abrange tanto o ponto de vista médico quanto ergonômico. De forma a facilitar o entendimento dos trabalhos relacionados, eles foram classificados de acordo com três pontos de vista:

- Conforto como parte do conceitos de qualidade de serviço, como iluminação, limpeza e pontualidade;
- Conforto como uma medida de possível dano físico ao passageiro, como a distensão de um músculo causada por uma guinada repentina;
- Conforto como resultado do efeito de fatores cinéticos, como velocidade ou aceleração.

Os trabalhos que estudam o aspecto de qualidade de serviço em transporte coletivo estão interessados em elementos do conforto como a lotação do veículo (PALMA; KILANI; PROOST, 2015) (ESSADEQ; DUBAIL; JEANNIERE, 2016) (TIRACHINI et al., 2017), limpeza e segurança pessoal, percepção de segurança contra assalto ou agressão (SHAM; SAMSUDIN; RAHMAN, 2013) (CHOWDHURY, 2016) (MONZÓN; HERNÁNDEZ; Di Ciommo, 2016) (SOHONI; THOMAS; RAO, 2017) (IMRE; ÇELEBI, 2017), pontualidade e otimização de rotas (De Oña et al., 2013) (NOOR; NASRUDIN; FOO, 2014) (DEB; Ali Ahmed, 2018) e, finalmente, no conforto térmico dos passageiros (SHEK; CHAN, 2008) (PIMENTA; ASSUNÇÃO, 2015) (VELT; DAA-NEN, 2017) (CARDOSO et al., 2018) (PALA; OZ, 2015).

Os trabalhos que estudam o conforto como um fator de importância para medição de danos físicos se interessam pelos efeitos que os fatores cinéticos de uma viagem causam no corpo humano (CASICIOLI; HEUSCH; MCCARTHY, 2011) (SCHUBERT et al., 2017). Os estudos com esse ponto de vista são encontrados com menor frequência, mas trazem resultados interessantes para outros pesquisadores como, por exemplo, o fato de que situações que causam desconforto momentâneo para pessoas jovens podem gerar danos físicos a idosos, devido ao fato de que idosos costumam ter força reduzida nos membros superiores e inferiores, ficando, portanto, mais suscetíveis a quedas em situações de frenagem brusca (CASICIOLI; HEUSCH; MCCARTHY, 2011).

Os trabalhos que estudam os aspectos cinéticos do conforto, em geral, estão relacionadas à coleta e análise de dados de veículos de transporte coletivo que são, na sua maioria, ônibus. Elas podem ser subdivididas nas seguintes categorias: conforto relacionado à trepidação (aceleração vertical) (SEKULIĆ et al., 2013) (SEKULIĆ et al., 2016) (SEKULIĆ et al., 2018) (ERIKSSON; FRIBERG, 2000), conforto relacionado à frenagem, aceleração e curvatura (acelerações longitudinais e laterais) (HOBEROCK, 1977) (BEURIER, 2012) (FREYSS; GIESEN; MUÑOZ, 2013), conforto relacionado

à velocidade (SEREDYNSKI et al., 2014) e, finalmente, conforto relacionado tanto à velocidade quanto à aceleração em qualquer um dos três eixos (ZEEMAN; BOOYSEN, 2013) (SIMONYI; FAZEKAS; GÁSPÁR, 2014).

Os trabalhos citados nessa seção apresentam diversas perspectivas sobre o conforto, tanto sob o ponto de vista psicológico, focando no passageiro e seus hábitos, quanto sob a ótica física, centrada na análise de dados coletados de veículos. Ainda que tenham visões consideravelmente distintas, ambas complementam a percepção do conforto do passageiro.

2.2.1 Percepção de Conforto

Por definição, a percepção é uma opinião pessoal. Dessa forma, uma percepção deve ser analisada com cuidado e, sempre que possível, considerando outros aspectos além daquele que a pesquisa está centrada. A percepção de uma pessoa pode ser coletada de várias formas, sendo elas objetivas ou livres (MARCONI; LAKATOS, 2003), e conscientes ou não (RICHARDSON, 2017). Como exemplo de técnica objetiva, tem-se a aplicação de questionário com questões de múltipla escolha, como técnica livre, tem-se a realização de entrevistas. Como exemplo de técnica consciente, tem-se as técnicas onde a pessoa observada sabe que está sendo analisada, como na aplicação de um questionário e/ou entrevista, já na não consciente a pessoa é observada à distância e seu comportamento é anotado como resposta a percepção do que ela está sentindo, um exemplo é a pesquisa etnográfica.

Em qualquer das técnicas de pesquisa utilizada, são aspectos possíveis de serem considerados na coleta: idade, sexo, origem, situação sócio-econômica (OBORNE; CLARKE, 1975). No entanto, somente esses dados não são suficientes para o entendimento do meio cultural que implicam na visão de mundo do entrevistado. O meio cultural em que um sujeito está inserido é responsável, entre outras coisas, por seus níveis de expectativa. Uma pessoa que é nascida e criada em um país economicamente desenvolvido tem expectativas diferentes de outra nascida em um país subdesenvolvido. Hagerty (1999) aponta que os habitantes de países desenvolvidos, por terem suas necessidades básicas atendidas, esperam por serviços de alta qualidade. Por outro lado, Malhotra et al. (2005) demonstram que pessoas que habitam países menos desenvolvidos, que ainda não alcançaram a satisfação de suas necessidades básicas, toleram serviços de menor qualidade. Por exemplo, no Brasil a infraestrutura viária oferecida aos cidadãos é diferente da oferecida em cidades da Europa e América do Norte, devido à negligência do setor público no oferecimento desse serviço (HERINGER, 2002). Portanto, não é possível generalizar percepções, sejam elas de conforto ou não, entre pessoas que não tenham as mesmas expectativas.

Como visto, é possível perceber que a análise de conforto de passageiros de

transporte coletivo é uma área de pesquisa que tem muitos aspectos a serem considerados. Ela está centrada na coleta das impressões do passageiro e isso pode variar dependendo do contexto e das expectativas de quem avalia. No entanto, partindo da premissa de que as avaliações expressam o que de melhor ou pior foi assimilado pelo passageiro, sempre que uma viagem for dita desconfortável, o passageiro provavelmente estará se referindo ao pior momento da viagem, enquanto que se ele avaliá-la como confortável, estará se referindo ao melhor momento da viagem. Identificar as variáveis que caracterizam uma viagem desconfortável e confortável é um desafio que pode ser mitigado com a utilização de informações de telemetria (ver seção 3.1).

2.2.2 Padrões de Conforto

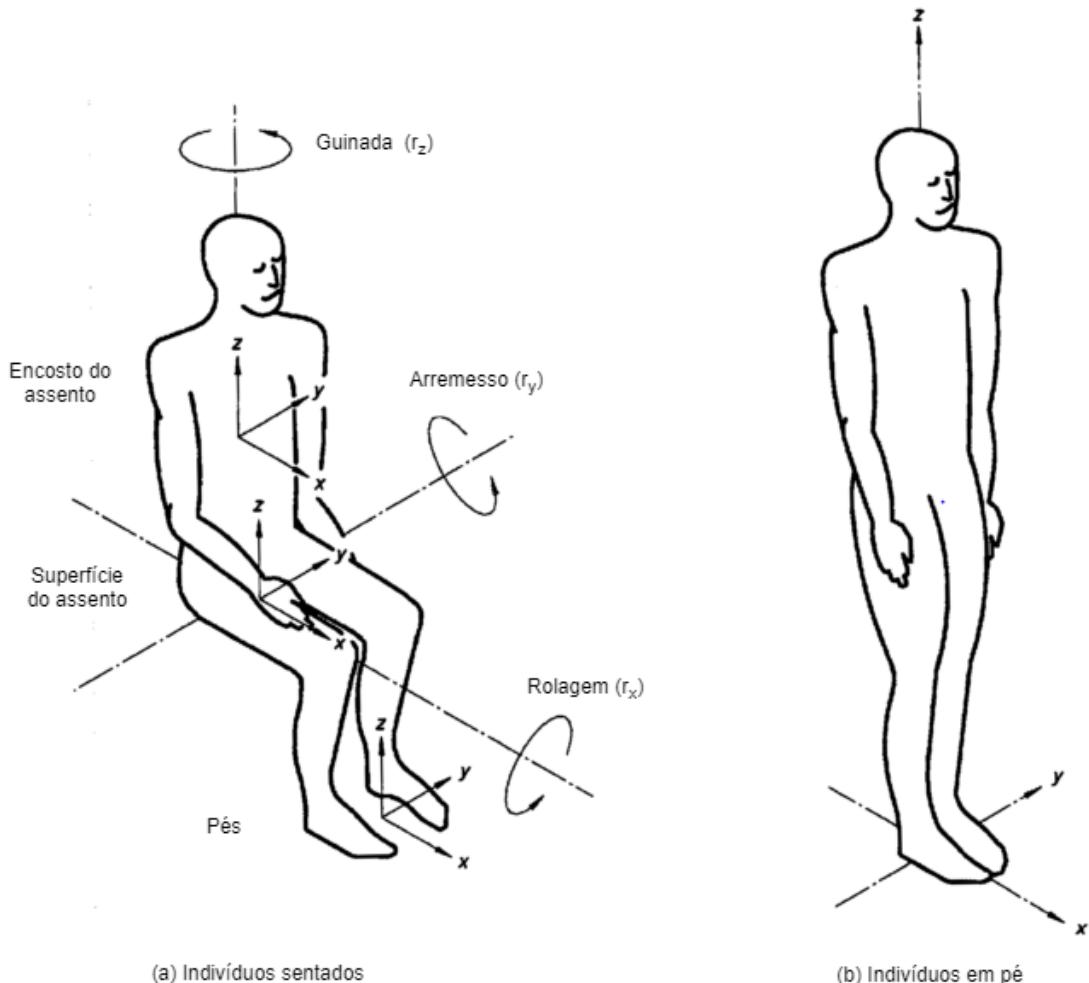
Padrões são documentos que compilam requisitos, especificações, guias ou características que podem ser usadas consistentemente por empresas ou indivíduos, de forma a garantir que materiais, produtos, processos e serviços estejam adequados ao seus propósitos de utilização (ISO, 2018). A *International Organization for Standardization* (ISO) é uma organização mundial dedicada à especificação de padrões internacionalmente reconhecidos e utilizados. Dessa forma, as instituições nacionais costumam utilizar as recomendações da ISO para a criação de leis e regulamentações próprias.

Os padrões ISO com relação ao conforto de passageiros de veículos de transporte coletivo se concentram, em sua maioria, no conforto térmico, tais como a ISO 14505-3:2006 e a ISO 19659-1:2017. Existem ainda padrões relacionados ao conforto no sentido de pontualidade e disseminação de informação, como a ISO/TS 14823:2008.

Em um sentido mais relacionado com o trabalho em tela, podem-se citar a ISO/TR 10687:2012 que discorre sobre as posturas ideais dos passageiros para a diminuição do desconforto e danos físicos relacionados com a vibração do corpo. Além disso, a ISO 26311:1997 estuda passageiros de veículo e seu conforto relacionado à vibração. Ela tem como objetivo principal definir métodos para quantificar as vibrações de corpo inteiro relacionadas com saúde e conforto humano, a probabilidade da percepção da vibração e a incidência de enjoos causados pela vibração. Seu foco é concentrado em veículos que expõem as pessoas a vibrações aleatórias que podem interferir no conforto e saúde, como é o caso do ônibus. A Figura 1 apresenta as forças de aceleração que influenciam no conforto descritas na ISO 26311:1997.

A Figura 1(a) ilustra os três pontos que sofrem efeitos da aceleração em passageiros sentados: dorso, nádegas e pés. Para o indivíduo em pé, Figura 1(b), somente um ponto pode sofrer efeitos da aceleração: os pés. As acelerações que afetam os passageiros são nomeadas pelas letras x , y e z , e representam:

Figura 1 – Acelerações e seus efeitos sobre os seres humanos.



(a) Indivíduos sentados

(b) Indivíduos em pé

Fonte: ISO 26311:1997. Adaptado pela autora.

- x : aceleração horizontal, aquela ocasionada por frenagens e retomadas de aceleração do veículo;
- y : aceleração lateral, aquela que ocorre em curvas;
- z : aceleração vertical, provocada pelas imperfeições da via.

Os efeitos danosos para a saúde do indivíduo são representados pelos movimentos de Guinada, Arremesso e Rolagem. A Guinada é um movimento ocasionado pela aceleração exacerbada em x e y , com manutenção estável da aceleração em z . Esse movimento causa danos ao pescoço e dorso do indivíduo. O Arremesso é um movimento resultante do aumento abrupto das forças em x e z , com manutenção da aceleração em y , fazendo com o que o corpo do indivíduo seja lançado para frente e para cima. Por fim, a Rolagem é um movimento resultante da aceleração súbita em y e z , com manutenção estável da aceleração em x , ocasionando o lançamento lateral e para cima do indivíduo.

De forma a avaliar as medições dos eixos citados acima, a ISO 26311:1997 propõe diferentes métodos matemáticos, que têm por objetivo auxiliar no cálculo do conforto no momento em que a medição é obtida. Além disso, a ISO 26311:1997 esclarece que devido à subjetividade da percepção do conforto, não existe consenso com relação aos limites exatos de conforto de passageiros, em particular pela necessidade de análise da percepção de conforto sentida pelo passageiro. Entretanto, ela fornece limites para prováveis percepções de conforto, considerando a combinação das acelerações nos três eixos, conforme Tabela 1.

Tabela 1 – Limites de provável percepção de conforto de passageiros, como indicados pela ISO 26311:1997.

Limite	Reação
Menor que $0,315 \text{ m/s}^2$	Não desconfortável
De $0,315 \text{ m/s}^2$ a $0,63 \text{ m/s}^2$	Um pouco desconfortável
De $0,5 \text{ m/s}^2$ a 1 m/s^2	Razoavelmente desconfortável
De $0,8 \text{ m/s}^2$ a $1,6 \text{ m/s}^2$	Desconfortável
De $1,25 \text{ m/s}^2$ a $2,5 \text{ m/s}^2$	Muito desconfortável
Maior que 2 m/s^2	Extremamente desconfortável

Fonte: ISO 26311:1997

É importante ressaltar que os limites apresentados na Tabela 1 se sobrepõem, dificultando a descoberta exata da percepção de desconforto dos passageiros. Dessa forma, mesmo com a existência da ISO 26311:1997, diversos autores buscam novas formas de descobrir esses limites e medir com maior exatidão o nível de desconforto de um passageiro.

2.3 MÉTODO DE CLASSIFICAÇÃO

Como não existem limites bem definidos de conforto de passageiros de transporte coletivo, o trabalho em tela buscou técnicas de classificação de dados que permitissem a categorização de dados obtidos por meio da telemetria em função do conforto. Os métodos utilizados na solução proposta neste trabalho são, as árvore de decisão e os Modelos Ocultos de Markov, detalhados nas seções 2.3.2 e 2.3.3, respectivamente.

2.3.1 Tipos de classificação de dados

A classificação de dados se refere ao problema do aprendizado das relações entre um conjunto de atributos e uma variável alvo, classe da instância. Aggarwal (2014) o descreve como: partindo de um conjunto de dados treinamento rotulados com suas respectivas classes que o categorizam, determine a classe de uma instância não classificada.

A classificação de dados é dividida comumente em dois estágios: fase de treino e fase de testes. A fase de treino é responsável pela criação de um modelo que seja capaz de agrupar dados de acordo com suas características; a fase de testes, por sua vez, utiliza o modelo treinado para classificar instâncias não utilizadas no treinamento do modelo (HASTIE et al., 2005).

Os conjuntos de dados utilizados para a classificação podem ser de diferentes tipos: rotulados, não rotulados e parcialmente rotulados. Dessa maneira, existem diferentes técnicas para classificar os dados com base nos tipos citados:

- **Classificação supervisionada:** Utiliza dados marcados ou anotados para o treinamento do modelo. Nessa situação, o modelo de classificação tem à disposição uma classe de anotação para cada instância de dados utilizada para o treinamento (RUSSELL; NORVIG, 2016).
- **Classificação não supervisionada:** Utiliza dados não rotulados para a classificação. Esse tipo de classificação agrupa os dados de acordo com as características de maior significância (HINTON; SEJNOWSKI; POGGIO, 1999).
- **Classificação semi supervisionada:** Utiliza um conjunto de dados que contém um subconjunto de instâncias classificadas e não classificadas. O modelo deve utilizar diferentes técnicas para inferir a classe dos dados não classificados, inclusive uma combinação entre técnicas supervisionadas e não supervisionadas. (ZHU, 2005).

2.3.2 Árvores de Decisão

Uma árvore de decisão é uma ferramenta de suporte à decisão que expressa regras e condições para que um determinado estado seja atingido. São utilizadas para diversos fins e, entre eles, para a previsão de um resultado baseado na análise de dados de entrada (REZENDE, 2003).

Uma árvore de decisão é um tipo de grafo $G = (V, E)$, em que V é um conjunto de vértices (v_1, v_2, \dots, v_n) e E é um conjunto de arestas (e_1, e_2, \dots, e_n) . Cada aresta liga um par de vértices $e_m = (v_i, v_j)$, em que v_i é o vértice de origem e v_j é o vértice de destino da aresta. Dessa forma, as árvores de decisão são grafos direcionados acíclicos, cujos vértices folha são a classificação de determinada instância (WITTEN et al., 2016).

Segundo Witten et al. (2016), para classificar uma instância, deve-se seguir um caminho pela árvore, iniciando pela raiz. Esse caminho é percorrido por nós sucessivos, em que a cada nó deve-se comparar o atributo da instância com o atributo contido no nó. Para nós que contém atributos numéricos, é comum que as arestas

representem comparações “de maior que o atributo” e “menor que o atributo”. Nós que contêm atributos nominais, por sua vez, costumam ter tantos filhos quantos forem os valores possíveis para o atributo. Ainda, é possível que um atributo seja testado múltiplas vezes no mesmo caminho.

Existem diversas situações em que as regras de classificação das instâncias não são conhecidas *a priori* ou, ainda, existem tantas condições de classificação que a combinação delas é trabalhosa e pode ser intratável manualmente por seres humanos. Nesses casos, é possível utilizar algoritmos de aprendizado de máquina para construir a árvore de classificação sem a intervenção de um ser humano.

O problema de construção de uma árvore de decisão pode ser expressado da seguinte forma:

1. Selecione um atributo para ser inserido na posição atual. Se a árvore estiver vazia, esse item será a raiz, senão será um filho de outro vértice;
2. Crie uma aresta para cada avaliação possível do atributo;
3. Para cada aresta, retorne ao passo 1.

O processo acima deve ser repetido recursivamente até que todos os caminhos que saiam da raiz tenham como resultado uma classificação. Esse processo apresenta dois desafios: i) qual é o melhor atributo para ser posicionado na raiz da árvore, e, ii) quais são as avaliações que devem ser feitas em cada ramo da árvore para torná-la o mais acurada e concisa possível.

A medida comumente utilizada para selecionar o atributo escolhido como vértice de um determinado ramo da árvore é aquele que agrupa o maior ganho de informação possível IG , calculado conforme 2.4.

$$IG(S, x) = H(S) - \sum_{v \in x} \frac{|S_v|}{|S|} H(S_v) \quad (2.4)$$

O cálculo de IG é realizado sobre todo o conjunto de amostra de dados $S = \{s_1, s_2, \dots, s_n\}$. Cada amostra s_i é um conjunto que contém atributos x_{ji} e uma classe y_i , tal que $s_i = \{x_{1i}, x_{2i}, \dots, x_{mi}, y_i\}$. Além disso, o cálculo de ganho de informação considera a entropia $H(x)$ para cada atributo x . A entropia pode ser calculada como em 2.5.

$$H(x) = -\sum P(x) \log_2(P(x)) \quad (2.5)$$

A entropia está relacionada com a quantidade de valores que x pode assumir na instância avaliada. Atributos que tem muitas possibilidades de valores em determinada altura da árvore terão maior entropia e, por consequência, maior aleatoriedade, que, por sua vez, implica menor predicabilidade. Dessa forma, a cada iteração do algoritmo será escolhido o atributo que tem o maior IG e, normalmente, o que tiver o menor $H(x)$.

2.3.2.1 Poda de Árvores de Decisão

O processo de construção de árvores de decisão por meio de algoritmos de aprendizado de máquina busca descobrir árvores resultantes que sejam as mais genéricas e mais compactas possível. Esse objetivo se deve ao fato de que árvores muito longas se tornam muito especializadas (*overfitting*), e costumam classificar corretamente apenas os dados da amostra de treinamento, mostrando-se ineficientes para a classificação de outras instâncias. Além disso, árvores mais curtas são mais facilmente interpretadas por seres humanos.

A poda de árvore é utilizada para obter árvores mais curtas e com maior chance de acerto. A poda pode ocorrer durante o processo de construção da árvore, chamado de pré-poda, ou após a construção da árvore, chamado de pós-poda.

A pré-poda é o processo de parada de divisão do conjunto de instâncias e transformação do nó atual em um nó folha (MONARD; BARANAUSKAS, 2003). Essa decisão pode ser tomada pelo algoritmo com base no cálculo do IG . Se o IG de um determinado vértice for menor que um valor estabelecido *a priori*, e não houverem outros atributos possíveis para aquela posição da árvore, o vértice é retirado e a classe com maior frequência passa a ser o vértice folha. Esse método pode ser problemático no sentido de determinar um valor adequado para a poda, pois se esse valor for muito elevado a árvore resultante será simplificada demais, enquanto que se o valor for muito baixo, a árvore será especializada para o conjunto de treino, sem a capacidade de previsão de outras instâncias.

Por sua vez, a pós-poda retira ramos completos da árvore após sua construção (GOUDA; ZAKI, 2001). O algoritmo calcula a taxa de erro com e sem a existência do ramo. Se a taxa não sofrer variação maior que um limite estabelecido *a priori*, o ramo é retirado por completo e a classe com maior frequência entre seus filhos toma o lugar do nó pai. Esse processo pode gerar diferentes árvores, que contêm diferentes ramos. A acurácia das árvores para um conjunto de verificação é calculada, e é escolhida a árvore com maior acurácia. Esse método sofre com os mesmos problemas que o anterior, mas, em geral, apresenta resultados melhores.

2.3.2.2 Algoritmo C4.5

O algoritmo de aprendizado de máquina C4.5, proposto por Quinlan (1993), é uma evolução do algoritmo ID3, proposto pelo mesmo autor em 1975. O ID3 utiliza apenas a métrica de entropia para a construção das árvores de decisão. Dessa forma, esse algoritmo é muito sensível a atributos nominais que possam assumir muitos valores diferentes. O algoritmo não tem capacidade de descobrir regras com base em atributos contínuos. Além disso, dados faltantes não são interpretados por esse algoritmo, o que em aplicações de mundo real é uma situação comum. Por fim, o ID3 tem dificuldades em balancear a importância dos atributos, considerando o mesmo peso para todos.

O C4.5 resolve os problemas citados acima, utilizando a métrica de ganho de informação para diminuir sua sensibilidade; regras a partir de atributos contínuos; criação de condições para o treinamento de dados com atributos faltantes; uso de pesos diferentes para atributos com importâncias diferentes para a árvore. Por outro lado, esse algoritmo não prevê paralelização de processamento e pode ser lento com relação a outras soluções que utilizam técnicas como a de divisão e conquista.

2.3.3 Modelos Ocultos de Markov

O Modelo Oculto de Markov, ou Hidden Markov Model (HMM), é uma especialização do modelo estatístico de Markov em que o sistema a ser modelado contém estados não observados (RABINER; JUANG, 1986). São considerados estados não observados aqueles que não estão visíveis ao observador, ou seja, apenas o comportamento do modelo é visível e a partir dele o estado real pode ser inferido.

Um modelo estatístico de Markov é um processo que tem a capacidade de transacionar os estados do modelo. Além disso, a cada momento todos os estados estão disponíveis e a única dependência para a transição de estado é o estado atual (BILMES et al., 1998). Os modelos estatísticos de Markov são também chamados de cadeias de Markov. Esses modelos têm como uma de suas características o fato de serem estocásticos, ou seja, não determinístico, que ainda que se conheça o estado inicial, não se pode prever o estado final para todas as ocorrências (GARDINER et al., 1985). Um HMM segue essas mesmas características, com estados observados e não observados, em que os não observados devem ser inferidos.

Um HMM é um conjunto finito de estados em que cada estado está associado a uma distribuição de probabilidade que, em geral, é multidimensional. As transições entre os estados são governadas por um conjunto de probabilidades chamado de probabilidade de transição. Em um estado qualquer, uma saída ou observação pode ser gerada de acordo com a probabilidade de transição de estado associada ao dado

de entrada analisado. Portanto, o modelo representa o comportamento dos dados e os estados associados a determinados padrões desses dados.

O principal objetivo de um HMM é inferir, com base na probabilidade de mudança de estado, qual o próximo estado de um sistema, levando em consideração o histórico conhecido de estados representado pelo HMM construído e o último estado observado. Dessa forma, esse modelo é utilizado para a descoberta de padrões em amostras de dados tanto discretos quanto contínuos. Como exemplos de descoberta de padrões discretos têm-se a identificação de movimento com base em fotos ou *frames* parados (WHORISKEY et al., 2017), e a análise de sentimentos (MEDHAT; HAS-SAN; KORASHY, 2014). Como exemplos de descoberta de padrões contínuos têm-se a identificação de padrões de discurso (NAGY; NÉMETH, 2016; HADDAD et al., 2015), e o próximo gene em uma cadeia genética (NIELSEN; SIMONSEN; HOBOLTH, 2016; ELLIOTT; TEH et al., 2016).

Os HMMS podem ser calculados sobre diferentes conjuntos de dados, sejam eles discretos ou contínuos. Para modelos contínuos, é necessário que as distribuições de probabilidade sejam definidas sobre espaços contínuos, de forma que a distribuição seja definida como uma matriz multidimensional de estados e momentos. Para o trabalho em tela, esse tipo de cálculo não é necessário, nem será abordado, portanto as formalizações a seguir são apenas para modelos discretos.

- N : A quantidade de estados possíveis naquele modelo;
- M : A quantidade de observações. Se o modelo for contínuo, M é infinito.
- Λ : O conjunto de probabilidades de transição de estado, tal que $\Lambda = \{a_{ij}\}$, no qual:

$$a_{ij} = p\{q_{t+1} = j | q_t = i\}, 1 \leq i, j \leq N \quad (2.6)$$

em que q_t é o estado atual e que observe as seguintes limitações:

$$\begin{aligned} \sum_{j=1}^N a_{ij} &= 1, 1 \leq i \leq N, \\ a_{ij} &\geq 0, \\ 1 &\leq i, j \leq N \end{aligned} \quad (2.7)$$

- β : A probabilidade de distribuição de cada estado, tal que $\beta = \{b_j(k)\}$, em que:

$$b_j(k) = p\{o_t = v_k | q_t = j\}, 1 \leq j \leq N, 1 \leq k \leq M \quad (2.8)$$

onde v_k denota a k -ésima observação, e o_t é o vetor de parâmetros atual. Além disso, as limitações a seguir devem ser observadas:

$$\begin{aligned} \sum_{k=1}^N b_j(k) &= 1, \\ b_j(k) &\geq 0, \\ 1 &\leq j \leq N, \\ 1 &\leq k \leq M \end{aligned} \tag{2.9}$$

- π : A distribuição inicial de estados, tal que $\pi = \{\pi_i\}$, onde:

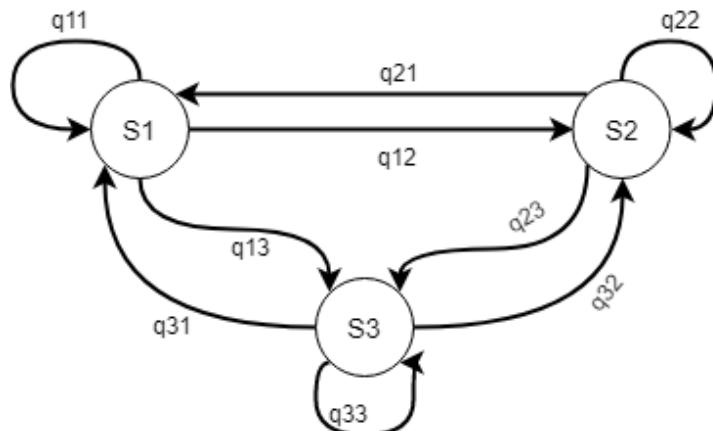
$$\theta_i = p\{q_1 = i\}, 1 \leq i \leq N \tag{2.10}$$

Dessa forma, é possível definir um HMM por meio da notação compacta

$$\lambda = (\Lambda, \beta, \pi) \tag{2.11}$$

Para a compreensão do funcionamento de um HMM, os próximos parágrafos descrevem um exemplo baseado em Rabiner (1989). Primeiramente, considera-se que um determinado sistema pode ser descrito a partir de um conjunto $\{S_1, S_2, \dots, S_n\}$ de N estados distintos, e $\Lambda = \{q_1, q_2, \dots, q_{n^2}\}$ transições de estados. Na Figura 2 o modelo tem $N = 3$, onde cada estado é representado por um nodo, a probabilidade q_{xy} de transição entre dois estados quaisquer é representado por uma aresta e o modelo como um todo é representado pelo grafo completo.

Figura 2 – Modelo Oculto de Markov com $N = 3$ representado como um grafo



Fonte: Autora.

Em momentos discretos M , conforme a segunda formalização da lista apresentada, de forma regularmente espaçados, t_1, t_2, \dots, t_m , o sistema pode mudar de estado, inclusive retornando para o mesmo estado da observação anterior. A cada observação, em cada momento discreto, denota-se o estado atual, também chamado de

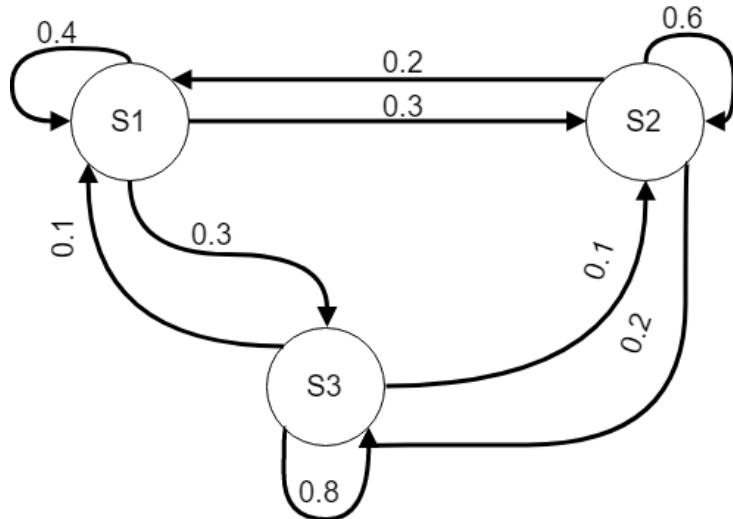
estado real, como q_t . A descrição probabilística de troca de estado, em cada momento t é dada por meio da Tabela 2.

De forma prática, pode-se considerar que um modelo com $N = 3$ estados tenha matriz de probabilidade de transição Λ representada pela Figura 3:

Tabela 2 – Matriz de probabilidade Λ .

	S1	S2	S3
S1	0.4	0.3	0.3
S2	0.2	0.6	0.2
S3	0.1	0.1	0.8

Figura 3 – Modelo de Markov cuja matriz de transição Λ é representada nas arestas do grafo



Fonte: Autora.

A matriz da Tabela 2 representa a probabilidade de cada estado, S_1 , S_2 e S_3 , transacionarem entre si, ou seja, a probabilidade de S_1 retornar a S_1 é 0.4, enquanto que uma transição de S_1 para S_2 é de 0.3 e de S_1 para S_3 é de 0.3. Com base nessa matriz, espera-se descobrir qual a probabilidade da sequência $O = S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3$, considerando-se que no momento observado atual (q_t) se tem S_3 , ou seja, a probabilidade de S_3 em t é $\pi_3 = 1$, pode-se realizar o seguinte cálculo:

$$\begin{aligned}
 P(O) &= P[S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3] \\
 &= P[S_3] \cdot P[S_3|S_3] \cdot P[S_3|S_3] \cdot P[S_3|S_1] \\
 &\quad \cdot P[S_1|S_1] \cdot P[S_1|S_3] \cdot P[S_3|S_2] \cdot P[S_2|S_3] \\
 &= \pi_3 \cdot a_{33} \cdot a_{33} \cdot a_{31} \cdot a_{13} \cdot a_{32} \cdot a_{23} \\
 &= 1 \cdot (0.8) \cdot (0.8) \cdot (0.1) \cdot (0.4) \cdot (0.3) \cdot (0.1) \cdot (0.2) \\
 &= 1.536 \cdot 10^{-4}
 \end{aligned} \tag{2.12}$$

Dessa forma, a probabilidade daquela sequência ser observada naquele momento é de $1.536 \cdot 10^{-4}$, ou seja, 0.15%. Entretanto, a simples inversão de dois estados

pode afetar o resultado, aumentando ou diminuindo a probabilidade de ocorrência. De um modo geral, conforme esperado pela matriz de transição, Tabela 2, a transição de S3 para ele mesmo ocorre mais de uma vez, o que aumenta a chance de ocorrência da observação como um todo, enquanto que a transição de S3 para S1 diminui a chance de ocorrência da observação. Dessa maneira, observa-se que a probabilidade de ocorrência da observação é uma função de combinação das transições.

2.3.3.1 Classificação de dados

Um HMM pode ser utilizado para classificar dados, ou seja, para agrupar dados de acordo com características em comum e utilizar esses grupos para classificar novos indivíduos. Para tanto, a estratégia comumente adotada é a de representar cada classe esperada como um estado. Com essa estratégia, tem-se uma situação similar à classificação supervisionada, pois os dados são anotados *a priori* do treinamento do modelo.

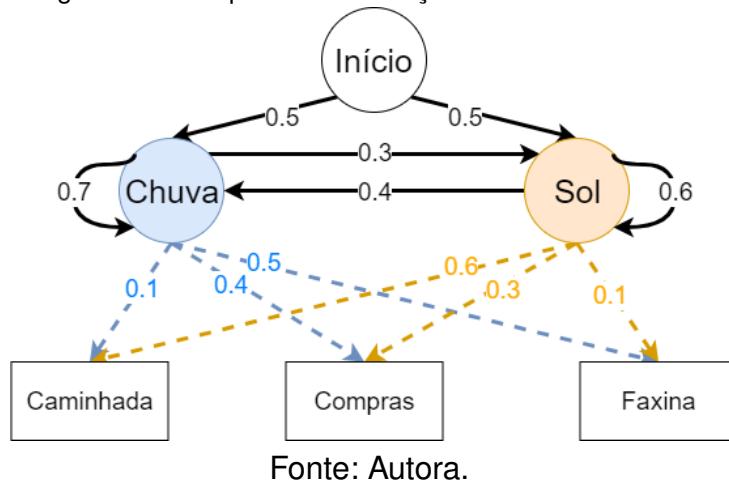
Dessa forma cada dado ou conjunto de dados pode ser classificado em uma classe específica. Para tanto os atributos das instâncias do conjunto de dados são utilizados para o cálculo da probabilidade de emissão entre os estados. Além disso, é possível classificar sequências de instâncias, ou seja, observações realizadas durante um determinado período de tempo, em que o estado imediatamente anterior à observação é utilizado para a classificação. Assim, o dado é classificado como a classe que tem maior probabilidade de ocorrência.

Em um exemplo simples e clássico, apresentado pela primeira vez por Rabiner (1989), a classificação de dados por meio de um HMM pode ser representada pela classificação do clima em determinado dia com base nas observações de hábitos de uma pessoa. A Figura 4 representa esse exemplo. Suponha que a pessoa tenha 3 ações observáveis: fazer uma caminhada, ir às compras ou fazer uma faxina em sua casa, esses são os estados observáveis e são representados na Figura 4 pelos retângulos brancos.

A cada dia essa pessoa realiza apenas essas ações específicas, e essas ações são influenciadas apenas pelo clima do dia em questão, que são chamados de estados ocultos, representados na Figura 4 pelos círculos coloridos. De acordo com a observação do comportamento da pessoa e da anotação do clima no dia, é possível, para dias futuros, calcular a probabilidade de o dia ser chuvoso ou ensolarado, ou ainda, se o estado chuvoso passou a ser ensolarado ou vice versa, representados pela setas sólidas que contém a probabilidade dessas emissões entre os estados, que varia de 0 a 1. Todos os dias iniciam com a mesma chance de serem ensolarados ou chuvosos, o início do dia é representado pelo círculo branco e a probabilidade de emissão do clima é representada pelas setas sólidas que iniciam no estado inicial.

O modelo representado na Figura 4 é baseado no histórico de observações da pessoa e do clima. Dessa maneira, suponha que não é mais possível observar o clima, somente a pessoa e que a pessoa saiu para fazer uma caminhada e depois fez uma faxina. Com esse comportamento, é possível estimar que o dia se iniciou ensolarado, pois a pessoa tem 60% de chance de sair para uma caminhada em dias de sol, versus 10% em dia de chuva, e que após determinado período o dia passa a ser chuvoso, pois existe 40% de chance de mudança do clima ensolarado para o chuvoso, e a pessoa tem 50% de chance de realizar faxina em dias chuvosos, versus 10% em dias ensolarados.

Figura 4 – Exemplo de classificação de dados com HMM.



Fonte: Autora.

A resolução de um HMM implica na resolução de três problemas:

- **Problema da Avaliação:** Dado um HMM λ e uma sequência de observações ω , qual a probabilidade que as observações sejam geradas pelo modelo λ ?
- **Problema da Decodificação:** Dado um modelo λ e uma sequência de observações ω , qual a sequência de estados mais provável que produziria as observações ω ?
- **Problema do Aprendizado:** Dado um modelo λ e uma sequência de observações ω , como o modelo deve se ajustar para maximizar a probabilidade de ω ?

2.3.3.1.1 Problema da Avaliação

O primeiro problema, pode ser resolvido usando algoritmos probabilísticos simples, inclusive soluções determinísticas. Essas soluções podem ser inviáveis, no entanto, quando se considera que serão N^T operações para a descoberta desse resultado, em que N é a quantidade de estados disponíveis e T a quantidade de instantes, ou observações esperadas (SILANON; SUVONVORN, 2011).

Os HMMs, por sua vez, utilizam o algoritmo *forward–backward*, ou algoritmo “para frente e para trás”. Esse algoritmo, conforme seu nome indica, faz duas avaliações dos dados ordenados. Essas avaliações são realizadas primeiro “para frente”, ou seja, da primeira observação para a última, e depois “para trás”, ou seja, da última observação para a primeira (BINDER; MURPHY; RUSSELL, 1997).

Cada avaliação calcula a probabilidade de ocorrência da observação avaliada com base nas observações anteriores. Assim, na primeira avaliação, a primeira observação terá 100% de probabilidade de ocorrência, enquanto que a segunda, se for diferente da primeira, terá 50% e assim sucessivamente.

Uma vez que a avaliação “para frente” é finalizada, o mesmo processo é realizado “para trás”. Finalmente, um processo de combinação das probabilidades é computado. A combinação combina, para cada observação, sua probabilidade “para frente”, multiplicada pela observação “para trás”, multiplicadas pela probabilidade de ocorrência da observação anterior. Esse processo pode ocorrer ao mesmo tempo que o passo “para trás”, uma vez que nesse instante, todas as probabilidades já foram calculadas.

2.3.3.1.2 Problema da Decodificação

Esse problema precisa, primariamente, da definição de maior probabilidade, que, por sua vez, necessita da solução do problema da avaliação anteriormente. Para a resolução desse problema, comumente, se utiliza do algoritmo de *Viterbi* (VITERBI, 2006).

Esse algoritmo também é chamado de algoritmo de soma máxima, ou de produto máximo. Ele cria duas matrizes de duas dimensões, $K \times T$, em que K é a quantidade de estados disponíveis e T é a quantidade de observações disponíveis. A primeira matriz armazena o caminho de estados mais provável até o momento.

A segunda matriz armazena os atributos em que o caminho foi maximizado, ou seja, quais os atributos que criam o caminho de estados mais provável. De maneira geral, o algoritmo retorna a maior probabilidade a cada observação avaliada, portanto a maior probabilidade parcial do modelo.

2.3.3.1.3 Problema do Aprendizado

De maneira geral, esse problema se refere ao ajuste dos parâmetros do HMM, ou seja, de que maneira o modelo deve ser construído para que represente os dados de treinamento da melhor maneira possível, ou ainda, como maximizar a probabilidade de que uma determinada sequência de observações ocorra. Para resolver esse

problema, é comum que seja aplicado o algoritmo *forward–backward* com a adição de duas variáveis: a primeira é a probabilidade de que um determinado estado i em um instante t leve a um estado j em um instante $t + 1$; a segunda é a probabilidade de que um estado j em um instante $t + 1$ seja precedido por um estado i em um instante t .

Observa-se, claramente, que a descrição das duas variáveis pode ser calculada pelo próprio algoritmo, uma vez que ele percorre o conjunto de dados “para frente” e “para trás”. De posse dessas duas variáveis, o algoritmo pode maximizar os parâmetros com base na maximização das variáveis.

2.3.3.2 HMM univariado e HMM multivariado

Um HMM pode ser univariado ou multivariado. Um HMM univariado tem uma única variável associada a todos os estados e suas possíveis emissões dependem apenas dela. Modelos univariados podem ser representados por meio de uma matriz de dados, em que cada posição da matriz representa a probabilidade de transição de cada um dos estados, conforme ilustrado pela Figura 5.

Como exemplo de um HMM univariado, pode-se considerar a modelagem de um sistema que identifica se no momento seguinte estará chovendo ou não, considerando apenas a variável de umidade relativa do ar, em que cada estado (chovendo ou não chovendo) está relacionada com o valor dessa variável.

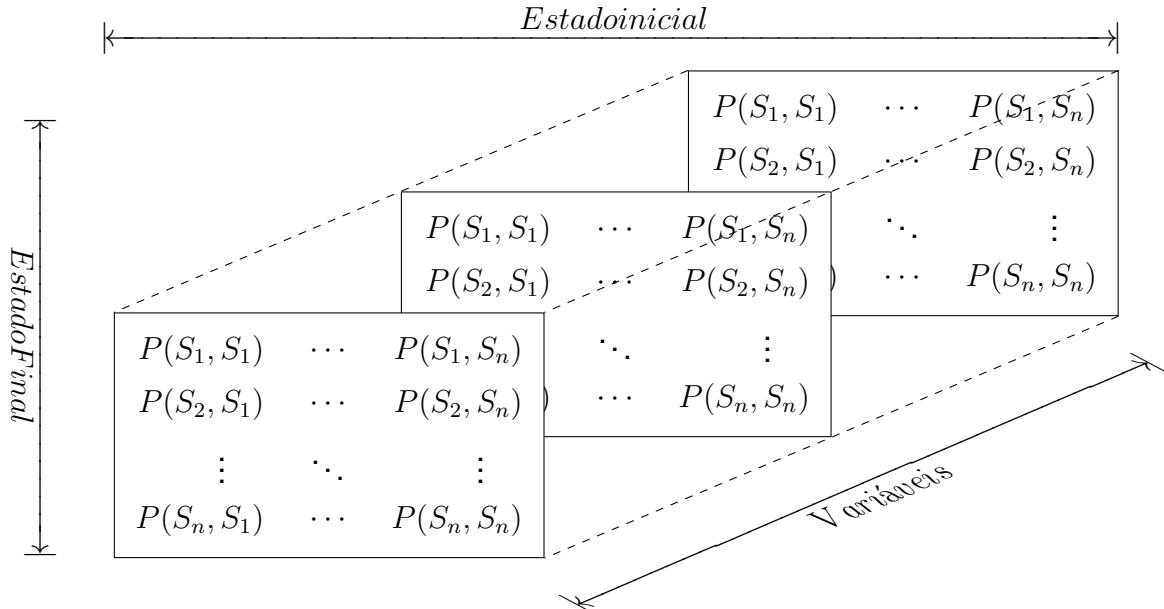
Figura 5 – HMM Univariado, representado por uma matriz de estados.

$$\begin{array}{c} \text{Estado Final} \\ \uparrow \\ \left(\begin{array}{ccc} P(S_1, S_1) & \cdots & P(S_1, S_n) \\ P(S_2, S_1) & \cdots & P(S_2, S_n) \\ \vdots & \ddots & \vdots \\ P(S_n, S_1) & \cdots & P(S_n, S_n) \end{array} \right) \\ \downarrow \end{array}$$

Fonte: Adaptado de (ZUCCHINI; MACDONALD; LANGROCK, 2017).

Por outro lado, um modelo multivariado deve ser representado por uma matriz multidimensional, em que cada dimensão da matriz representa as probabilidades de transição associadas a uma determinada variável, conforme representado pela Figura 6. Utilizando o exemplo da ocorrência de chuva, é possível que o modelo considere diversas variáveis para cada um dos estados, como pluviosidade, temperatura, variação de temperatura, luminosidade, entre quaisquer outras que possam influenciar o estado de chuva ou não. Dessa forma, cada estado está associado a diversas variáveis, com seus diferentes valores, assim como a probabilidade de emissão entre um estado e qualquer outro.

Figura 6 – HMM multivariado representado por uma matriz multimensional de estados.



Fonte: Adaptado de (ZUCCHINI; MACDONALD; LANGROCK, 2017).

2.3.4 Meta Modelo de Classificação

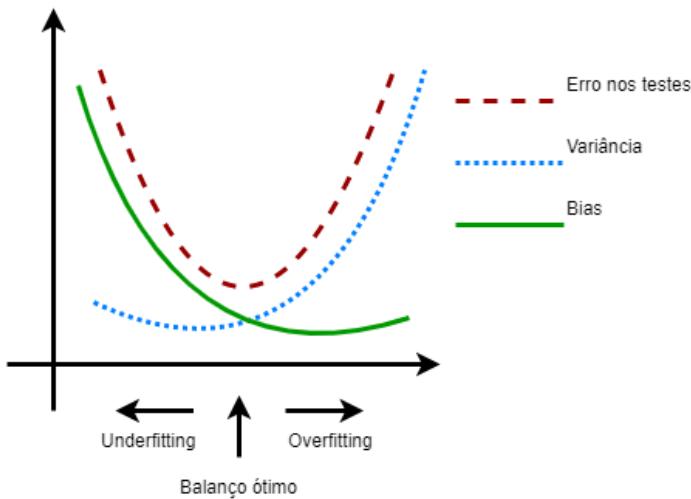
Os meta modelos de classificação, também conhecidos como *ensembles*, são modelos preditivos compostos pela integração de diversos outros modelos (DIETTERICH, 2000a). A principal motivação para o uso desse tipo de abordagem é considerar o peso de diferentes classificadores individuais e combiná-los a fim de obter um modelo que tenha melhor desempenho que cada um dos modelos individuais (POLIKAR, 2006).

Um dos grandes problemas enfrentados pelo aprendizado de máquinas está relacionado com o tratamento de *bias* e da alta variância do resultado entre treinamentos (RAYANA; ZHONG; AKOGLU, 2016). É esperado que exista variância no resultado de diferentes treinamentos com o mesmo dataset, mas essa variância costuma ser apenas marginal (DOSIO, 2016). Alto *bias* implica baixa variância, ou seja, diferentes treinos e testes terão resultados similares, no entanto o modelo dificilmente será generalizável para todo o conjunto de dados, esse fenômeno é chamado de *underfitting*.

Por outro lado, *bias* muito baixo e variância muito alta implicam em treinos e testes muito diferentes, o que também significa acentuada dificuldade de generalização. A solução para essa questão está no equilíbrio entre *bias* e variância, conforme demonstrado na Figura 7.

Os resultados dos modelos devem ser diferentes para que a técnica de meta classificação seja utilizada. Se um conjunto de modelos obtiverem o mesmo resultado, a utilização de uma meta classificação terá o mesmo resultado que os modelos

Figura 7 – Equilíbrio entre variância e *bias*.



Fonte: Autora.

integrados (BUIZZA; MILLEER; PALMER, 1999).

Um *ensemble* precisa de diversidade entre os modelos, ou seja, cada modelo deve ser especialista em classificar diferentes tipos de instâncias, de forma que o conjunto de modelos tenha menor dificuldade em classificar corretamente o maior número de instâncias. Dessa maneira, o meta modelo tem um resultado melhor que cada um dos modelos individuais.

Existem diferentes formas de construir um *ensemble*, as principais são *bagging*, *boosting* e *stacking*. As subseções a seguir resumem cada uma delas.

2.3.4.1 *Bagging*

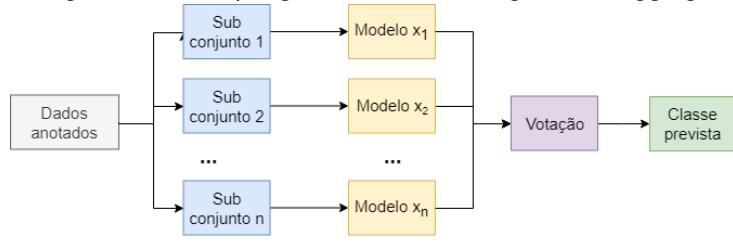
Bagging, sigla derivada da expressão *Bootstrap Aggregating*, é a técnica mais antiga de *ensemble*, em que o conjunto de treinamento é dividido em diversos subconjuntos e cada subconjunto é utilizado para o treinamento de diferentes modelos que utilizam o mesmo algoritmo (DIETTERICH, 2000b). Após o treinamento, esses modelos são utilizados para a previsão de cada instância de teste. Os resultados são comparados de forma que a classe prevista pela maioria dos modelos é considerada correta. A Figura 8 apresenta a dinâmica de funcionamento desse *ensemble*, desde a divisão dos dados até a votação para a seleção da classe adequada para a instância.

Essa abordagem deve ser utilizada em situações em que o uso de um modelo único resulta em *overfitting*, devido a presença de um menor *bias* (GALAR et al., 2011).

2.3.4.2 *Boosting*

Essa técnica também realiza a divisão do conjunto de treinamento e combina seus resultados. Diferente do *bagging*, o *boosting* treina diferentes modelos “fracos” e

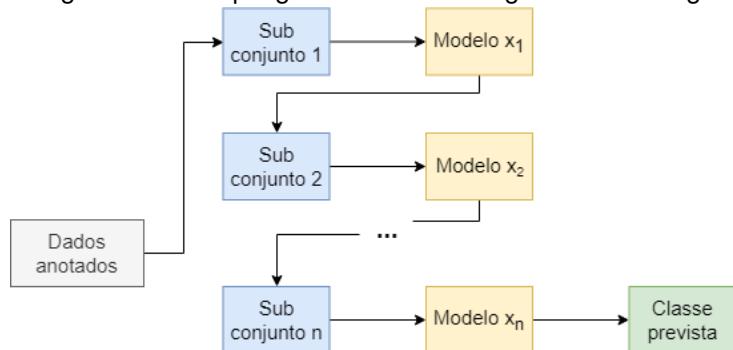
Figura 8 – Exemplo genérico da abordagem de bagging.



Fonte: Autora.

os combina deixando-os mais “fortes”. Um modelo é considerado fraco quando seus resultados são apenas marginalmente melhores do que a escolha aleatória de uma classe. Por exemplo, em um conjunto com duas classes possíveis, o modelo poderia ter uma acurácia de 55%, enquanto que a seleção aleatória seria próxima de 50%. Dessa maneira, um classificador fraco C_1 é criado com um subconjunto de treinamento, e um modelo fraco C_2 é treinado considerando as instâncias mais informativas para C_1 , e assim subsequentemente para cada modelo do *ensemble* (DIETTERICH, 2000a). A Figura 9 apresenta o fluxo de dados entre os modelos.

Figura 9 – Exemplo genérico da abordagem de boosting.

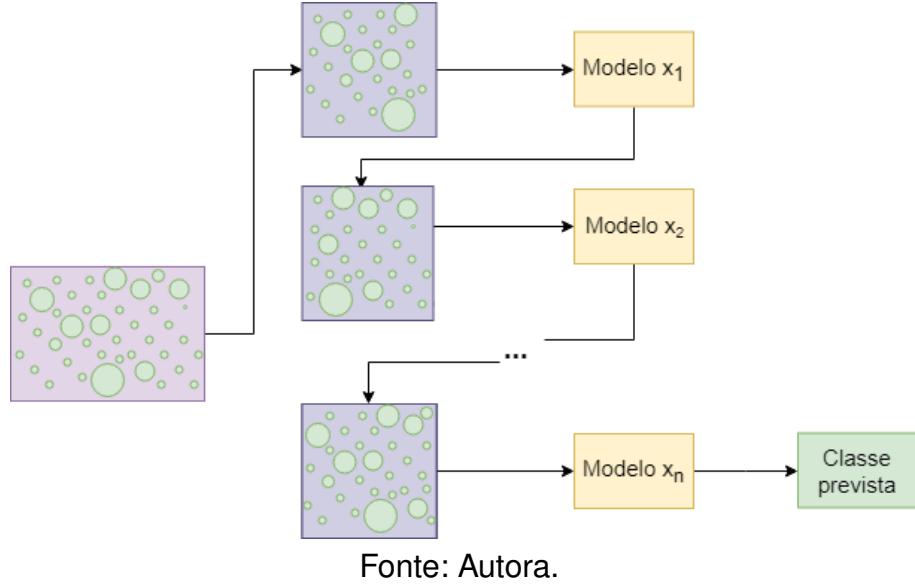


Fonte: Autora.

Os dados mais significativos para um modelo qualquer não são necessariamente mais significativos para o próximo, uma vez que o objetivo da utilização dessa abordagem é refinar os resultados com base nos mesmos dados, com diferentes recortes e pesos (OPITZ; MACLIN, 1999). A Figura 10 representa os dados e seus pesos com relação ao conjunto como um todo, dessa forma, um dado significativo é representado por um círculo maior do que outros dados.

Boosting tem melhor desempenho em situações em que o modelo único tem baixa variância. Essa abordagem, por sua vez, pode resultar em maior *overfitting*, pois os dados utilizados para os modelos subsequentes sempre serão os mais significativos para o modelo anterior (DRUCKER et al., 1994). Dessa forma, é possível que o modelo se especialize nesses casos significativos e ignore os casos menos significativos.

Figura 10 – Exemplo genérico do refinamento de resultados com os subconjuntos de dados e seus pesos.

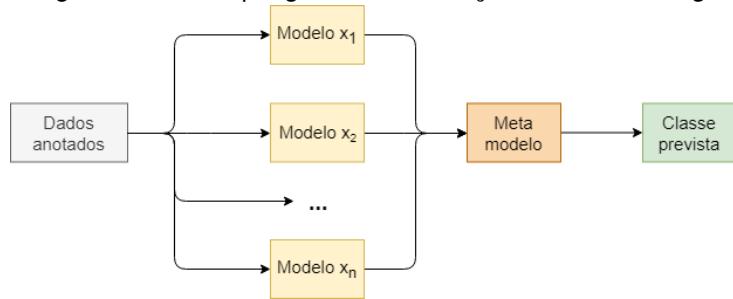


Fonte: Autora.

2.3.4.3 *Stacking*

Enquanto as outras técnicas costumam utilizar o mesmo algoritmo para o treino de todos os modelos, o *stacking* utiliza diferentes algoritmos. Os resultados desses algoritmos são combinados de forma mais sofisticada do que em um *bagging*, pois é utilizado um modelo agregador para computar a predição, com base nos resultados dos classificadores base (XING et al., 2016). Dessa forma, as diferentes características dos dados podem ser exploradas pelos diferentes modelos, e serem consideradas com seus pesos de forma ponderada. A Figura 11 apresenta um modelo genérico de um *ensemble* do tipo *Stacking*.

Figura 11 – Exemplo genérico de criação de um *Stacking*.



Fonte: Autora.

2.4 TRABALHOS RELACIONADOS

Nesta seção são discutidos os trabalhos encontrados na literatura que se relacionam com o problema objeto desta pesquisa. Esses trabalhos tratam sob algum aspecto o problema da análise de conforto de passageiros em veículos de transporte coletivo.

O trabalho de Hassan, Hawas e Ahmed (2013) objetiva auxiliar engenheiros e autoridades a medir a eficiência e eficácia dos serviços de transporte coletivo. O trabalho propõe um arcabouço para a medição de diversos fatores que influenciam na qualidade do transporte coletivo organizados em classes de critérios. Como exemplos de critérios relacionado a rota, têm-se: o estado das vias, o estado dos pontos de ônibus e o estado dos terminais. Já como exemplos de critérios associados ao sistema, têm-se: o tempo de viagem, a sensação de segurança, o conforto e a limpeza. Uma vez que os critérios são definidos para o contexto a ser avaliado, um grupo de interessados e especialistas devem definir a importância de cada critério. Então, os critérios devem ser divididos em características a serem medidas, tanto de forma subjetiva quanto objetiva. Segundo os autores, dados de conforto e sensação de segurança podem ser colhidos com base em questionários e entrevistas, ou seja, de forma subjetiva, enquanto dados de tempo e distância podem ser colhidos de forma objetiva, com uso de ferramentas próprias para esse fim. Na sequência, os dados são analisados considerando as médias e são agrupados por meio de uma média ponderada, em que os pesos são definidos pelos especialistas. Os dados são agrupados por meio da aplicação do algoritmo *K-means*, de forma a encontrar a similaridade entre os dados subjetivos e objetivos, e uma nota de desempenho é atribuída à viagem. Os autores concluem que esse arcabouço pode ser utilizado para a identificação de aspectos do transporte que estão falhando qualitativa e quantitativamente. O arcabouço apresentado considera a opinião de especialistas, mas não a percepção dos usuários. Dessa forma, é possível que a nota de desempenho atribuída pelos especialistas não reflita a realidade experimentada pelos passageiros.

O trabalho de Innamaa e Penttinen (2013) estuda o efeito da utilização de um sistema de *feedback* para o motorista. O *feedback* exibido para o motorista contém dados carregados no sistema *a priori*, com informações de rotas e limites de velocidade. O sistema sugere acelerações e velocidades para cada trecho da viagem e guarda os dados coletados da viagem para análise posterior. Os autores compararam dados de cada viagem com e sem a utilização do sistema por meio de distribuições estatísticas. Os autores discutem que os motoristas que tiveram acesso ao sistema de *feedback* tiveram diminuição no consumo de combustível, na velocidade de cada trecho em comparação com a velocidade esperada e melhoria na avaliação dos passageiros sob o conforto percebido. O trabalho demonstra que a ausência de acelerações bruscas melhora a percepção de conforto dos passageiros, mas não considera outras variáveis que podem ser importantes para o conforto da viagem, como aquelas propostas pela ISO 26311:1997, como discutido na seção 2.2.2. Além disso, o trabalho agrupa os dados por categorias de motorista, que são motoristas sem experiência, com pouca experiência e com muita experiência de direção. Por fim, os autores concluem que o sistema de *feedback* tem impacto proporcional à experiência de direção do motorista,

ou seja, quanto maior a experiência, mais rápida é a resposta do motorista e maiores são os benefícios ao passageiro. O trabalho analisa conforto de passageiros, mas não apresenta um modelo replicável. Sua principal preocupação é como o *feedback* instantâneo pode afetar a condução do veículo.

O trabalho de Castellanos e Fruett (2014) descreve um sistema que mede os fatores que afetam o conforto no transporte coletivo para auxiliar na criação de políticas públicas que promovam melhorias no transporte coletivo. O trabalho discute que o levantamento desses fatores com base em pesquisas e questionários aplicados ao público em geral são trabalhosos e pouco confiáveis. Além disso, menciona a dificuldade em descobrir métricas para percepções individuais. De forma a automatizar o processo de coleta de dados, o trabalho descreve uma ferramenta que evidencia os fatores de conforto com base em um equipamento que coleta informações do ambiente e da percepção do usuário. Os dados de ambiente são coletados por sensores de aceleração triaxial, de geolocalização e de temperatura. Os dados de percepção do usuário são coletados por um equipamento instalado na parte central do ônibus, onde os passageiros podem acionar diferentes botões para informar seu nível de desconforto naquele momento. Com base nos dados coletados dos sensores e dos passageiros, um conjunto de algoritmos calcula os limites de conforto e desconforto da viagem nos momentos apontados pelos passageiros. Os limites de conforto considerados inicialmente pelo algoritmo são baseados nos apresentados pela ISO 26311:1997, e são recalculados a cada entrada realizada pelos passageiros. Os resultados são transmitidos para um software que exibe, em uma plataforma on-line, os momentos de desconforto da viagem. Apesar de apresentar uma abordagem para a medição de conforto, esse trabalho não generaliza essa medição na forma de um modelo que possa ser replicado para outros casos semelhantes. Portanto, sempre que se deseja identificar o conforto dos passageiros em um viagem é necessário coletar a percepção deles e calibrar os limites definidos inicialmente pela ISO 26311:1997.

O trabalho de Zhang, Zhou e Zhang (2014) propõe um modelo geral de previsão de conforto em ônibus aplicado em cidades chinesas. Para a criação do modelo foram colhidos dados de som, temperatura, vibração e aceleração. Os dados são agrupados e comparados com diferentes métricas e medidas disponíveis em padrões nacionais e internacionais. Os dados das variáveis são associados com base na correlação de Pearson e sofrem múltiplas análise de regressão para analisar a dependência estatística entre as variáveis. Além da coleta e análise de dados, os autores aplicaram um questionário sobre o conforto dos passageiros e sua situação sócio-econômica. A partir da análise dos dados, os autores propuseram uma fórmula matemática para o conforto de passageiros. Essa fórmula considera as sete variáveis citadas (som, temperatura, vibração, aceleração em três eixos) e, além disso, a idade, o gênero e

o estado de saúde do passageiro. A conclusão dos autores é que a variável que tem maior efeito sobre o conforto é a aceleração, que tem 88% de importância, seguida pelo gênero do passageiro com 5.6% de importância, e pela saúde com 2.8% de importância. Como o modelo resultante desse trabalho é uma fórmula fixa, inclusive para os dados sócio-econômicos, sua aplicação em outras regiões pode ser limitada, uma vez que conforto é um aspecto subjetivo e, portanto, afetado por questões culturais. Além disso, os autores admitem que a fórmula pode não ser válida para todos os casos, pois: 1) não é possível generalizar um sentimento e 2) pode existir outras variáveis que afetam o conforto de passageiros que não foram medidas. Como exemplos de variáveis a serem adicionados à formula têm-se a aceleração lateral e a aceleração longitudinal.

O trabalho de Eboli, Mazzulla e Pungillo (2016) teve como propósito encontrar métricas e medidas de conforto que possam ser utilizadas no sentido de atrair mais passageiros para os meios de transporte coletivo. O trabalho discute que no passado o conforto era medido apenas com base nas recomendações de normas internacionais, mas que atualmente é possível utilizar dados e técnicas algorítmicas para inferir o conforto em determinado momento do tempo. O trabalho considera medidas cinéticas, como vibração e curvas, para a medida de conforto. Para tanto, os autores coletam dados subjetivos, por meio de questionários e entrevistas, e dados objetivos, por meio de sensores de dispositivos móveis. São considerados os dados de aceleração nos três eixos, velocidade e curvatura. Os dados objetivos são analisados e um limite de conforto é definido. Esse limite é calculado pela média dos pontos que estão na parte superior da distribuição dos dados, ou seja, aqueles dados que são maiores do que a média somada ao desvio padrão. Os autores escolheram analisar apenas as entrevistas que expressaram desconforto. Os dados objetivos são relacionados com as entrevistas selecionadas por meio da observação da distribuição de dados com relação a média. Nas situações em que os dados estão acima da média, os autores consideram que houve desconforto. O trabalho aborda conforto de um ponto de vista simplista, pois considera apenas a média dos dados objetivos coletados e verifica se nesses casos a percepção dos passageiros foi similar ao resultado observado. O trabalho não apresenta um método para a avaliação de conforto.

O trabalho de Park et al. (2017) analisa o efeito das vibrações em veículos estacionários no conforto de passageiros. As vibrações analisadas nesse trabalho podem ser provenientes de uma ventoinha de refrigeração do motor, da rotação do motor ainda que o veículo esteja estacionado ou de qualquer outra parte que possa causar vibração do chassis do veículo. Para a análise dessa situação, um simulador foi construído e diversos voluntários foram entrevistados após utilizarem o simulador. Os voluntários foram expostos a vibrações equivalentes às reais de um veículo e a

outras vibrações de maior amplitude. O estudo concluiu que os limites de conforto de vibração apresentados pela ISO 26311:1997 não são suficientes para a avaliação do impacto da vibração no conforto. As respostas obtidas pela pesquisa sugerem a diminuição do limite inferior da caracterização do início de desconforto de 4Db para 3,5Db. Como conforto é uma percepção particular e cultural, estudos com mais passageiros em situações reais são necessários para a validação das conclusões do estudo.

2.4.1 Considerações sobre os Trabalhos Relacionados

De forma geral, os trabalhos relacionados demonstram que a análise de conforto no transporte coletivo vem sendo gradual e consistentemente abordada em pesquisas científicas, e que os avanços das Tecnologias de Informação e Comunicação (TICs) têm auxiliado diretamente na proposição de novas soluções. Dentre os principais avanços, pode-se destacar a ampliação do uso de sensores embarcados nos veículos, particularmente os sensores de aceleração e de posicionamento (GPS). Além disso, dispositivos móveis como *smartphones* têm auxiliado consideravelmente na interação entre os passageiros e os sistemas de análise de qualidade, visto que alguns dos trabalhos relacionados se beneficiaram do uso de equipamentos eletrônicos instalados nos ônibus para coletar a avaliação dos passageiros.

As abordagens mais recentes estão centradas na utilização de informações de aceleração, vibração e velocidade do veículo combinadas com o *feedback* dos passageiros sobre sua percepção de conforme. Essas informações são utilizadas com entrada de modelos estatísticos, matemáticos ou de aprendizado de máquina que têm por objetivo estimar o nível de conforto dos passageiros. Ainda, grande parte dos trabalhos estão interessados no aumento do conforto dos passageiros como catalisador para aumentar o interesse pelo uso do transporte coletivo.

A Tabela 3 apresenta um resumo dos trabalhos analisados na seção 2.4. A primeira coluna lista os trabalhos em ordem cronológica. A segunda coluna resume as abordagens de avaliação do conforto utilizadas pelos trabalhos. São consideradas abordagens matemáticas aquelas que utilizam fórmula fixas para avaliação do conforto definidas pelos seus autores, como análise empírica aquelas que utilizam observações pontuais para definir métodos de avaliação do conforto e, por fim, de aprendizado de máquina aquelas que utilizam ampla quantidade de dados para construção automática de modelos de avaliação do conforto. A terceira coluna identifica quais trabalhos, de alguma forma, coletam a percepção dos passageiros. A quarta coluna destaca quais trabalhos utilizam algum aspecto cinético durante a análise dos dados. A quinta coluna classifica os trabalhos de acordo com a apresentação explícita da estratificação dos limites de conforto. Por fim, a sexta coluna elenca a quantidade de aspectos cinéticos considerados por cada trabalho. A última linha da tabela classi-

fica o trabalho em tela em comparação com os discutidos nessa seção.

Conforme pode ser observado na Tabela 3, dentre os trabalhos revisados, apenas o trabalho de Innamaa e Penttinen (2013) não coleta a percepção dos passageiros para analisar o conforto. Esse mesmo trabalho é o único a não considerar aspectos cinéticos da viagem na análise. Percebe-se, portanto, que a combinação entre dados de percepção e coleta de sensores é largamente utilizada na literatura. Entretanto, o trabalho que não considera a percepção de passageiros está amparado no entendimento de especialistas sobre os aspectos relevantes a serem analisados e seus níveis de importância. Uma vez de posse das informações dos especialistas, é gerado um modelo matemático que considera indiretamente aspectos cinéticos para a avaliação de conforto.

Apesar de suas relevantes contribuições, algumas limitações são encontradas nos trabalhos revisados. Dentre elas, pode-se destacar a utilização de abordagens matemáticas que usam fórmulas fixas. Esse tipo de abordagem dificulta a generalização ou mesmo especialização do cálculo de conforto para outros cenários particulares de transporte coletivo. Ainda, outra limitação é que nenhuma das abordagens se beneficia da utilização de grandes massas de dados para determinar os níveis de conforto dos passageiros. Esse tipo de informação é de fácil obtenção por meio de sensores. Por fim, nenhuma delas combina as interferências entre as variáveis associadas ao conforto, apenas avaliam o impacto de cada variável isoladamente sobre o conforto do passageiro. Ou seja, avaliam o impacto das acelerações independente da velocidade do veículo e vice-versa. Por meio da utilização de métodos de descoberta de padrões providos pela área de aprendizado de máquina é possível combinar diferentes aspectos de conforto de forma automática, o que representa uma contribuição até então não utilizada pelos trabalhos relacionados.

Portanto, o trabalho em tela propõe a utilização de técnicas de classificação providas pela área de aprendizado de máquina para a criação de um modelo de classificação do conforto que leve em consideração as dependências entre as diferentes variáveis associadas à condução do veículo que afetam o conforto dos passageiros. Essa proposta leva em consideração que a percepção dos passageiros e a análise dos aspectos cinéticos dos trajetos são parte integrante do cálculo do conforto.

Tabela 3 – Resumo dos trabalhos relacionados

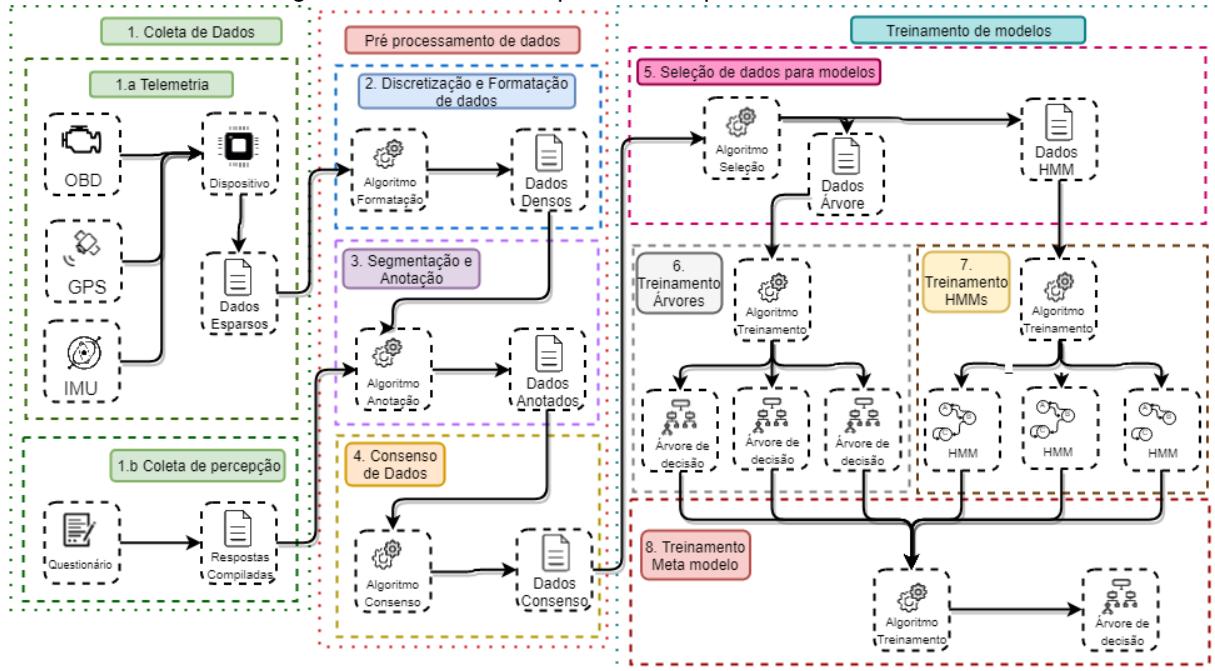
Trabalho	Modelo Proposto	Coleta percepção	Considera aspectos cinéticos	Apresenta limites de conforto	Número de características cinéticas avaliadas
Hassan, Hawas e Ahmed (2013)	Análise TOPSSIS - Algoritmo multicritério de tomada de decisão	Sim	Sim	Não	0
Innamaa e Penttinen (2013)	Cálculo Matemático	Não	Não	Não	0
Castellanos e Fruett (2014)	Cálculo baseado na ISO2631-1	Sim	Sim	Sim	1
Zhang, Zhou e Zhang (2014)	Análise estatística da correlação das variáveis de conforto	Sim	Sim	Não	2
Eboli, Mazzulla e Pungillo (2016)	Observação das distribuições dos dados e suas médias	Sim	Sim	Não	1
Park et al. (2017)	Análise empírica da reação dos passageiros à vibração	Sim	Sim	Não	1
Trabalho em tela	Algoritmos de aprendizado de máquina para classificação dos dados	Sim	Sim	Sim	7

Fonte: Tabela gerada pela autora.

3 MODELO PROPOSTO

O modelo proposto neste trabalho é baseado em um conjunto de atividades apresentado na Figura 12 e descritos nas subseções deste capítulo. O processo começa pela identificação da fonte de dados adequada para a assimilação dos aspectos cinéticos que impactam no conforto do passageiro, Figura 12.1.a, apresentada e discutida na seção 3.1. A seguir, é necessário aferir o conforto percebido pelos usuários do transporte coletivo por meio de um questionário, Figura 12.1.b, apresentado na seção 3.2. A partir da obtenção dos dados de telemetria e das percepções dos passageiros, é necessário proceder a limpeza, a segmentação e a anotação dos dados para posterior treinamento, Figuras 12.2, 12.3 e 12.4, conforme descrito na seção 3.3. De posse do conjunto de dados preparados e anotados, realiza-se a divisão dele em dois conjuntos, um utilizado para o treinamento dos modelos de árvore de decisão e ocultos de Markov e o outro utilizado para a aferição da acurácia, Figura 12.5, conforme descrito na seção 3.4. Por fim, são realizados os treinamentos dos diferentes modelos de aprendizado de máquina utilizados para a identificação do nível de conforto dos passageiro, Figuras 12.6, 12.7 e 12.8, conforme descrito na seção 3.5. A implementação do modelo proposto, bem como sua validação, são discutidos nos capítulos 4 e 5, respectivamente.

Figura 12 – Processo implementado pelo trabalho em tela.



Fonte: Imagem gerada pela autora.

3.1 DADOS DE TRAJETÓRIAS

Os aspectos cinéticos que podem influenciar as percepções de conforto de passageiros são variados e têm diferentes fontes. Para a construção do modelo apresentado no trabalho em tela, foi selecionado um conjunto de dados provindo de um dispositivo instalado em um ônibus da cidade de Joinville. Esse dispositivo, bem como a forma com a qual ele foi construído, é discutido no trabalho de Bayer (2017).

De forma prática, os dados disponíveis para a construção do modelo são descritos na Tabela 4.

3.2 QUESTIONÁRIO

Conforme discutido na seção 2.2.1, a percepção de conforto é fator importante para medição do conforto do passageiro. Dessa forma, um questionário foi desenvolvido e aplicado aos passageiros do veículo que estava equipado com o dispositivo de coleta.

A aplicação do questionário foi realizada por funcionários da empresa que administra os ônibus da cidade e tem como foco a percepção de conforto com relação a acelerações durante a viagem. As perguntas são centradas nas acelerações devido ao exposto no capítulo 2, em particular devido ao fato de que as normas internacionais, seção 2.2.2, dão grande ênfase a esse tipo de dado. Dessa forma, este trabalho também assume que os efeitos das acelerações podem ser determinantes na percepção de conforto dos passageiros e, portanto, os atributos associados a forças cinéticas aplicadas sobre o passageiro são o foco principal das perguntas contidas no questionário.

O questionário é composto por 3 perguntas descritivas, que objetivaram posicionar geograficamente o trajeto no qual o passageiro viajou e que formaram sua percepção, além de 4 perguntas objetivas com escala de *likert*, indo de muito confortável a muito desconfortável, que objetivaram coletar a percepção do passageiro com relação ao conforto durante a viagem sob a perspectiva de trepidações, curvas, freadas e arrancadas, e da viagem como um todo. Essas percepções estão diretamente relacionadas com os dados dos eixos de aceleração x, y e z, que, por sua vez, estão relacionados com o conforto como representado na Figura 1.

3.3 PRÉ-PROCESSAMENTO

O conjunto de dados resultante da coleta de telemetria são dados brutos que necessitam de tratamento para garantir que os resultados de um modelo construído com base neles não seja influenciado por valores espúrios, também chamados de

,
Tabela 4 – Dados disponibilizados pelo dispositivo instalado no veículo.

Origem #	Atributo	Frequência	Unidade	Precisão	Descrição
0	TIMESTAMP_MILLIS				Milissegundos a partir do inicio do sistema
Controle	INFO_CODE				Código de diagnóstico (erros/informações/outros)
2	INFO_EXTRA				Informações complementares sobre o INFO_CODE
OBD 3	MIL_STATUS	1/20m			Indicadores de auto-diagnóstico
4	OBD_STANDARD	1/20m			Padrão OBD da conexão
5	VEHICLE_SPEED	-	Km/h	1	Velocidade (não disponível)
6	ENGINE_RPM	2/s	RPM	0.25	Taxa de rotação do motor
7	ENGINE_LOAD	2/s	%	0.5	Carga do motor
8	BATTERY_VOLTAGE	1/2m	V	0.1	Tensão da bateria
9	COLANT_TEMPERATURE	1/2m	°C	1	Temperatura de arrefecimento
10	ENGINE_THROTTLE	-	%	0.5	Posição da borboleta (não coletado)
11	INTAKE_MANIFOLD_PRESSURE	2/s	kPa	1	Pressão do coletor
12	INTAKE_AIR_TEMPERATURE	1/2m	°C	1	Temperatura de admissão do ar
13	ACCELERATOR_PEDAL_D	2/s	%	0.5	Posição do pedal do acelerador
14	ENGINE_RUN_TIME	1/20m	s	1	Tempo de motor ligado
GPS 15	DATE	1s	YY-MM-DD	1s	Data UTC
16	TIME	1/s	HH:MM:SS	1s	Hora UTC
17	LATITUDE	1/s	°	10 ⁻⁷	Latitude
18	LONGITUDE	1/s	°	10 ⁻⁷	Longitude
19	ALTITUDE	1/s	m	5*HDOP	Altitude acima de WGS84
20	SPEED	1/s	km/h	5*VDOP	Velocidade
21	FIX_QUALITY	1/s			Booleano indicando conexão GPS
22	PRN	1/s			Identificador de satélites fixação
23	PDOP	1/s		10 ⁻¹	Diluição da precisão 3D
24	HDOP	1/s		10 ⁻¹	Diluição da precisão horizontal
25	VDOP	1/s		10 ⁻¹	Diluição da precisão vertical
IMU 26	ACCELEROMETER_X	10/s	g	10 ⁻⁴	Acelerômetro em eixo X (vertical)
27	ACCELEROMETER_Y	10/s	g	10 ⁻⁴	Acelerômetro em eixo Y (lateral)
28	ACCELEROMETER_Z	10/s	g	10 ⁻⁴	Acelerômetro em eixo Z (frontal)
29	GYROSCOPE_X	10/s	°/sec	10 ⁻³	Giroscópio em eixo X (vertical)
30	GYROSCOPE_Y	10/s	°/sec	10 ⁻³	Giroscópio em eixo Y (lateral)
31	GYROSCOPE_Z	10/s	°/sec	10 ⁻³	Giroscópio em eixo Z (frontal)
32	MAGNETOMETER_X	-	°C	10 ⁻³	Magnetômetro em eixo X (vertical, não coletado)
33	MAGNETOMETER_Y	-	°	10 ⁻³	Magnetômetro em eixo Y (lateral, não coletado)
34	MAGNETOMETER_Z	-	°	10 ⁻³	Magnetômetro em eixo Z (frontal, não coletado)
35	TEMPERATURE	10/s	°	10 ⁻²	Temperatura IMU

Fonte: (BAYER, 2017)

outliers, ou subconjuntos demasiadamente grandes que não representem informações significativas para o modelo e causem desperdício de recursos, como é o caso dos *staypoints*.

Para tanto, no trabalho em tela foram utilizadas técnicas de interpolação e limpeza de dados:

- **Transformação do arquivo de entrada:** O arquivo resultante da coleta de dados de telemetria é processado de forma que seja transformado em um arquivo denso, em que cada linha de dados tenha todas as colunas de informação preenchidas. Além disso, novas colunas são inseridas, derivadas de colunas de dados já existentes, como médias e desvios padrão de variáveis como velocidade e aceleração.
- **Retirada de *staypoints*:** O conjunto denso de dados tem milhões de entradas, algumas das quais não têm valor significativo para a análise de conforto, pois são idênticas a instâncias já analisadas. Para diminuir o conjunto de dados e minimizar a existência desse tipo de dado, foi utilizada uma técnica de retirada de *staypoints*, ou pontos de parada. Essa técnica foi aplicada sobre os dados geográficos do conjunto, retirando aqueles pontos que ocorrem muito próximos ao ponto imediatamente anterior em um curto espaço de tempo. Além de considerar a distância geográfica entre os pontos, a retirada considera também pontos sem flutuações significativas de qualquer eixo de aceleração. Dessa forma, o conjunto é reduzido sem que características importantes da aceleração sejam perdidas.

3.4 CRIAÇÃO DO CONJUNTO DE TREINAMENTO

Uma vez os dados formatados e limpos, o desafio a seguir é correlacionar a percepção de conforto dos passageiros e os dados da telemetria. O método proposto para esse cruzamento de dados se concentra na seleção dos dados compreendidos entre o momento de embarque e desembarque do passageiro no ônibus.

Os dados de percepção são coletados através do questionário, em conjunto com o local aproximado de embarque. O local de embarque pode ser o nome de uma via ou um ponto de referência conhecido na região. Como os dados de momento de embarque e desembarque do veículo são coletados por pessoas e são repostas aproximadas, com base na lembrança daquele momento que não teve grande significância na vida do passageiro, esses dados não são precisos.

Além disso, os passageiros não têm, necessariamente, a mesma percepção da viagem, ainda que tenham utilizado o mesmo ônibus, no mesmo percurso, dia e horário que outros passageiros. Dessa forma, um mesmo trecho da viagem pode ser

avaliado de forma diferente por passageiros que percorreram juntos o mesmo trajeto. Para que um modelo seja treinado, é importante que cada dado tenha apenas uma classificação. Portanto, para cada dado de telemetria coletado em um instante é necessário que apenas uma percepção de conforto esteja associada a ele.

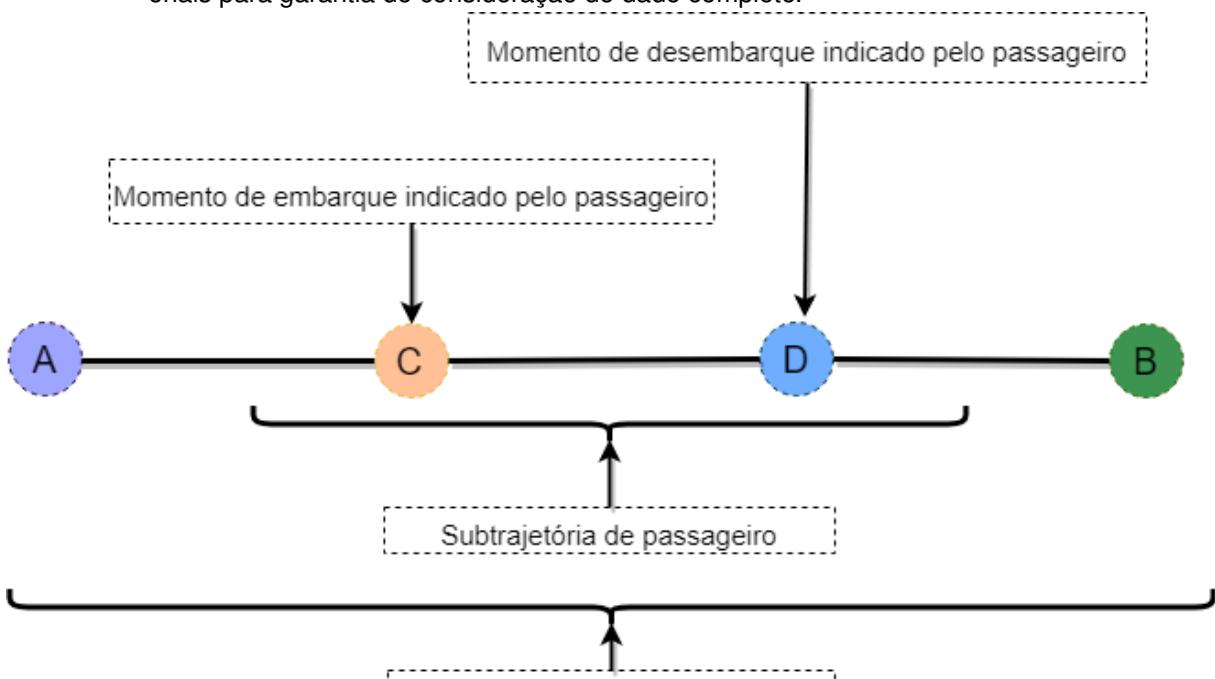
A solução as divergências de percepção de conforto, assim como a criação da base de conhecimento, são descritas nas próximas subseções.

3.4.1 Segmentação da Trajetória do Ônibus com Base na Viagem do Passageiro

Conforme já mencionado, a coleta da percepção considera o horário informado pelo passageiro como correspondente ao seu embarque. Entretanto, sabendo que essa informação é imprecisa devido a forma empírica pela qual ela é coletada, é necessário utilizar algum intervalo de tempo anterior ao horário informado pelo passageiro para garantir que toda a trajetória do passageiro seja considerada para a anotação.

Dessa forma, o método de anotação considera que a trajetória do veículo é subdividida em uma subtrajetória de passageiro no momento de embarque informado pelo passageiro subtraído de 10 minutos. A Figura 13 mostra uma representação gráfica da lógica de segmentação utilizada no trabalho em tela, em que A é o ponto de partida do veículo, B é o ponto final da linha do ônibus, C é o local de embarque do passageiro e o D o local de desembarque.

Figura 13 – Segmentação de trajetória em subtrajetórias de passageiro, com ilustração de limites adicionais para garantia de consideração de dado completo.

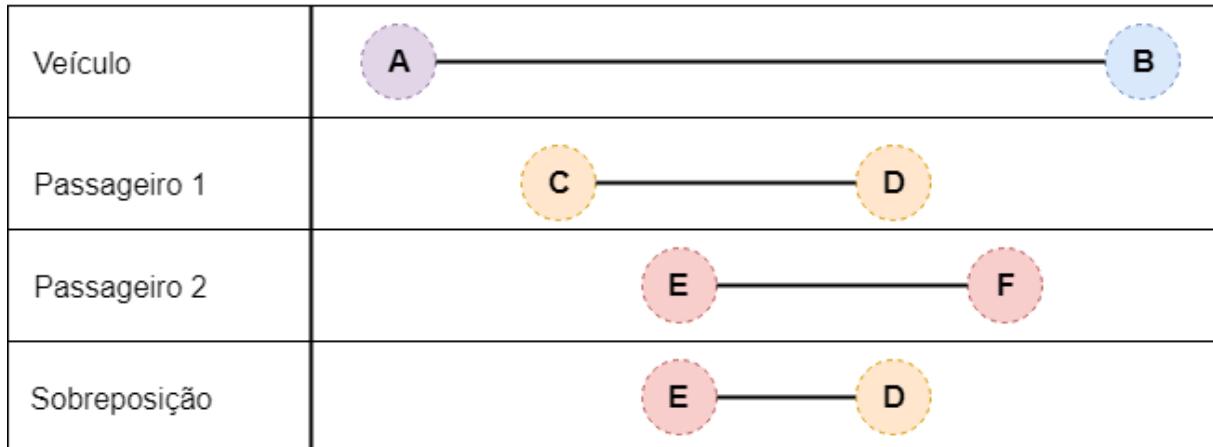


Fonte: Imagem gerada pela autora.

3.4.2 Consenso entre Percepções de Conforto em cada Período da Viagem

Cada subtrajetória de passageiro, por sua vez, pode se sobrepor a alguma outra subtrajetória de outro passageiro. A Figura 14 mostra uma representação da possível sobreposição entre duas subtrajetórias de passageiros.

Figura 14 – Representação de sobreposição de subtrajetórias de passageiros.



Fonte: Imagem gerada pela autora.

Conforme visto na Figura 14, é possível que haja sobreposição entre as viagens dos passageiros e, ainda, esses passageiros podem ter percepções diferentes de conforto. Entretanto, para garantir o bom desempenho do modelo de classificação, é importante que alguma técnica seja empregada para garantir que uma mesma instância de dados não seja classificada de duas ou mais formas distintas. Para tanto, utiliza-se o método do consenso por média. O consenso por média considera que a classificação correta para a instância é a média de todas as percepções coletadas para aquele período, com arredondamento para baixo.

Utilizando a Figura 14 como exemplo, se o Passageiro 1 classificou sua viagem como Muito Confortável, numericamente, sua percepção é 5. Se o Passageiro 2 classificou sua viagem como Razoavelmente Confortável, numericamente tem-se percepção 4. A média entre as duas classificações é 4.5, que, quando arredondada para baixo representa 4. Portanto, o período entre os pontos E e D seria classificado como 4, Razoavelmente Confortável, enquanto que o trecho entre C e E seria classificado pela percepção do Passageiro 1, e entre D e F pelo Passageiro 2.

3.4.3 Anotação da Base com a Percepção de Conforto do Passageiro

O maior desafio dessa etapa do trabalho está relacionado ao fato de que o passageiro é questionado sobre o conforto percebido apenas uma vez durante sua viagem, o que significa que não é possível saber qual momento da viagem teve maior influência em sua percepção de conforto. Para entender a percepção dos passageiros

as notas atribuídas pela escala de *likert* foram estratificadas por meio de uma distribuição normal, conforme apresentado na Tabela 5.

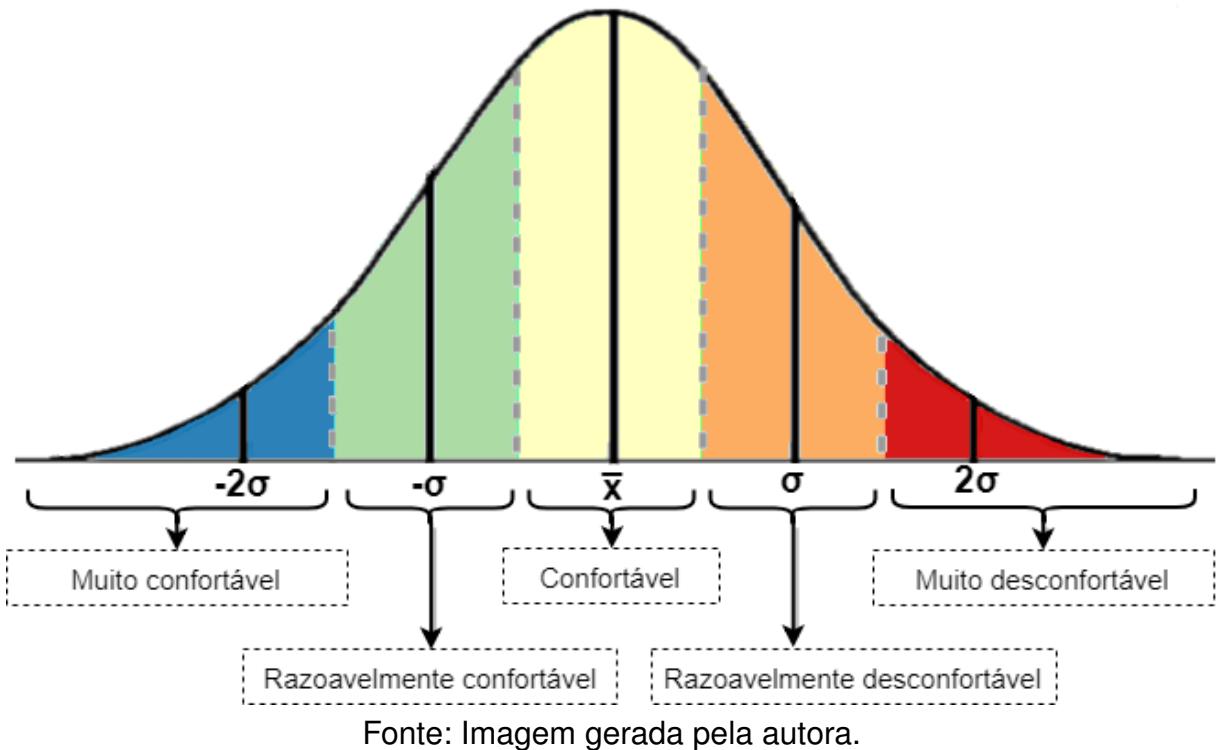
Para classificar adequadamente as viagens com relação à percepção de conforto do passageiro, uma abordagem de classificação com base na distribuição estatística dos dados foi proposta. As distribuições foram calculadas sobre as acelerações em cada um dos três eixos, pois são essas as características avaliadas pelos passageiros durante a aplicação do questionário.

O cálculo da distribuições e respectiva classificação da viagem é realizado da seguinte forma. Para cada passageiro, em seu respectivo segmento de trajetória, calcula-se a média e desvio padrão de cada eixo de aceleração. Esse cálculo é realizado com os dados em módulo, ou seja, as acelerações negativas são consideradas positivas. Dessa forma, todos os dados serão maiores que zero. Os dados são modelados como uma distribuição normal, de forma que a maior parte dos dados está alocada no centro da distribuição. Os dados que estão nessa região são considerados da classe Confortável. Essa classe se inicia na média subtraída de meio desvio padrão, $\bar{x} - \frac{\sigma}{2}$ e se encerra na média somada a meio desvio padrão, $\bar{x} + \frac{\sigma}{2}$. A utilização de $\frac{\sigma}{2}$ é uma decisão de projeto que tem por objetivo garantir que os valores que se aproximam da média por qualquer um dos lados ainda sejam considerados da mesma classe.

O mesmo raciocínio é aplicado às classes Razoavelmente Confortável e Razoavelmente Desconfortável, com a exceção de que essas classes utilizam o desvio padrão no lugar da média, ou seja, ficam contidas entre $\bar{x} - \sigma - \frac{\sigma}{2}$ e $\bar{x} - \sigma + \frac{\sigma}{2}$, e $\bar{x} + \sigma - \frac{\sigma}{2}$ e $\bar{x} + \sigma + \frac{\sigma}{2}$, respectivamente. Por fim, a classe Muito Confortável se inicia no menor dado existente na distribuição e se estende até $-2\sigma + \frac{\sigma}{2}$, enquanto que a classe Muito Desconfortável tem inicio em $+2\sigma - \frac{\sigma}{2}$ e se estende até o maior dado disponível naquele segmento. A Figura 15 mostra uma representação gráfica das regiões consideradas para cada classe.

Um conjunto de dados que contenha n linhas de coletas realizadas no período de viagem de um passageiro p , após a aplicação desse processo, resultará em conjuntos com m linhas, em que $m < n$, pois uma vez que os limites de cada classe são calculados somente as linhas que tenham dados dentro dos limites da classe correspondente à percepção de p serão anotadas e armazenadas no conjunto de saída.

Figura 15 – Anotação de percepção de passageiros nas trajetórias com base na distribuição de acelerações em cada eixo, classificadas por cores.



3.5 TREINAMENTO DOS MODELOS

Para a realização do treino dos modelos, a base de conhecimento é dividida em dois subconjuntos: o de treino e o de teste. Esses subconjuntos podem variar em volume, dependendo da natureza dos testes a serem realizados, mas uma vez que um subconjunto é utilizado para o treinamento de um modelo, ele é reutilizado para o treinamento dos outros modelos.

O treinamento dos submodelos é realizado de forma independente. Somente o treinamento do meta modelo depende do treinamento dos outros modelos. Os dados utilizados para o treinamento dos modelos são idênticos, com exceção do dado de anotação de cada instância, que varia de acordo com o modelo treinado.

3.5.1 Treinamento das árvores de decisão

De posse da base de conhecimento anotada, é possível partir para a inferência de modelos baseados em aprendizado de máquina. Dada a natureza do problema de classificação de conforto, sendo ele um problema de múltiplas variáveis, em particular no trabalho em tela, com ênfase nos dados de aceleração, optou-se pela utilização de diferentes modelos de classificação de dados, de forma que as características de cada conjunto de variáveis pudesse ser melhor considerado de acordo com suas particularidades.

Diferentes configurações de modelos foram analisadas até que o modelo atual seja apresentado. Dessa maneira, a primeira solução encontrada foi a utilização de apenas uma instâncias de cada classificador para a simplificação do modelo. No entanto essa solução não obteve resultados satisfatórios, uma vez que uma árvore de decisão anotada apenas com a classe de conforto geral e utilizando o conjunto completo de dados obteve apenas 56% de acurácia após diferentes tentativas.

Portanto são treinadas três árvores de decisão, uma para cada eixo de aceleração coletado pela telemetria e correspondentes às classificações coletadas por meio dos questionários. Cada árvore é treinada em um mesmo subconjunto da base de conhecimento, utilizando os mesmos dados, com exceção do dado de anotação de conforto.

Os modelos são nomeados de acordo com o eixo de aceleração utilizado para sua anotação, dessa forma, daqui para frente serão referidos como árvore x, referente ao eixo de aceleração x que é correspondente à aceleração frontal ou acelerar e frear do veículo; árvore y, referente ao eixo de aceleração y que é correspondente à aceleração lateral ou curvas; e árvore z, referente ao eixo de aceleração z que é correspondente à aceleração vertical ou trepidação.

3.5.2 Treinamento dos Modelos Ocultos de Markov - HMM

Também em razão da natureza do problema de classificação de conforto, optou-se pela utilização de Modelos Ocultos de Markov. Essa escolha foi feita com base na característica de descoberta de influências entre as diferentes variáveis de um mesmo conjunto de dados. Essa característica do HMM é útil para a inferência de influências não percebidas pelas árvores de decisão.

Da mesma forma que as árvores, os HMMs são treinados com o mesmo subconjunto da base de conhecimento, com a diferença de que para esses modelos não são utilizadas todas as variáveis disponíveis, apenas os dados relacionados com giroscópio, acelerações e percepção de conforto. A implementação foi realizada dessa forma devido à grande sensibilidade do HMM a cada um dos atributos. Essa sensibilidade pode ser explicada pelo fato de que um modelo multivariado é uma combinação de $n \times n$ atributos, em que cada dimensão representa uma variável sendo combinada com todas as outras. Utilizar mais variáveis implica na consideração de mais dimensões e, por consequência, em mais combinações. Para o conjunto de treinamento utilizado no trabalho em tela, a elevada quantidade de atributos dificultou a generalização dos modelos, ocasionando comportamento de *underfitting*. Entretanto, a medida que o conjunto de variáveis foi reduzido, reduziu-se proporcionalmente a sensibilidade do modelo, proporcionando os níveis de generalização desejado.

Para o treinamento de um HMM é necessário definir os seus estados observáveis e não observáveis. Para o trabalho em tela, similarmente à Figura 4, considera-se que os estados oculto são as possíveis classificações de conforto definidas na Tabela 5, que vão de 1 a 5, e cada instância do conjunto da base de conhecimento é um estado observado. Dessa maneira, o modelo será inferido com base nas observações obtidas pelos dados de sensores.

Uma vez de posse dos estados, deve-se inicializar a matriz de emissão. No caso de um modelo multivariado, conforme discutido na seção 2.3.3, é uma matriz de n atributos por n atributos, que deve ser utilizada para o treinamento do modelo e cálculo da probabilidade de transição de estados. Essa matriz é inicializada de forma genérica e é, então, treinada com o algoritmo de Viterbi (VITERBI, 2006).

Por fim, após o treinamento, o modelo pode ser utilizado para a previsão de novas instâncias. O modelo retorna a probabilidade de que uma instância pertença à uma classe, ou seja, uma instância terá a probabilidade de pertencer a cada uma das 5 classes, e é necessário escolher a classe com maior probabilidade de ocorrência. Como cada instância é composta por um conjunto de atributos, cada atributo recebe uma probabilidade, dessa maneira um simples cálculo da mediana das classes previstas tem a capacidade de retornar a classe com maior probabilidade.

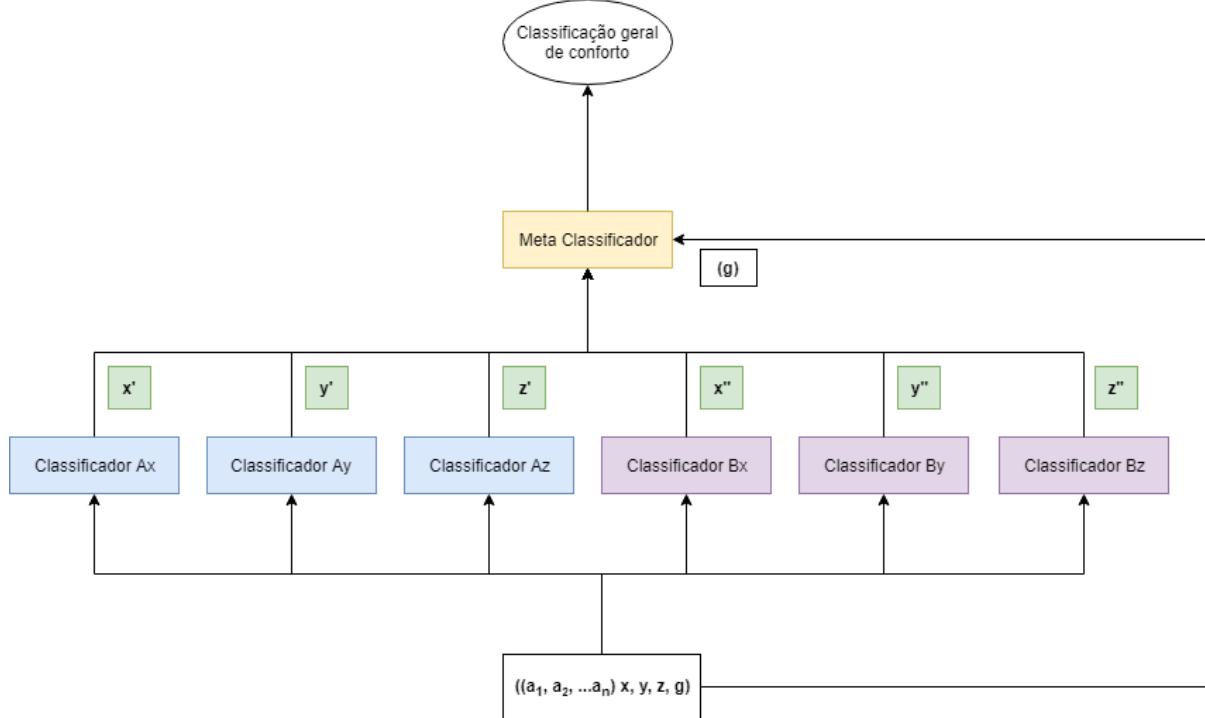
Semelhante às árvores de decisão, os três HMMs treinados são nomeados de acordo com o eixo de aceleração utilizado para a anotação do subconjunto da base de conhecimento. Portanto, os HMMs treinados são denominados da seguinte forma: HMM x, referente ao eixo de aceleração x, HMM y, referente ao eixo y e HMM z referente ao eixo z.

3.5.3 Treinamento do metamodelo

Uma vez treinados os modelos em cada um dos eixos, é necessário combiná-los para a classificação da percepção de conforto com relação à viagem como um todo, também referido aqui como conforto geral. Para tanto, selecionou-se o modelo de árvore de decisão, já utilizado para o treinamento especializado, uma vez que esse modelo apresentou generalização suficiente para ser considerado capaz de combinar as classificações realizadas pelo modelos especializados. A Figura 16 representa a combinação dos diferentes modelos em um único meta classificador, em que cada sub modelo tem como saída uma classificação de conforto para um eixo específico e o meta modelo os recebe como entrada, bem como a anotação de conforto geral.

Para o treinamento desse metamodelo uma nova base de conhecimento é criada, em que cada modelo anterior é utilizado para a previsão de uma classe de conforto para o eixo em questão, em cada uma das instâncias utilizadas para o trei-

Figura 16 – Representação de metamodelo de classificação.



Fonte: Imagem gerada pela autora.

namento dos modelos. Essas instâncias passam, nessa nova base de conhecimento, a serem anotadas com a nova classe prevista, de forma que cada instância tem 6 classes previstas pelos modelos e uma classe geral anotada com base na percepção dos passageiros. Essa base de conhecimento é, então, utilizada para o treinamento do metamodelo.

4 IMPLEMENTAÇÃO

O modelo foi posto em prática por meio de uma série de experimentos realizados de forma iterativa, com base na metodologia da Pesquisa-Ação. Os experimentos, bem como cada uma das etapas ilustradas pela Figura 12 são discutidos com detalhes de implementação, parametrização e resultados obtidos. Este capítulo cobre todas as etapas do método proposto, desde a obtenção e preparação dos dados cinéticos, seção 4.1, e de percepção, seção 4.2, cobrindo o pré-processamento dos dados, seção 4.3, passando pela construção da base de conhecimento, seção 4.4, seguindo para a seleção dos subconjuntos de treinamento dos modelos, seção 4.5, e finalizando com o treinamento de cada um dos 7 modelos de aprendizado de máquina, seções 4.6, 4.7 e 4.8.

4.1 COLETA DE DADOS DE SENSORES

A coleta de dados, resumida na seção 3.1, foi realizada por meio de um dispositivo Arduino, modelo Due, instalado em um veículo de transporte coletivo na cidade de Joinville. Conectados ao Arduino estavam um módulo ELM327, responsável pela coleta de dados do ODB do veículo por meio de uma conexão Bluetooth implementada por um módulo HC05, um módulo de GPS modelo GY-NEO6MV2, responsável pela coleta de dados espaço temporais e, por fim, um módulo MPU6050, para coleta de dados de acelerômetro e giroscópio.

Como o dispositivo não tinha conexão com a Internet, os dados eram persistidos *in loco* em um cartão de memória do tipo *MicroSD*. O cartão de memória foi escolhido por facilitar a coleta periódica dos dados por meio da substituição do cartão conectado ao dispositivo. O dispositivo foi instalado em um compartimento sobre o assento do motorista para facilitar o acesso à interface OBD e diminuir a interferência no sinal de GPS.

Os dados de ônibus foram obtidos por meio de um dispositivo instalado em um ônibus na cidade de Joinville. A instalação do dispositivo foi autorizada pela empresa concessionária do serviço na cidade e responsável pelo ônibus em questão. O dispositivo ficou instalado por cerca de 6 meses no veículo coletando os dados. Ele foi construído de forma modularizada. Sua arquitetura pode ser vista de forma detalhada em Bayer (2017).

As fontes de dados coletam e enviam os dados em momentos distintos e não coordenados. Portanto, é necessário que o arquivo de armazenamento dos dados tenha um formato que permita que todos os dados sejam gravados sem o risco de perda

da inter-relação entre as informações. Dessa forma, o dispositivo utiliza o formato TSV, que é um formato de armazenamento de texto tabular, onde cada dado é separado por um caractere de tabulação ('\t').

Como os dados são coletados em momentos distintos, o arquivo resultante é uma matriz esparsa de dados. Cada linha contém o registro de um evento e do instante de tempo em que ocorreu, no formato de *timestamp*. São considerados eventos a coleta de dados dos sensores e o recebimento de informação de funcionamento dos módulos. Cada dado provido pelos sensores é alocado em uma coluna específica, que nunca é utilizada para dados de outros sensores. Ainda, os intervalos de coleta são determinados pelo conhecimento *a priori* da taxa de modificação dos dados. Dessa forma, dados de acelerômetro, que são muito sensíveis, são coletados com maior frequência do que, por exemplo, os dados de localização GPS.

Para facilitar a geração da base de conhecimento utilizada para o treinamento dos modelos, é necessário criar um conjunto denso de dados a partir dos arquivos esparsos de dados. Dessa forma, para que o conjunto denso seja criado, os dados providos por sensores com menor frequência são interpolados para preencher as lacunas dos registros de maior frequência. Ainda, para sensores com dados muito sensíveis, ou seja, que variam rapidamente em pouco intervalo de tempo, são calculadas a média e o desvio padrão. Para os dados de acelerômetro utilizado no trabalho em tela foram utilizadas as médias das acelerações coletadas a cada minuto. Após esse tratamento, o conjunto de dados denso tem cerca de 1 milhão e meio de instâncias.

4.2 QUESTIONÁRIO

O questionário, descrito brevemente na seção 3.2, é composto por 7 perguntas classificadas em duas categorias. A primeira categoria é formada por 3 perguntas descriptivas, enquanto a segunda é formada por 4 perguntas objetivas na forma da escala de *likert*, onde o significado de cada índice da escala é apresentado na Tabela 5. Abaixo são listadas as perguntas:

1. Descriptivas: Qual foi o horário de embarque no veículo?
2. Descriptivas: Qual foi o local de embarque?
3. Descriptivas: Onde pretende desembarcar?
4. Objetiva: Qual é sua percepção de conforto em relação à trepidação (acelerações verticais)?
5. Objetiva: Qual é sua percepção de conforto em relação às curvas (acelerações laterais)?

6. Objetiva: Qual é sua percepção de conforto em relação às freadas e arrancadas (acelerações horizontais)?
7. Objetiva: Qual sua percepção de conforto em relação à viagem como um todo?

Tabela 5 – Classes de estratificação das notas de percepção de conforto comunicadas pelos passageiros durante a aplicação do questionário.

Nota	Descrição
1	Muito desconfortável
2	Razoavelmente desconfortável
3	Confortável
4	Razoavelmente confortável
5	Muito confortável

Fonte: Tabela gerada pela autora.

As perguntas relacionadas ao horário de embarque e local de desembarque têm como principal objetivo facilitar a segmentação da trajetória do veículo para anotá-la com a percepção do passageiro, ver seção 3.4.1. As demais perguntas são utilizadas para classificar a subtrajetória segmentada com a percepção de conforto do passageiro.

Foram obtidas 247 respostas de passageiros, durante um período de 9 dias, entre os dias 16 e 25 de setembro de 2017. As coletas de percepção foram realizadas em diferentes períodos do dia, entre as 06 horas da manhã e as 11 horas da noite. A Figura 17 apresenta as distribuições das respostas dos passageiros. As respostas se concentram, majoritariamente, na classificação 4 (razoavelmente confortável) para todas as acelerações.

Considerando que cerca de 120.000 pessoas utilizam o transporte coletivo em Joinville, uma amostra de 247 implica em um grau de confiança de 90%, com margem de erro de 5,2% para mais ou para menos. Como já discutido na seção 2.2.1, essa significância estatística exprime a percepção dos cidadãos da cidade de Joinville, e não devem ser tomadas como significativa para outras localidades.

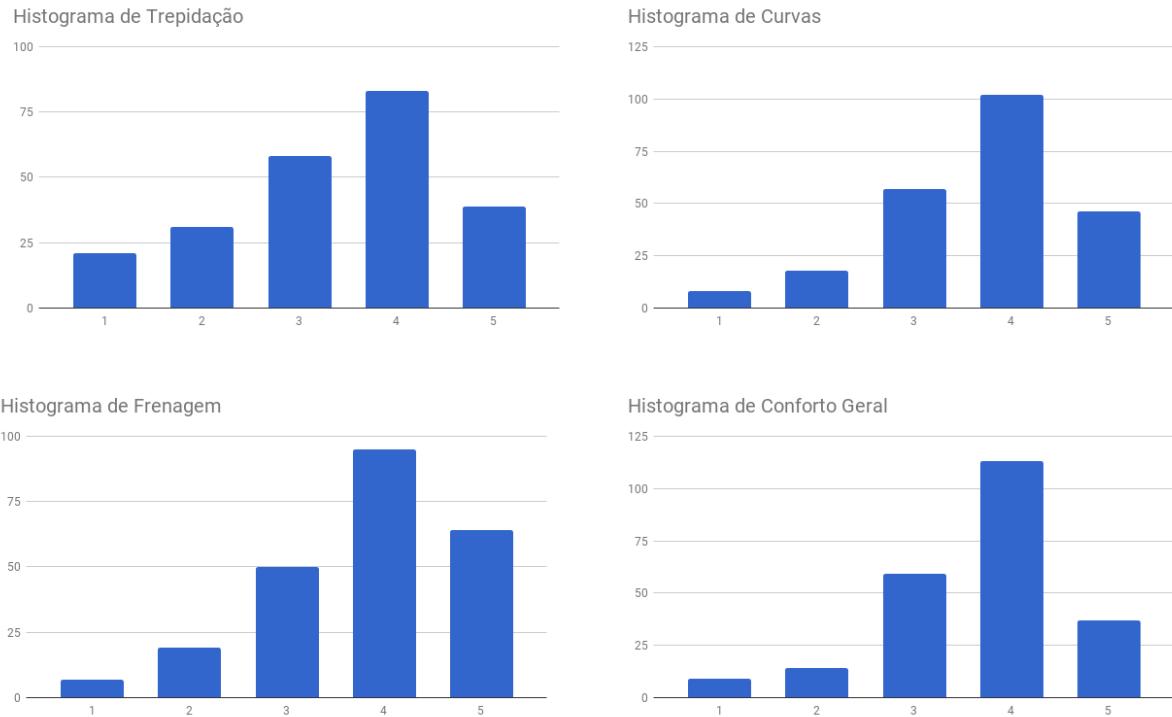
4.3 PRÉ-PROCESSAMENTO

Esta seção detalha a implementação da retirada dos pontos sem significância da trajetória (*staypoints*) e o consenso sobre a percepção de conforto entre os passageiros, apresentadas nas subseções 4.3.1 e 4.3.2, respectivamente.

4.3.1 Retirada de *staypoints*

A retirada de pontos de parada, ou *staypoints*, está relacionada à exclusão de pontos não significativos para o conjunto de dados, conforme discutido na seção

Figura 17 – Distribuição das respostas dos usuários com relação à percepção de conforto.



Fonte: Imagem gerada pela autora.

3.3. Portanto, pontos muito próximos a um ponto relevante e que não tenham alteração significativa nas acelerações são excluídos da trajetória, permanecendo apenas pontos que tenham valor informativo para a anotação do conforto e treinamento dos modelos.

Para realizar a retirada dos *staypoints*, o Algoritmo 1 recebe os seguintes parâmetros: i o índice da instância sendo avaliada, t o instante de tempo em que a instância ocorre, $a_{x,y,z}$ o conjunto de acelerações relacionado com a instância, d_{max} a distância máxima para que um ponto seja considerado um *staypoint*, t_{max} o intervalo de tempo máximo para que um ponto seja considerado um *staypoint*, e a_{max} o máximo de variação nas acelerações para que um ponto possa ser considerado um *staypoint*. Cada instância é comparada com a instância que ocorreu no momento anterior, de forma a verificar se as condições d_{max} , t_{max} e a_{max} são respeitadas.

Se qualquer uma dessas condições for violada, o ponto é válido e deve ser considerado para o treinamento dos modelos, caso contrário, ele é descartado. Por fim, o algoritmo retorna tanto a lista de pontos válidos quanto a de pontos descartados, de forma que seja possível analisar os dados ignorados e garantir que os limites parametrizados são suficientemente representativos.

A implementação desse algoritmo na situação do trabalho em tela considerou a distância máxima 5 metros, o período máximo de 4 segundos, e a variação má-

Algoritmo 1 Algoritmo para retirada de stay points.

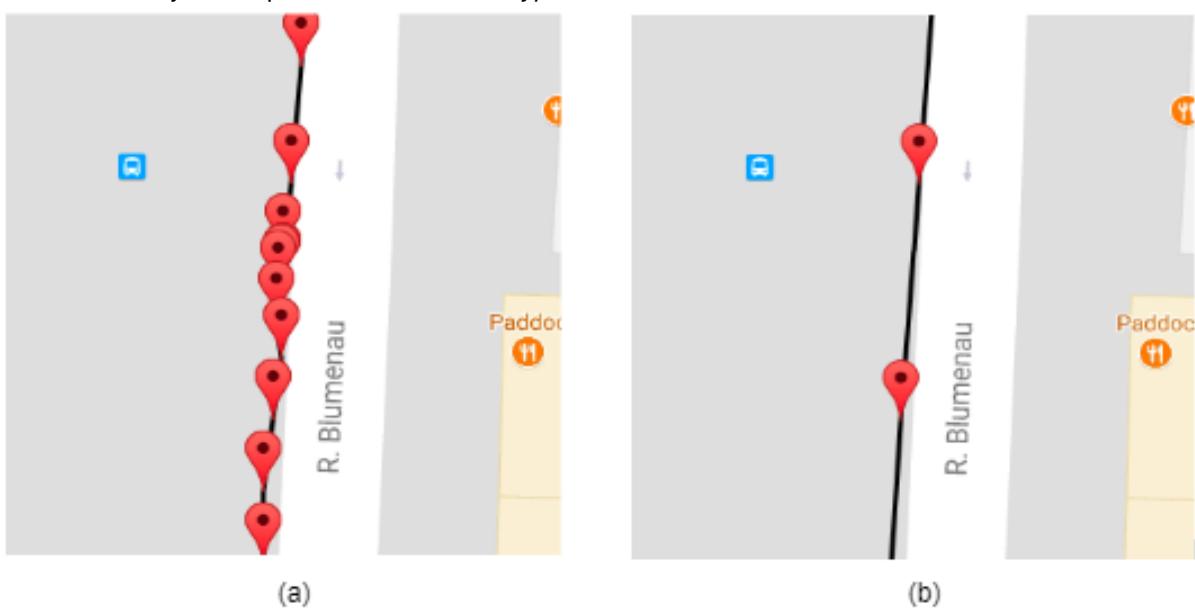
Entrada: $i, t, p_i, a_{x,y,z}, d_{max}, t_{max}, a_{max}$

- 1: **se** $distancia(p_i, p_{i-1}) < d_{max}$ and
 $periodo(t_i, t_{i-1}) < t_{max}$ and
 $a_{xi,yi,zi} - a_{xi-1,yi-1,zi-1} < a_{max}$ **então**
- 2: $staypoints \leftarrow i$
- 3: **senão**
- 4: $validos \leftarrow i$
- 5: **fim se**
- 6: **devolve** $validos, staypoints$

xima de 0.1g para todas as acelerações. Além disso, a distância entre os pontos foi calculada com base na função *vicenty*, do pacote *geopy* da linguagem Python. Os dados de tempo foram convertidos para o tipo data e hora da linguagem utilizada na implementação e calculados com as funções apropriadas provindas da biblioteca *datetime*, padrão da linguagem. Esses limites foram definidos com base na observação da base de dados coletados e em diversos experimentos baseados na metodologia Pesquisa-Ação.

A aplicação desse algoritmo no conjunto de dados resultou em um subconjunto de cerca de 700 mil instâncias de pontos significativos. Para ilustrar o resultado da aplicação do algoritmo, a Figura 18 apresenta um exemplo gráfico da eliminação dos *staypoints*, onde a Figura 18(a) apresenta a trajetória antes e a Figura 18(b) após a eliminação dos *staypoints*.

Figura 18 – Comparação de uma trajetória antes e após a retirada de *staypoints*. A Figura (a) representa uma trajetória completa, com todos os dados coletados. A Figura (b) representa a mesma trajetória após a retirada dos *staypoints*.



Fonte: Imagem gerada pela autora.

A Figura 19 representa uma trajetória que se inicia em um terminal de ônibus após a retirada de *staypoints*. Essa representação demonstra que nas situações em que existem arrancadas e freadas bruscas o método mantém os pontos significativos, mesmo estando relativamente próximos. O terminal de ônibus é um local em que as arrancadas e freadas bruscas ocorrem com frequência devido ao volume de pedestres que precisam realizar a travessia entre as plataformas de parada dos veículos.

Figura 19 – Trajetória de um veículo após a retirada de *staypoints*, mantendo os pontos de aceleração brusca comumente encontrados em terminais de ônibus.



Fonte: Imagem gerada pela autora.

4.3.2 Consenso de Percepções

Conforme detalhado na seção 3.4.2, é necessário garantir que cada instância de dados coletados tenha apenas uma classificação de percepção associada, de forma que durante o treinamento dos modelos de aprendizagem de máquina não exista ambiguidade em relação a qual classe cada instância pertença.

Para tanto, foi necessário implementar um algoritmo de consenso de percepções, considerando a sobreposição temporal e geográfica entre as coletas de percepção. O Algoritmo 2 descreve a forma de realização do consenso. Ele recebe como entrada um conjunto de percepções PE e identifica em quais períodos existe sobreposição de coleta de percepção, então, calcula o consenso nesse período.

O conjunto $PE = pe_1, pe_2, \dots, pe_n$, em que cada percepção é uma tupla $pe_i = (he_i, hd_i, le_i, ld_i, pt_i, pc_i, pa_i, pg_i)$, composta pelo horário de embarque he_i ; horário de desembarque hd_i ; percepção de conforto de trepidações pt_i ; percepção de conforto de curvas pc_i ; percepção de conforto de acelerações e frenagens pa_i ; e, finalmente, percepção de conforto geral pg_i .

O algoritmo percorre todas as instâncias em PE , verifica se existe sobreposição de coleta entre essas instâncias, anotando o momento de início e fim da sobre-

posição, uma vez que a sobreposição não tem necessariamente a mesma duração que as trajetórias analisadas. Portanto, as percepções são analisadas, os períodos de sobreposição são percorridos, e a média das percepções é calculada nos momentos de sobreposição, verificando qual trajetória finaliza primeiro. Para o trabalho em tela, somente um veículo foi utilizado para a coleta de dados cinéticos, portanto, não é necessário considerar os pontos geográficos, pois os momentos de tempo estão relacionados com uma única trajetória realizada pelo mesmo veículo.

Por fim, o algoritmo retorna a lista de percepções consensuadas. Além disso, é necessário manter os pontos que não estão no subconjunto de sobreposição. Após a aplicação do consenso, das 247 percepções coletadas foram obtidas 72 instâncias de percepções em consenso.

Algoritmo 2 Consenso de percepções

Entrada: PE

```

1: para  $i \leftarrow 0$  até  $Length(PE)$  faça
2:   para  $j \leftarrow i$  até  $Length(PE)$  faça
3:     enquanto  $he_j \geq he_i$  e  $he_j < hs_i$  e faça
4:        $overlap \leftarrow j$ 
5:     fim enquanto
6:   fim para
7:   para  $k \leftarrow 0$  até  $Length(overlap)$  faça
8:     se  $hs_k \geq hs_i$  então
9:        $c \leftarrow avg(pe_i, pe_k), he_i, hs_i$ 
10:      senão
11:         $c \leftarrow avg(pe_i, pe_k), he_i, hs_k$ 
12:      fim se
13:    fim para
14:  fim para
15:  devolve  $c$ 
```

4.4 CRIAÇÃO DE BASE DE CONHECIMENTO

De posse do conjunto de dados cinéticos pré-processados e o conjunto de percepções consensuado, é necessário realizar a combinação desses dados, conforme discutido na seção 3.4.3. Para tanto, utiliza-se o Algoritmo 3, que recebe como entrada as percepções P após o consenso e os conjuntos denso de dados de telemetria T .

Os dados de percepção são percorridos e, para cada tupla de percepção, os dados de telemetria ordenados pelo momento da coleta são avaliados, linhas 1 e 2. Se o dado de telemetria ocorreu no período da tupla de percepção, então ele é armazenado até que algum dado não esteja mais nesse período, linhas 3 a 5. A média e desvio padrão do conjunto é calculado, linhas 6 e 7, e todos os dados que estejam na área sob a curva de distribuição referente à percepção avaliada são anotados, linhas

9 a 38, conforme a Figura 15. O resultado do algoritmo de anotação é um conjunto de cerca de 15 mil instâncias.

A Tabela 6 apresenta a quantidade de registros de dados anotados por classe, baseado no cruzamento dos dados relacionado a qualificação do usuário por viagem realizada. Observando a Tabela 6, é possível perceber que algumas classes têm maior volume de dados classificados que outras. Tal resultado é esperado, uma vez que as respostas dos questionários, conforme Figura 17, apresentam o mesmo comportamento, pois os passageiros responderam mais perguntas com classes com conotação positiva de conforto.

Tabela 6 – Distribuição da anotação percepção de conforto geral por classes.

Classe	Quantidade	Percentual
ViagemMuitoConfortavel	3487	22,17%
ViagemRazoavelmenteConfortavel	3673	23,36%
ViagemConfortavel	3258	20,72%
ViagemRazoavelmenteDesconfortavel	3192	20,30%
ViagemMuitoDesconfortavel	2116	13,46%

Fonte: Tabela gerada pela autora.

4.5 SELEÇÃO DE SUBCONJUNTOS

Com a base de conhecimento criada, é necessário garantir que o treinamento dos modelos considere todas as classes de maneira equilibrada. Para tanto, foram selecionadas instâncias aleatórias de cada uma das classes, de forma que todas as classes tivessem a mesma quantidade de instâncias para o treinamento. Dessa maneira, cada classe representa 20% do conjunto de treinamento, com 2116 instâncias.

Além disso, nem todos os atributos disponíveis têm relevância para o treinamento do modelo. As árvores de decisão foram treinadas repetidamente retirando um atributo a cada treinamento, de forma a selecionar os atributos que apresentassem efetivamente relevância no treinamento dos modelos. Outra maneira possível de selecionar os atributos seria por meio da análise da associação linear entre as variáveis. Os atributos selecionados são apresentados na Tabela 7. Além desses atributos, também fazem parte do conjunto as anotações de conforto.

Para o treinamento dos HMMs, a utilização de todos os atributos apresentados na Tabela 7 resultou em modelos com baixa acurácia, dessa maneira o subconjunto de atributos de treinamento foi reduzido, considerando apenas os dados do IMU e derivados. É possível que esse comportamento seja resultado de um *underfitting* para esse tipo de modelo de aprendizado de máquina. Entretanto, mesmo utilizando um número reduzido de atributos, ainda assim foram utilizadas as mesmas instâncias do

Algoritmo 3 Anotação de percepções

Entrada: P, T

```

1: para  $i \leftarrow 0$  até  $\text{Length}(P)$  faca
2:   para  $j \leftarrow i$  até  $\text{Length}(T)$  faca
3:     enquanto  $he_i \geq \text{datetime}_j \wedge hs_i > \text{datetime}_j$  faca
4:        $periode \leftarrow j$ 
5:     fim enquanto
6:      $\bar{x} \leftarrow \text{avg}(periode)$ 
7:      $\sigma \leftarrow \text{std}(periode)$ 
8:     se  $p_i = 1$  então
9:       para  $j \leftarrow i$  até  $\text{Length}(periode)$  faca
10:      se  $ac_j \leq -1,5 * \sigma$  então
11:         $noted \leftarrow (j, 1)$ 
12:      fim se
13:    fim para
14:    senão se  $p_i = 2$  então
15:      para  $j \leftarrow i$  até  $\text{Length}(periode)$  faca
16:        se  $ac_j > -1,5 * \sigma \wedge ac_j \leq -\sigma/2$  então
17:           $noted \leftarrow (j, 2)$ 
18:        fim se
19:      fim para
20:    senão se  $p_i = 3$  então
21:      para  $j \leftarrow i$  até  $\text{Length}(periode)$  faca
22:        se  $ac_j > -\sigma/2 \wedge ac_j \leq \sigma/2$  então
23:           $noted \leftarrow (j, 3)$ 
24:        fim se
25:      fim para
26:    senão se  $p_i == 4$  então
27:      para  $j \leftarrow i$  até  $\text{Length}(periode)$  faca
28:        se  $ac_j > \sigma/2 \wedge ac_j \leq 1.5 * \sigma$  então
29:           $noted \leftarrow (j, 4)$ 
30:        fim se
31:      fim para
32:    senão se  $p_i == 5$  então
33:      para  $j \leftarrow i$  até  $\text{Length}(periode)$  faca
34:        se  $ac_j > 1.5 * \sigma$  então
35:           $noted \leftarrow (j, 5)$ 
36:        fim se
37:      fim para
38:    fim se
39:  fim para
40: fim para
41: devolve  $noted$ 

```

Tabela 7 – Atributos selecionados para o método de classificação proposto.

Origem	Atributo	Unidade	Domínio
OBD	Rotações do motor	RPM	Discreto
	Carga do motor	%	Discreto
	Pressão no coletor	kPa	Contínuo
	Pressão no pedal do acelerador	%	Discreto
GPS	Velocidade	km/h	Contínuo
IMU	Acelerômetro no eixo X	g	Contínuo
	Acelerômetro no eixo Y	g	Contínuo
	Acelerômetro no eixo Z	g	Contínuo
	Giroscópio no eixo X	°/s	Discreto
	Giroscópio no eixo Y	°/s	Discreto
	Giroscópio no eixo Z	°/s	Discreto
Derivados	Média acelerações	g	Contínuo
	Desvio padrão acelerações	g	Contínuo

Fonte: Tabela gerada pela autora.

conjunto de treinamento das árvores de decisão. Por fim, a base de dados é repartida em dois subconjuntos com o mesmo volume de instâncias, sendo um utilizado para o treinamento e o outro para o teste dos modelos.

4.6 TREINAMENTO DAS ÁRVORES DE DECISÃO

O treinamento das árvores de decisão ocorreu em ciclos incrementais de experimentos que foram evoluídos com base na observação dos resultados obtidos até que um resultado aceitável fosse alcançado. Esse é o raciocínio utilizado pelo método pesquisa-ação, utilizado como procedimento de pesquisa deste trabalho. Essa seção apresenta os resultados finais alcançados pelo treinamento das árvores de decisão utilizadas pelo metamodelo.

As árvores foram treinadas com o algoritmo J48, utilizando os seguintes parâmetros: *entropy* como o critério de divisão de ramos, *best* como a estratégia de escolha de cada ramo, 2 como o mínimo de amostras para a divisão de ramos e 1 como o mínimo de amostras para a criação de nós folha.

Todas as árvores foram treinadas com a mesma base de conhecimento, que representa metade da base de conhecimento completa. A única diferença entre os subconjuntos de treinamento é a classe anotada, que sempre será a classe relacionada com o eixo de decisão. Os valores de acurácia citados em cada uma das subseções é relativo com ao treinamento do modelo, a acurácia dos testes será discutida no próximo capítulo.

4.6.1 Árvore de decisão eixo x - acelerações e frenagens

Essa árvore foi treinada com 5290 instâncias, 60 vezes e obteve 80,3% de acurácia média, com 1,44 de desvio padrão. Para a análise dessa seção, um caso particular foi selecionado, que obteve de 80,4%. A matriz de confusão desse modelo é apresentada na Tabela 8, onde as colunas da tabela representam a quantidade de instâncias anotadas naquela classe e as linhas representam as classes nas quais elas deveriam ter sido anotadas.

Tabela 8 – Matriz de confusão da árvore de decisão do eixo x.

a	b	c	d	e	Classificado/ Esperado	Acurácia
861	82	54	35	26	a = MuitoConfortavel	81,38
58	768	119	68	45	b = RazoavelmenteConfortavel	72,59
7	27	976	12	36	c = Confortavel	92,25
18	37	46	926	31	d = RazoavelmenteDesconfortavel	87,52
59	64	79	134	722	e = MuitoDesconfortavel	68,24

Fonte: Tabela gerada pela autora.

A análise da diagonal principal da matriz apresentada na Tabela 8, bem como a coluna acurácia, denotam que a classe Confortavel eleva a média de acertos, enquanto que a classe MuitoDesconfortavel diminui a média de acertos do modelo como um todo. Além disso, a Tabela ilustra que o erro é bem distribuído entre todas as classes, porém em todas as classes existe erro na classe contrária, com exceção da classe Confortavel.

4.6.2 Árvore de decisão eixo y - curvas

Essa árvore foi treinada com 5290 instâncias, 60 vezes e obteve 77,7% de acurácia média, com 2,22 de desvio padrão. Para a análise dessa seção, um caso particular foi selecionado, que obteve de 77,2%. A matriz de confusão desse modelo é apresentada na Tabela 9, onde as colunas da tabela representam a quantidade de instâncias anotadas naquela classe e as linhas representam as classes nas quais elas deveriam ter sido anotadas.

Tabela 9 – Matriz de confusão da árvore de decisão do eixo y.

a	b	c	d	e	Classificado/ Esperado	Acurácia
892	74	51	29	12	a = MuitoConfortavel	84,31
24	52	21	934	27	b = RazoavelmenteConfortavel	88,28
33	98	786	83	58	c = Confortavel	74,29
116	660	101	87	94	d = RazoavelmenteDesconfortavel	62,38
31	46	71	99	811	e = MuitoDesconfortavel	76,65

Fonte: Tabela gerada pela autora.

Nesse modelo, a matriz de confusão, Tabela 9, tem duas classes com acerto acima de 80%, nas classes de conotação mais positiva para o passageiro. Além disso, duas classes, Confortavel e MuitoDesconfortavel, tem acurácia muito próxima à acurácia do modelo. A classe RazoavelmenteDesconfortavel, por sua vez, tem quase 15 pontos percentuais a menos do que o modelo como um todo.

O erro nessa árvore, na maior parte das classes, está nas classes imediatamente adjacentes, mas, novamente, nenhuma classe deixa de ter erro em classes extremas, como é o caso da classe MuitoConfortavel, que tem 7,23% do erro na classe MuitoDesconfortavel. Esses volumes não são particularmente representativos, no entanto demonstram que pode existir confusão com classes que não são próximas à classe analisada.

4.6.3 Árvore de decisão eixo z - trepidação

Assim como as anteriores, essa árvore também foi treinada com o mesmo conjunto de 5290 instâncias e obteve 79,56% de acurácia média, com 3 pontos de desvio padrão. Para a análise dessa seção, um caso particular foi selecionado, que obteve de 79%. A matriz de confusão desse modelo é apresentada na Tabela 10, onde as colunas da tabela representam a quantidade de instâncias anotadas naquela classe e as linhas representam as classes nas quais elas deveriam ter sido anotadas.

Tabela 10 – Matriz de confusão da árvore de decisão do eixo z.

a	b	c	d	e	Classificado/ Esperado	Acurácia
901	85	32	23	17	a = MuitoConfortavel	85%
84	763	91	72	48	b = RazoavelmenteConfortavel	72%
29	82	834	79	34	c = Confortavel	79%
19	58	138	701	142	d = RazoavelmenteDesconfortavel	66%
7	11	17	42	921	e = MuitoDesconfortavel	92,7%

Fonte: Tabela gerada pela autora.

A análise da diagonal principal da matriz apresentada na Tabela 10, bem como da coluna acurácia, demonstram que a classe MuitoDesconfortavel tem a maior acurácia do grupo, 13 pontos percentuais acima do modelo como um todo. Ainda, observa-se que a classe RazoavelmenteDesconfortavel tem quase 15 pontos percentuais abaixo do modelo, com erro bem distribuído entre as classes adjacentes, com 80% do erro dessa classes dividido entre elas.

4.7 TREINAMENTO DOS MODELOS OCULTOS DE MARKOV

O treinamento dos HMMs também ocorreu em ciclos incrementais, de tal forma que foi possível perceber que a utilização de todos os atributos utilizados no

treinamento das árvores de decisão não resultaria em modelos de Markov generalizáveis. Essa conclusão foi atingida pela análise dos modelos treinados repetidamente.

Esses treinamentos repetidos demonstraram que o modelo poderia sofrer de *underfitting*, uma vez que os resultados, sem qualquer alteração do conjunto de treinamento, variavam até 15 pontos percentuais com o mesmo conjunto de teste. Dessa maneira, a solução identificada foi a redução do conjunto de treinamento, que, quando retreinado repetidamente com o mesmo conjunto de treinamento teve resultados que variavam cerca de 3 pontos percentuais.

Ainda assim, os resultados de todos os HMMs são menos acurados que os resultados das árvores de decisão porém, conforme a própria literatura descreve, foi citado brevemente na seção 2.3.4, modelos com resultados muito próximos não necessitam de um metamodelo, uma vez que os seus resultados já representam a previsão para o conjunto.

Os HMMs foram construídos com base na biblioteca disponibilizada por Schreiber (2018), chamada *pomegranate*. Essa biblioteca implementa tanto o algoritmo de *Baum-Welch* quanto o *Viterbi*, ambos discutidos na seção 2.3.3. A previsão das instâncias é realizada por meio da mediana de classes previstas para cada um dos atributos na instância.

4.7.1 HMM eixo x - acelerações e frenagens

Esse HMM foi treinado com o mesmo conjunto de 5290 instâncias utilizado para o treinamento das árvores de decisão e obteve acurácia média de 67,5% após 60 treinos, além de 3 pontos de desvio padrão. Novamente, para a análise dessa seção um modelo foi selecionado, que apresenta 67,7% de acurácia. A matriz de confusão desse modelo é apresentada na Tabela 11, onde as colunas da tabela representam a quantidade de instâncias anotadas naquela classe e as linhas representam as classes nas quais elas deveriam ter sido anotadas.

Tabela 11 – Matriz de confusão do HMM do eixo x.

a	b	c	d	e	Classificado/ Esperado	Acurácia
681	219	73	56	29	a = MuitoConfortavel	64,3%
137	686	162	63	10	b = RazoavelmenteConfortavel	64,8%
52	106	765	121	14	c = Confortavel	72,3%
35	44	183	657	139	d = RazoavelmenteDesconfortavel	62%
0	33	45	186	794	e = MuitoDesconfortavel	75%

Fonte: Tabela gerada pela autora.

Essa matriz de confusão representa um comportamento recorrente nos HMMs, em que o erro fica agrupado quase que exclusivamente nas classes adjacentes à

classe esperada. Nesses modelos esse comportamento é ainda mais presente, uma vez que o acerto do modelo é significativamente menor do que o das árvores. Ainda, observa-se que o erro nas classes extremas é relativamente menor do que nas árvores de decisão.

4.7.2 HMM eixo y - curvas

Esse modelo também foi treinado com as mesmas 5290 instâncias e obteve acurácia média de 70% e 2,48 pontos de desvio padrão. O caso particular selecionado para essa seção obteve 70% de acurácia e a matriz de confusão desse modelo é apresentada na Tabela 12, onde as colunas da tabela representam a quantidade de instâncias anotadas naquela classe e as linhas representam as classes nas quais elas deveriam ter sido anotadas.

Tabela 12 – Matriz de confusão do HMM do eixo y.

a	b	c	d	e	Classificado/ Esperado	Acurácia
709	126	99	81	43	a = MuitoConfortavel	67%
138	717	164	38	1	b = RazoavelmenteConfortavel	67,8%
8	108	833	98	11	c = Confortavel	78,7%
34	39	84	782	119	d = RazoavelmenteDesconfortavel	73,9%
28	39	91	185	715	e = MuitoDesconfortavel	67,6%

Fonte: Tabela gerada pela autora.

A matriz de confusão, Tabela 12, representa um modelo bastante equilibrado, com 3 classes orbitando no 67% de acurácia. No entanto, observa-se que duas classes elevam a média do modelo como um todo, em particular a classe Confortavel, que tem quase 80% de acerto, e RazoavelmenteDesconfortavel.

Novamente, observa-se que o erro está agrupado nas classes adjacentes, mesmo nas classes com maior acerto. Além disso, esse modelo tem menos dificuldade com relação à classificação nas classes extremas, com poucas instâncias nessa situação.

4.7.3 HMM eixo z - trepidação

Assim como os modelos anteriores, esse modelo também foi treinado com o conjunto de 5290 instâncias e obteve acurácia média de 67,35%, com desvio padrão de 4,1 pontos. O modelo selecionado para essa seção obteve 66,4% de acurácia e a matriz de confusão desse modelo é apresentada na Tabela 13, onde as colunas da tabela representam a quantidade de instâncias anotadas naquela classe e as linhas representam as classes nas quais elas deveriam ter sido anotadas.

Tabela 13 – Matriz de confusão do HMM do eixo z.

a	b	c	d	e	Classificado/ Esperado	Acurácia
601	306	81	52	18	a = MuitoConfortavel	56,8%
193	632	176	25	32	b = RazoavelmenteConfortavel	59,7%
7	136	726	169	20	c = Confortavel	68,6%
26	6	75	862	89	d = RazoavelmenteDesconfortavel	81,5%
12	19	39	293	695	e = MuitoDesconfortavel	65,7%

Fonte: Tabela gerada pela autora.

Esse modelo apresenta o pior resultado dos modelos treinados para o trabalho em tela, com uma classe se aproximando de 50% de acerto, em que a quantidade de instâncias classificadas erroneamente na classe adjacente é igual à metade de instâncias classificadas corretamente. Além disso, uma classe tem mais de 80% de acerto, no mesmo modelo.

4.8 META MODELO DE CLASSIFICAÇÃO

Assim como os sub modelos, esse modelo também foi treinado com o conjunto de 5290 instâncias por 60 vezes e obteve acurácia média de 83,04%, com 1,54 ponto de desvio padrão. O modelo selecionado para essa seção obteve 83,8% de acurácia e foi treinado com as classificações saídas dos submodelos e com a anotação de conforto geral provinda da percepção dos passageiros. A matriz de confusão desse modelo é apresentada na Tabela 14, onde as colunas da tabela representam a quantidade de instâncias anotadas naquela classe e as linhas representam as classes nas quais elas deveriam ter sido anotadas.

Tabela 14 – Matriz de confusão do meta modelo.

a	b	c	d	e	Classificado/ Esperado	Acurácia
972	58	20	7	1	a = MuitoConfortavel	91,8%
92	863	75	12	16	b = RazoavelmenteConfortavel	81,5%
19	66	907	51	15	c = Confortavel	85,7%
24	85	116	799	34	d = RazoavelmenteDesconfortavel	75,5%
8	29	46	81	894	e = MuitoDesconfortavel	83,9%

Fonte: Tabela gerada pela autora.

Observa-se que no treinamento desse modelo a classe que mais contribuiu para aumentar a média do modelo como um todo foi a classe MuitoConfortavel, no entanto a classe com pior desempenho é a RazoavelmenteDesconfortavel, que tem 16 pontos percentuais abaixo do desempenho da melhor classe. Como a acurácia foi muito alta, o erro é marginal na maior parte das classes, no entanto, conforme observado em modelos anteriores, ele se distribui nas classes adjacentes e tem baixa distribuição nas classes extremas à classe esperada. Além disso, no treinamento, é

claro que o meta modelo tem desempenho superior a todos os modelos que o compõe, sugerindo que a utilização de um metamodelo em *stacking* faz sentido para esse contexto.

4.9 RESUMO DOS RESULTADOS DE TREINAMENTO

Os resultados obtidos na fase de treinamento são apresentados na Tabela 15. Eles foram gerados a partir de 60 execuções do processo de treinamento sobre o mesmo conjunto de instâncias de treinamento. Os valores de acurácia média, máxima, mínima e desvio padrão são apresentados em valores percentuais.

Tabela 15 – Resumo dos resultados de treinamento.

Modelo	Média	Desvio	Máxima	Mínima
Árvore X	80,3%	1,44%	82,3%	78,6%
Árvore Y	77,7%	2,22%	80,8%	74%
Árvore Z	79,5%	3%	83%	75,7%
HMM X	67,5%	3%	71,2%	63,3%
HMM Y	70%	2,4%	73,1%	67%
HMM Z	67,3%	4%	72,1%	62%
Meta modelo	83%	1,5%	84,8%	80%

Fonte: Tabela gerada pela autora.

Os resultados demonstram que, em média, o meta modelo tem desempenho superior ao dos sub modelos que o compõe. Além disso, apesar de combinar modelos com resultados consideravelmente distintos providos pelos sub modelos, ele apresenta um pequeno desvio padrão. Esse resultado sugere que o uso de um meta modelo é adequado para o contexto estudado, pois consegue melhor o resultado obtido pelos sub modelos sem aumentar o desvio padrão da acurácia. Além disso, observa-se que as árvores alcançaram um desempenho melhor que os HMMs, tanto nos valores de média quanto de mínima e máxima acurácia.

5 ANÁLISE DOS RESULTADOS

Para avaliar a implementação e aplicação do modelo proposto, neste capítulo são apresentadas as análises realizadas sobre o protótipo implementado a fim de avaliar seus resultados. Portanto, este capítulo começa apresentando o protocolo de experimentação, que é composto pelas regras gerais de avaliação dos resultados, bem como as métricas de avaliação. A seguir, os cenários de teste são apresentados e avaliados com relação a cada eixo de aceleração e conforto geral. Dessa forma, cada um dos modelos que compõe o metamodelo, inclusive o próprio metamodelo, são avaliados. Por fim, os resultados obtidos pelo modelo são discutidos e as suas limitações são apontadas.

5.1 PROTOCOLO DE EXPERIMENTAÇÃO

Alguns cenários foram selecionados para a realização da experimentação do modelo proposto no trabalho em tela. Esses cenários têm como principal objetivo demonstrar a robustez e a capacidade de generalização do modelo proposto por meio da classificação de viagens com características opostas. Os cenários de avaliação são caracterizados por uma dupla de trajetórias previamente selecionadas com trajetos e horários das viagens semelhantes, em que cada trajetória tenha uma classificação oposta à outra. Portanto, por exemplo, se uma trajetória t_i tem a classificação “MuitoConfortável” para algum eixo de aceleração, a outra trajetória t_j tem a classificação “MuitoDesconfortável” para o mesmo eixo de aceleração analisado.

Cada cenário tem o foco em um eixo de aceleração, ou seja, na avaliação do eixo de aceleração x, as trajetórias terão as classes “MuitoConfortável” ou “Muito-Desconfortável” para o eixo x e quaisquer outras classificações para os outros eixos e para a classificação geral. Esses cenários são baseados na coleta da percepção dos passageiros, não sendo essas trajetórias compostas por classificações consensuadas e sim pela percepção do passageiro entrevistado.

Com base no exposto acima, a avaliação dos resultados se baseia na análise das seguintes métricas:

- Análise do comportamento da curva ROC e da área sob a curva ROC, também chamada de AUC, entre os modelos treinados;
- Avaliação das matrizes de confusão dos modelos, bem como da acurácia para comparação dos pontos de generalização entre os modelos;

- Análise da diversidade entre os modelos, buscando identificar se eles são efetivamente complementares e, em conjunto, conseguem resultados melhores de classificação.

A área sob a curva é calculada com base na binarização das classes possíveis, de forma que todas as instâncias classificadas corretamente ficam no grupo Verdadeiro Positivo, e as classificadas erroneamente, independentemente da classificação atribuída pelo modelo, são consideradas Falso Positivo.

A diversidade entre os modelos é calculada com base na proposta de Skalak et al. (1996), que define que dois modelos são diversos quando classificam o mesmo conjunto de dados de maneira diferente. Essa proposta soma o erro de cada classe em cada dupla de classificadores e o divide pelo tamanho do conjunto analisado. A diversidade entre os dois modelos é calculada pela média da diversidade entre as classes.

Para o trabalho em tela, o erro do modelo como um todo é a média da diversidade de todas as classes. Dessa maneira, a dupla de modelos que classifica cada eixo de aceleração é considerada para o cálculo e as médias entre todas as duplas é considerada como a diversidade do modelo como um todo.

5.2 CENÁRIOS

Com base nos critérios definidos anteriormente, os dados foram selecionados e classificados pelos modelos gerados nas seções 4.6 e 4.7. As seções a seguir descrevem os resultados observados após aplicação de cada modelo.

5.2.1 Cenário 1 - Conforto relacionado com o eixo x

Para essa avaliação foram selecionadas duas trajetórias, no mesmo percurso, de cerca de 6,5 km percorridos, colhidas entre as 10 e 11 horas da manhã. Uma foi anotada como “MuitoConfortavel” e a outra como “MuitoDesconfortavel”. A primeira trajetória, anotada como “MuitoConfortavel” tem 987 instâncias, enquanto que a segunda trajetória tem 1056 instâncias.

A Figura 20 apresenta as curvas ROC de cada uma das trajetórias analisadas, em que a Figura 20a representa a ROC da trajetória anotada como “MuitoConfortavel” e a Figura 20b representa a ROC da trajetória anotada como “MuitoDesconfortavel”.

Com base na Figura 20, é possível observar que o desempenho dos modelos foi similar para as duas trajetórias, com os HMMs tendo desempenho ligeiramente inferior ao apresentado pelas árvores de decisão. Entretanto, ainda assim os HMMs ti-

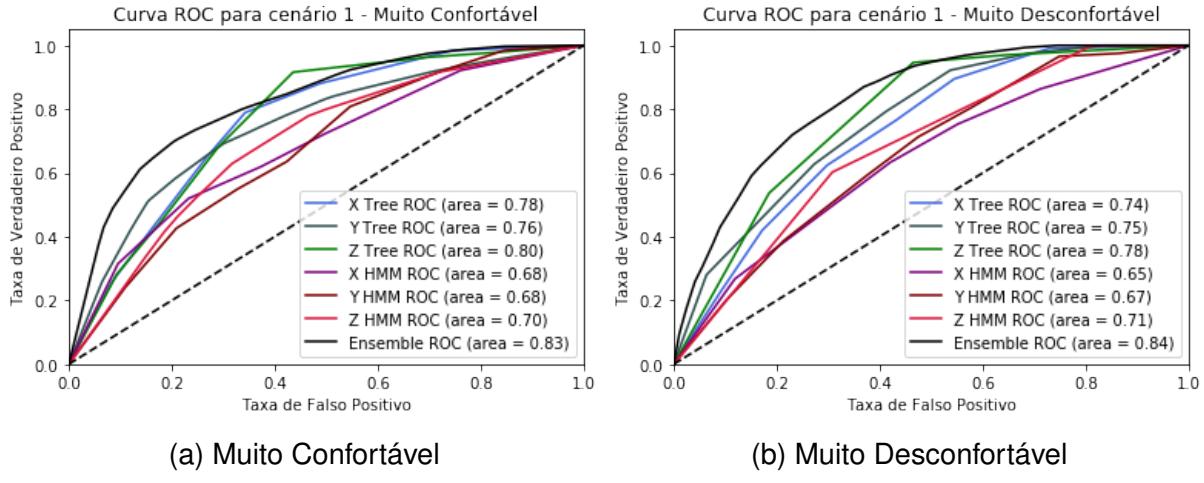


Figura 20 – Curvas ROCs do primeiro cenário de testes.

veram contribuição, pois influenciaram no desempenho do metamodelo, que teve AUC superior a todos os modelos para as duas trajetórias avaliadas.

5.2.1.1 Eixo x

A matriz de confusão da árvore de decisão do eixo x, representada pela Tabela 16, demonstra que apesar do erro ser equilibrado entre as classes, o erro na classe contrária da instância classificada como “MuitoConfortavel”, teve 38% do erro, enquanto que para a outra instância teve 26% de erro na classe contrária. Isso demonstra que o modelo tem alguma dificuldade de discernir entre duas classes diametralmente contrárias. Esse comportamento pode ser observado, em menor grau, no treinamento de modelo.

Tabela 16 – Matriz de confusão da árvore de decisão do eixo x do primeiro experimento

a	b	c	d	e	Classificado/ Esperado	Acurácia
760	67	35	39	86	a = MuitoConfortavel	77%
74	63	59	89	771	e = MuitoDesconfortavel	73%

Fonte: Tabela gerada pela autora.

Por sua vez, a matriz de confusão do HMM do eixo x, representada pela Tabela 17, demonstra que a maior parte do erro, para as duas trajetórias, está em uma classe adjacente à classe esperada. Isso demonstra que esse modelo tem mais dificuldade em discernir entre classe adjacentes. Dessa maneira, é possível inferir que esse modelo teve uma especialização maior que o gerado pela árvore de decisão.

Nesse cenário, o HMM teve para o eixo x uma média de acurácia de 68,4% para a classe “MuitoConfortavel” e de 64% para a classe “MuitoDesconfortavel”, que é um desempenho ligeiramente superior àquele alcançado pelo modelo no treinamento.

Tabela 17 – Matriz de confusão do HMM do eixo x do primeiro experimento

a	b	c	d	e	Classificado/ Esperado	Acurácia
676	125	82	68	36	a = MuitoConfortavel	68,4%
84	89	76	126	681	e = MuitoDesconfortavel	64%

Fonte: Tabela gerada pela autora.

Ainda que o resultado seja melhor que o do treinamento, ele é inferior ao resultado da árvore de decisão.

5.2.1.2 Eixo y

Com relação ao eixo y, a primeira trajetória foi anotada com a classe “RazoavelmenteConfortavel”, enquanto a segunda trajetória foi anotada com a classe “RazoavelmenteDesconfortavel”. Conforme foi discutido na introdução desse capítulo, é possível e esperado que as instâncias tenham anotações diferentes nos variados eixos de aceleração.

A Tabela 18 apresenta a matriz de confusão da árvore de decisão do eixo y. Essa tabela demonstra que a primeira trajetória tem a maior parte do erro, 36%, percebido em uma classe imediatamente adjacente à classe esperada. Comportamento similar é observado na segunda trajetória, com exceção de que boa parte do erro na trajetória está na classe com conotação mais positiva, “MuitoConfortavel”. Esse comportamento, por sua vez, pode demonstrar que os modelos de árvore podem confundir qualquer classe com as classes extremas. Isso pode contribuir para a classificação de instâncias como falsos positivos para a referida classe.

Tabela 18 – Matriz de confusão da árvore de decisão do eixo y do primeiro experimento

a	b	c	d	e	Classificado/ Esperado	Acurácia
89	740	53	48	57	b = RazoavelmenteConfortavel	74,9%
98	37	43	803	75	d = RazoavelmenteDesconfortavel	76,1%

Fonte: Tabela gerada pela autora.

Já a Tabela 19 representa a matriz de confusão do HMM do eixo y para esse cenário. Novamente, para esse tipo de modelo, a maior parte do erro está na classe imediatamente adjacente à classe esperada, com maior destaque para as classes extremas, tanto com conotação positiva quanto negativa. De modo geral, desconsideradas essas classes extremas, o erro é bem distribuído entre as classes restantes.

5.2.1.3 Eixo z

O eixo z foi anotado com a classe “MuitoConfortavel” na primeira trajetória e “RazoavelmenteDesconfortavel” na segunda trajetória.

Tabela 19 – Matriz de confusão do HMM do eixo y do primeiro experimento

a	b	c	d	e	Classificado/ Esperado	Acurácia
141	661	77	66	42	b = RazoavelmenteConfortavel	68,9%
6	65	96	702	187	d = RazoavelmenteDesconfortavel	66,5%

Fonte: Tabela gerada pela autora.

A matriz de confusão da árvore de decisão do eixo z, representada pela Tabela 20, demonstra comportamento similar à matriz da árvore de decisão do eixo y, com a maior parte do erro na classe adjacente, com 46% em média do erro nessa situação. Esse modelo tem o melhor desempenho entre as árvores de decisão, com 77% de acurácia. Além disso, é a primeira que tem melhor desempenho para a classe com conotação negativa, o que é incomum encontrar nos modelos dos outros cenários analisados.

Tabela 20 – Matriz de confusão da árvore de decisão do eixo z do primeiro experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
763	101	56	31	36	a = MuitoConfortavel	77,3%
35	34	50	824	113	d = RazoavelmenteDesconfortavel	78%

Fonte: Tabela gerada pela autora.

A Tabela 21 representa a matriz de confusão do HMM do eixo z. Essa matriz ilustra o melhor desempenho por parte dos modelos baseados no HMM, ainda que esse resultado seja, novamente, inferior ao modelo de árvore para as mesmas instâncias. Além disso, percebe-se um erro com volume maior que 25% do acerto na classe “RazoavelmenteDesconfortavel”. Esse resultado demonstra uma larga margem para erro nessa classe extrema, o que pode demonstrar que esse modelo é relativamente menos especializado que os outros.

Tabela 21 – Matriz de confusão do HMM do eixo z do primeiro experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
696	137	73	50	31	a = MuitoConfortavel	70,6%
15	28	57	755	201	d = RazoavelmenteDesconfortavel	71,5%

Fonte: Tabela gerada pela autora.

5.2.1.4 Metamodelo

As trajetórias para o conforto geral foram anotadas com as classes “MuitoConfortavel” e “RazoavelmenteDesconfortavel”.

A Tabela 22 representa a matriz de confusão do metamodelo do primeiro cenário. Para esse cenário, o metamodelo tem o erro bem distribuído em todas as classes, sem nenhuma classe se sobressaindo sobre as outras.

Tabela 22 – Matriz de confusão do metamodelo do primeiro experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
814	56	48	33	36	a = MuitoConfortavel	82,4%
44	29	38	882	53	d = RazoavelmenteDesconfortavel	83,7%

Fonte: Tabela gerada pela autora.

Como resultado da análise do metamodelo, é possível identificar que ele apresenta cerca de 25% de diversidade entre os modelos que o compõe. Essa diversidade é calculada com base na proposta descrita no protocolo de experimentação. Os modelos de classificação do eixo x têm diversidade de quase 30%, enquanto que os modelos de y têm apenas 21% e os de z têm cerca de 26%. Dessa maneira, percebe-se que a dupla de modelo que mais contribuiu para a diversidade do metamodelo foi a do eixo x.

5.2.1.5 Considerações sobre o primeiro cenário

De um modo geral, o metamodelo e os modelos que o compõe tiveram um desempenho adequado para a classificação das duas trajetórias. Todos os HMMs tiveram menor acurácia que as árvores de decisão. Os HMMs tiveram maior sucesso ao errar a classificação em classes próximas as classes esperadas. As árvores, por sua vez, tiveram, em média, 76% de acerto, 8 pontos percentuais acima da média de acertos dos HMMs. Ainda assim, o meta modelo teve desempenho 7 pontos percentuais a mais na acurácia quando comparado com as médias das árvores.

Esse resultado pode significar que os modelos são capazes de generalizar a classificação para instâncias classificadas em classes opostas. Além disso, o metamodelo teve desempenho superior a todos e a cada um dos modelos, demonstrando que a escolha de um *ensemble* foi adequada para o contexto no qual o modelo proposto nesse trabalho é aplicado.

5.2.2 Cenário 2 - Conforto relacionado com o eixo y

Para esse cenário foram selecionadas duas trajetórias, no mesmo percurso, de cerca de 3,5 km, colhidas entre as 13 e 14 horas da tarde. As anotações de cada trajetória são apresentadas na Tabela 23, em que a primeira coluna representa a trajetória a ser analisada, “X” representa a anotação do eixo x, “Y” a anotação do eixo y, “Z” a anotação do eixo z, “Geral” a anotação geral da trajetória e “Instâncias” a quantidade de instâncias que compõem cada trajetória. Além disso, as classificações foram resumidas, ou seja, “MC” significa “MuitoConfortavel”, “RC” “RazoavelmenteConfortavel”, “C” “Confortavel”, “RD” “RazoavelmenteDesconfortavel” e “MD” “MuitoDesconfortavel”.

A Figura 21 apresenta as curvas ROCs do segundo cenário analisado, que

Tabela 23 – Anotações das trajetórias do cenário dois relação aos eixos de aceleração e conforto geral.

Trajetória	X	Y	Z	Geral	Instâncias
1	MC	MC	RC	RC	1247
2	RD	MD	RD	C	1112

Fonte: Tabela gerada pela autora.

tem foco no eixo de aceleração Y, que representa a movimentação lateral do veículo. As curvas que representam os metamodelos, quando comparadas com a do primeiro cenário, apresentam maior distanciamento das curvas dos outros modelos, o que significa que para esses cenário os modelos tiveram desempenho inferior, enquanto que o desempenho do metamodelo é similar ao do cenário anterior.

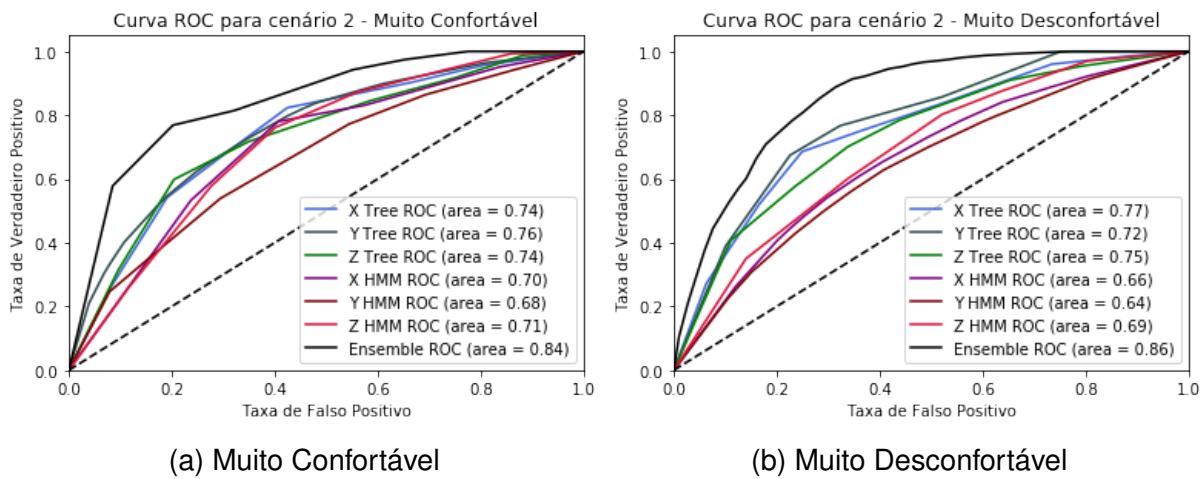


Figura 21 – Curvas ROCs do segundo cenário de testes.

5.2.2.1 Eixo x

A matriz de confusão da árvore de decisão do eixo x, representada pela Tabela 24, demonstra comportamento diferente para as duas trajetórias. Para a primeira, a maior parte do erro, 41%, está na classe oposta, enquanto que a segunda tem o erro muito equilibrado, variando entre 21 e 26% entre as classes. Esse comportamento do modelo na segunda trajetória pode ser interpretado como um alerta para a dificuldade de classificação de instâncias similares em uma mesma trajetória.

Tabela 24 – Matriz de confusão da árvore de decisão do eixo x do segundo experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
926	54	87	48	112	a = MuitoConfortavel	74,2%
68	55	55	856	67	d = RazoavelmenteDesconfortavel	76,9%

Fonte: Tabela gerada pela autora.

A matriz de confusão do HMM do eixo x, representada pela Tabela 25, demonstra que, mais uma vez, o HMM tem menor acurácia mas erra em classes mais

próximas às classes esperadas. No entanto, dessa vez o erro ocorre nas duas classes adjacentes à classe esperada de forma equilibrada.

Tabela 25 – Matriz de confusão do HMM do eixo x do segundo experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
878	127	95	102	45	a = MuitoConfortavel	70,4%
63	87	109	736	117	e = RazoavelmenteDesconfortavel	66,2%

Fonte: Tabela gerada pela autora.

5.2.2.2 Eixo y

A matriz de confusão da árvore de decisão do eixo y, representada pela Tabela 26, demonstra comportamento contrário à árvore do eixo x desse cenário, conforme a Tabela 24. Neste caso, a primeira instância teve erro equilibrado, com apenas 3 pontos percentuais a mais para a classe adjacente quando comparada com a próxima classe com maior distribuição de erro. Já a segunda instância, por sua vez, deslocou o erro majoritariamente para a classe oposta à classe esperada, com 66% do erro dividido entre as classes “MuitoConfortavel” e “RazoavelmenteConfortavel”.

Tabela 26 – Matriz de confusão da árvore de decisão do eixo y do segundo experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
952	88	56	79	72	a = MuitoConfortavel	76,4%
106	99	42	61	804	e = MuitoDesconfortavel	72,3%

Fonte: Tabela gerada pela autora.

A matriz de confusão do HMM do eixo y, representada pela Tabela 27, demonstra que novamente o HMM tem o erro distribuído nas classes próximas a classe esperada, para as duas trajetórias. Do ponto de vista da acurácia, o modelo tem desempenho razoavelmente inferior ao desempenho da árvore, com 66% de acurácia, em média, versus 74% em média para a árvore.

Tabela 27 – Matriz de confusão do HMM do eixo y do segundo experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
853	129	101	99	65	a = MuitoConfortavel	68,4%
75	78	94	150	715	e = MuitoDesconfortavel	64,3%

Fonte: Tabela gerada pela autora.

5.2.2.3 Eixo z

A matriz de confusão da árvore de decisão do eixo z, representada pela Tabela 28, evidencia que o erro ficou distribuído nas classes contíguas à classe esperada, com 63% do erro na primeira trajetória. No entanto, para a segunda trajetória o

erro foi bem distribuído, com duas classes recebendo 25% do erro, outra recebe 22% do erro e a última 28% do erro. Mais uma vez, percebe-se que, no mínimo em algumas situações, as árvores de decisão tem dificuldade de escolher a classe adequada, mesmo que marginalmente, ou, ainda, com maior agrupamento em alguma classe em particular.

Tabela 28 – Matriz de confusão da árvore de decisão do eixo z do segundo experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
96	921	112	65	53	b = RazoavelmenteConfortavel	73,4%
58	68	69	841	76	d = RazoavelmenteDesconfortavel	75,6%

Fonte: Tabela gerada pela autora.

A matriz de confusão do HMM do eixo z, representada pela Tabela 29, indica que para a primeira trajetória o HMM teve desempenho similar ao desempenho da árvore de decisão, com apenas 2 pontos percentuais de diferença entre eles. Além disso, nessa instância, mais de 45% do erro foi classificado em uma das classes adjacentes, enquanto que a outra classe adjacente admitiu apenas 25% do erro. Similarmente, a segunda trajetória teve o erro agrupado em uma das classes adjacentes de forma mais expressiva do que na outra, 43% contra 24%.

Tabela 29 – Matriz de confusão do HMM do eixo z do segundo experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
94	888	164	60	41	b = RazoavelmenteConfortavel	71,2%
46	66	84	764	152	d = RazoavelmenteDesconfortavel	68,7%

Fonte: Tabela gerada pela autora.

5.2.2.4 Metamodelo

A Tabela 30 representa a matriz de confusão do metamodelo do segundo cenário. Outra vez, o metamodelo tem erro bem distribuído entre as classes. Além disso, tem desempenho acima do desempenho dos modelos que o compõe, com 85% de acerto em média.

Tabela 30 – Matriz de confusão do meta modelo do segundo experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
58	1050	71	47	21	b = RazoavelmenteConfortavel	84,2%
22	42	956	55	37	c = Confortavel	86%

Fonte: Tabela gerada pela autora.

Além disso, esse metamodelo apresenta cerca de 29% de diversidade, composta por 28% da dupla do eixo x, 31% da dupla do eixo y e 28,5% da dupla do eixo z. Nesse cenário, o eixo y foi o que mais contribuiu para a diversidade, porém

apenas marginalmente, com apenas 3 pontos percentuais acima da diversidade dos outros eixos. Ainda assim, esse metamodelo apresenta maior diversidade do que o metamodelo do cenário anterior.

5.2.2.5 Considerações sobre o segundo cenário

De um modo geral o metamodelo e os modelos que o compõe tiveram um desempenho adequado para a classificação das duas trajetórias, com acurácia próxima as 70%. Novamente, o desempenho dos HMMs foi inferior ao das árvores de decisão. Além disso, em mais situações os modelos erram em classes mais próximas da classe esperada. Os HMMs tiveram, em média, 68% de acurácia, enquanto que as árvores tiveram, em média, 75% de acerto, 7 pontos percentuais de diferença entre os tipos de modelos. O metamodelo, por sua vez, teve média de acerto de 85%, 10 pontos percentuais acima das árvores.

Novamente, o metamodelo demonstrou que a combinação dos resultados de cada um dos modelos base tem capacidade de generalização das classificações opositas. Ainda, demonstram que no contexto do uso em mundo real, a classificação tem grande chance de estar correta, mas que na ocorrência de um erro, ele tem maior chance de ser em uma classe adjacente. Esse não é um comportamento desejável, pois dificulta a identificação de falsos positivos para as outras classes.

Portanto, os resultados desse cenário reforçam que o uso do metamodelo foi o mais adequado para esse contexto, uma vez que ele foi significativamente melhor na acurácia do que os modelos que o compõe.

5.2.3 Cenário 3 - Conforto relacionado com o eixo z

Para esse cenário foram selecionadas duas trajetórias, no mesmo percurso, de cerca de 5,5 km, colhidas entre as 11 da manhã e 13 horas da tarde. As anotações de cada trajetória são apresentadas na Tabela 31, em que a primeira coluna representa a trajetória a ser analisada, “X” representa a anotação do eixo x, “Y” a anotação do eixo y, “Z” a anotação do eixo z, “Geral” a anotação geral da trajetória e “Instâncias” a quantidade de instâncias que compõe cada trajetória. Além disso, as classificações foram resumidas, ou seja, “MC” significa “MuitoConfortavel”, “RC” “RazoavelmenteConfortavel”, “C” “Confortavel”, “RD” “RazoavelmenteDesconfortavel” e “MD” “MuitoDesconfortavel”.

A Figura 22 apresenta as curvas ROCs do terceiro experimento, que tem foco no eixo de aceleração Z, que representa a movimentação horizontal do veículo. Pela primeira vez, observa-se que todas as curvas de todos os modelos estão muito próximas, ou seja, todos os desempenhos foram similares, ainda que novamente os HMMs

Tabela 31 – Anotações das trajetórias do cenário três com relação aos eixos de aceleração e conforto geral.

Trajetória	X	Y	Z	Geral	Instâncias
1	MC	MC	MC	MC	1187
2	RD	C	MD	C	1338

Fonte: Tabela gerada pela autora.

tiveram desempenho inferior ao outros. Além disso, é a primeira vez que a curva de um metamodelo demonstra um desempenho praticamente igual ao dos modelos que o compõe.

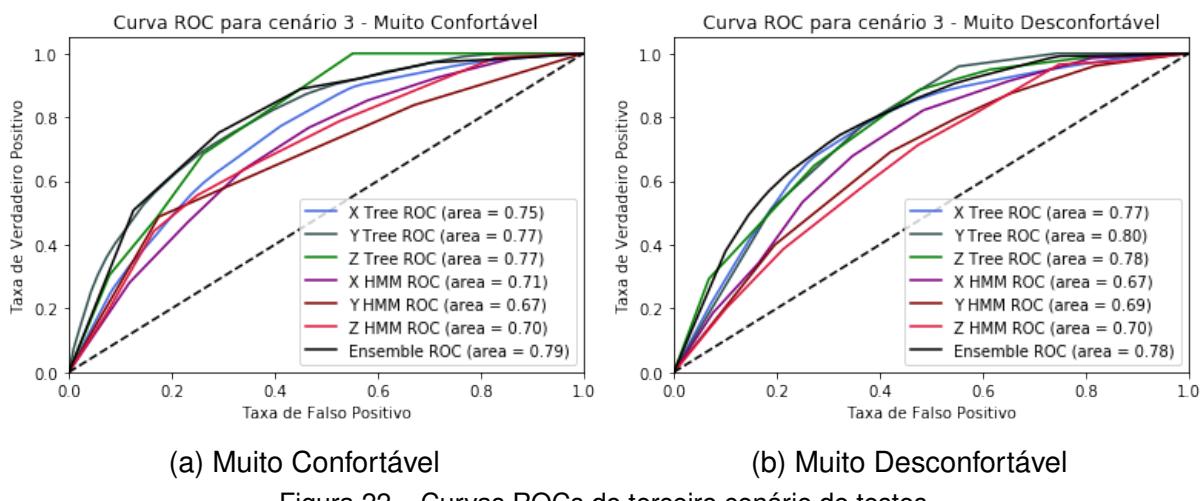


Figura 22 – Curvas ROCs do terceiro cenário de testes.

5.2.3.1 Eixo x

A matriz de confusão da árvore de decisão do eixo x, representada pela Tabela 32, aponta que para esse cenário, na primeira trajetória, o modelo erra com maior frequência na classe “Confortavel”, que não é adjacente à classe esperada, com 39% do erro. Por sua vez, a segunda trajetória tem o erro bem distribuído entre as classes, cuja ocorrência é ligeiramente maior nas classes adjacentes à classe esperada, com 27,6% para a classe “Confortavel” e 30% para a classe “MuitoDesconfortavel”.

Tabela 32 – Matriz de confusão da árvore de decisão o eixo x do terceiro cenário.

a	b	c	d	e	Classificado/ Esperado	Acurácia
893	76	115	58	45	a = MuitoConfortavel	75,2%
69	62	85	1030	92	d = RazoavelmenteDesconfortavel	77%

Fonte: Tabela gerada pela autora.

A matriz de confusão do HMM do eixo x, representada pela Tabela 33, destaca uma frequência significativa do erro para a classe adjacente, com 51,7% do erro,

por outro lado apenas 6% do erro é apontado na classe “MuitoDesconfortavel”. Na segunda trajetória o erro é novamente distribuído entre as classes adjacentes à classe esperada, com destaque para a classe “Confortavel”, com 35% do erro, 9 pontos percentuais acima da próxima classe com maior frequência de erro, a classe “MuitoDesconfortavel”.

Tabela 33 – Matriz de confusão do HMM do eixo x do terceiro experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
837	181	84	62	23	a = MuitoConfortavel	70,5%
78	89	157	896	118	d = RazoavelmenteDesconfortavel	66,9%

Fonte: Tabela gerada pela autora.

5.2.3.2 Eixo y

A matriz de confusão da árvore de decisão do eixo y, representada pela Tabela 34, segue o mesmo comportamento do primeiro modelo aplicado nesse cenário. O erro para as duas trajetórias está distribuído com maior frequência nas classes adjacentes às classes esperadas, enquanto que o erro nas classes opostas, ou extremas no sentido da percepção, apresentam frequência pouco significativa de erros, em particular quando comparadas com as classes de maior frequência.

Tabela 34 – Matriz de confusão da árvore de decisão o eixo y do terceiro cenário.

a	b	c	d	e	Classificado/ Esperado	Acurácia
918	98	77	54	40	a = MuitoConfortavel	77,3%
57	85	1068	76	52	c = Confortavel	79,8%

Fonte: Tabela gerada pela autora.

A matriz de confusão do HMM do eixo y, representada pela Tabela 35, teve desempenho significativamente abaixo do desempenho da árvore de decisão para esse cenário, com acurácia menor em 10 pontos percentuais, em média. Além disso, o modelo apresenta a maior dificuldade até o momento de distinguir entre a classe esperada e as classes próximas, ou adjacentes, a ela, com 43% do erro na classe adjacente à esperada, enquanto que a segunda instância tem 70% do erro nas duas classes adjacentes.

Tabela 35 – Matriz de confusão do HMM do eixo y do terceiro experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
790	172	89	76	60	a = MuitoConfortavel	66,5%
57	151	928	136	66	c = Confortavel	69,4%

Fonte: Tabela gerada pela autora.

5.2.3.3 Eixo z

A matriz de confusão da árvore de decisão do eixo z, representada pela Tabela 36, apresenta as melhores acurácia obtidas nesse cenário, com cerca de 77% em média. A primeira trajetória tem 46% do erro na classe adjacente e bem distribuído nas classes restantes, conforme já observado em outras aplicações desse modelo. A segunda trajetória tem comportamento ligeiramente diverso da primeira, com todo o erro bem distribuído entre as classes.

Tabela 36 – Matriz de confusão da árvore de decisão o eixo z do terceiro cenário.

a	b	c	d	e	Classificado/ Esperado	Acurácia
911	127	61	36	52	a = MuitoConfortavel	76,7%
54	81	73	84	1046	e = MuitoDesconfortavel	78,2%

Fonte: Tabela gerada pela autora.

A matriz de confusão do HMM do eixo z, representada pela Tabela 37, aponta resultados consistentes com os cenários apresentados anteriormente, com 7 pontos percentuais abaixo da acurácia da árvore de decisão, em média. A primeira instância tem 67% do erro dividido entre a classe adjacente e a classe “Confortavel”. Enquanto que a segunda instância tem 47% do erro presente na classe adjacente à classe esperada. Esse comportamento também é consistente com o comportamento desse modelo nos cenários apresentados anteriormente.

Tabela 37 – Matriz de confusão do HMM do eixo z do terceiro experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
827	124	116	77	43	a = MuitoConfortavel	69,7%
75	51	87	186	942	e = MuitoDesconfortavel	70,4%

Fonte: Tabela gerada pela autora.

5.2.3.4 Metamodelo

A matriz de confusão do metamodelo do terceiro cenário, representada pela Tabela 38, aponta o pior desempenho do metamodelo nos cenários apresentados. Além disso, ele apresenta comportamento similar ou pior do que o dos modelos que o compõe.

Assim como nos modelos que o integram, o metamodelo erra com maior frequência nas classes adjacentes às classes esperadas. Isso demonstra que todos os modelos, de um modo geral, sofrem do mesmo problema que é distinguir entre classes próximas. Portanto, percebe-se aqui um problema de baixa especialização do metamodelo.

Tabela 38 – Matriz de confusão do meta modelo do terceiro cenário.

a	b	c	d	e	Classificado/ Esperado	Acurácia
941	86	71	49	40	a = MuitoConfortavel	79,3%
47	90	1044	92	65	c = Confortavel	78%

Fonte: Tabela gerada pela autora.

Esse modelo apresenta diversidade de 23%, em média. Essa diversidade é composta por 25% de diversidade na dupla do eixo x, 24% de diversidade na dupla do eixo y e 21% de diversidade na dupla do eixo z. Observa-se que esse modelo tem diversidade similar a do metamodelo apresentado no cenário 1, seção 5.2.1, no entanto, naquele cenário um dos modelos contribuía com quase 30% de diversidade, enquanto que nesse modelo as diversidades são próximas de 20%.

5.2.3.5 Considerações sobre o terceiro cenário

Esse cenário apresenta um metamodelo com desempenho aquém do esperado, pois o metamodelo não tem, nesse caso, maior acurácia que os modelos que o compõe. Esse resultado pode ser justificado pela baixa diversidade entre os modelos, como pode ser observado nas curvas ROC desse cenário, apresentadas na Figura 22. O fato de que todos os modelos são muito similares oferece menos subsídios para que o metamodelo tenha a capacidade de discernir entre as classes qual é a correta. Todos os modelos tiveram acurácia acima de 65%, com melhor desempenho para a classificação das trajetórias anotada como “MuitoDesconfortavel” para o eixo z.

Ainda sobre a avaliação do eixo z, para as duas trajetórias os modelos obtiveram seus melhores resultados. Talvez esse comportamento seja advindo do fato de que a anotação de conforto se baseia preponderantemente nas arrancadas e freadas que o passageiro sentiu, o que parece ser uma sensação mais sutil e subjetiva de se interpretar, fazendo com que os modelos tenham maior dificuldade em diferenciarem entre as classes. Isso pode explicar porque os modelos erram tanto em classes adjacentes à esperada. Por outro lado, isso evidencia que eles têm, no mínimo, capacidade de errar próximo à classe esperada.

Finalmente, esse resultado não implica que necessariamente o metamodelo é uma solução ruim para a classificação dos dados desse contexto, mas que para a tomada de decisão baseada nos dados previstos pelo metamodelo é importante considerar a diversidade entre os modelos que o compõe. Talvez a aplicação de técnicas de aumento de diversidade, como a divisão de conjuntos de treino para cada modelo, como é feito nas técnicas de *bagging* e *boosting*, apresentadas na seção 2.3.4, em conjunto com o *ensemble*, possa apresentar melhores resultados.

5.2.4 Cenário 4 - Conforto geral

Duas trajetórias foram selecionadas para esse cenário, ocorridas no mesmo percurso, de cerca de 6 km, colhidas entre as 10 e 12 horas da manhã. As anotações de cada trajetória são apresentadas na Tabela 39, em que a primeira coluna representa a trajetória a ser analisada, “X” representa a anotação do eixo x, “Y” o eixo y, “Z” o eixo z, “Geral” a anotação de conforto geral e “Instâncias” a quantidade de instâncias em cada trajetória. Além disso, as classificações foram resumidas, em que “MC” significa “MuitoConfortavel”, “RC” “RazoavelmenteConfortavel” e “RD” “RazoavelmenteDesconfortavel”.

Tabela 39 – Anotações das trajetórias do cenário quatro com relação aos eixos de aceleração e conforto geral.

Trajetória	X	Y	Z	Geral	Instâncias
1	RC	MC	MC	MC	1469
2	RD	RD	RD	RD	1351

Fonte: Tabela gerada pela autora.

A Figura 23 apresenta as curvas ROCs do quarto experimento, que foca no conforto geral, que é composto pelos confortos colhidos para os 3 eixos de aceleração. A curva do metamodelo apresenta visível descolamento das curvas de outros modelos, enquanto que as curvas dos modelos apresentam similaridade de comportamento por tipo de modelo (árvore ou HMM), em que as árvores de decisão têm curvas próximas e os HMMs tem o mesmo comportamento, com as 3 curvas de cada trajetória próximas entre si no plano cartesiano.

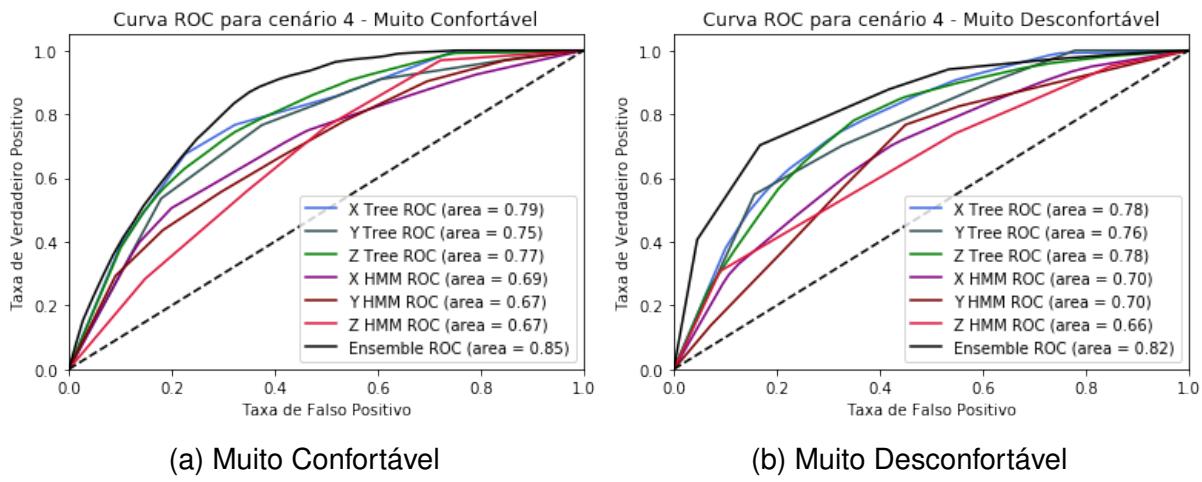


Figura 23 – Curvas ROCs do quarto cenário de testes.

5.2.4.1 Eixo x

A matriz de confusão da árvore de decisão do eixo x, representada pela Tabela 40, apresenta um desempenho muito similar entre as duas trajetórias, com apenas

0,8 ponto percentual de diferença. Além disso, na duas trajetórias têm, novamente, dificuldade de diferenciar entre as classes adjacentes à classe esperada.

Tabela 40 – Matriz de confusão da árvore de decisão o eixo x do quarto cenário.

a	b	c	d	e	Classificado/ Esperado	Acurácia
92	1158	106	62	51	b = RazoavelmenteConfortavel	78,8%
31	48	84	1053	135	d = RazoavelmenteDesconfortavel	78%

Fonte: Tabela gerada pela autora.

A matriz de confusão do HMM do eixo x, representada pela Tabela 41, evidencia um desempenho bastante inferior do que a árvore de decisão desse eixo, com cerca de 10 pontos percentuais de diferença entre os resultados. Novamente, o erro predomina nas classes adjacentes à classe esperada, com 64% do erro agrupado nas classes próximas nas duas trajetórias.

Tabela 41 – Matriz de confusão do HMM do eixo x do quarto experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
1007	185	114	85	78	a = MuitoConfortavel	68,6%
66	81	134	951	119	d = RazoavelmenteDesconfortavel	70,4%

Fonte: Tabela gerada pela autora.

5.2.4.2 Eixo y

A matriz de confusão da árvore de decisão do eixo y, representada pela Tabela 42, apresenta, novamente, um comportamento muito similar entre as trajetórias. O erro é distribuído majoritariamente nas classes próximas às classes esperadas.

Tabela 42 – Matriz de confusão da árvore de decisão o eixo y do quarto cenário.

a	b	c	d	e	Classificado/ Esperado	Acurácia
1104	132	88	76	69	a = MuitoConfortavel	75,1%
14	68	146	1026	97	d = RazoavelmenteDesconfortavel	76%

Fonte: Tabela gerada pela autora.

A matriz de confusão do HMM do eixo y, representada pela Tabela 43, também apresenta a maior parte do erro nas classes adjacentes. No entanto, aqui a frequência é ainda mais pronunciada, com 57% do erro da primeira trajetória atribuído à classe vizinha à classe esperada. A segunda trajetória apresenta comportamento similar, mas com 72% do erro dividido nas duas classes adjacentes à esperada.

5.2.4.3 Eixo z

A matriz de confusão da árvore de decisão do eixo z, representada pela Tabela 44, expõe outro modelo com o mesmo comportamento apresentado nos outros

Tabela 43 – Matriz de confusão do HMM do eixo y do quarto experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
988	276	105	78	22	a = MuitoConfortavel	67,2%
33	77	156	951	134	d = RazoavelmenteDesconfortavel	70,4%

Fonte: Tabela gerada pela autora.

modelos desse cenário. A diferença no desempenho entre as trajetórias analisadas é apenas marginal, apenas 0,4 pontos percentuais de diferença. Além disso, os modelos têm desempenho adequado, próximos a 77,5% de acurácia.

Tabela 44 – Matriz de confusão da árvore de decisão o eixo z do quarto cenário.

a	b	c	d	e	Classificado/ Esperado	Acurácia
1135	143	85	57	49	a = MuitoConfortavel	77,3%
21	54	106	1049	121	d = RazoavelmenteDesconfortavel	77,7%

Fonte: Tabela gerada pela autora.

A matriz de confusão do HMM do eixo z, representada pela Tabela 45, apresenta o mesmo comportamento dos HMMs apresentados nesse cenário, com boa parte do erro distribuído nas classes adjacentes. Para a primeira trajetória, 43% do erro está na classe vizinha à classe esperada, enquanto que para a segunda trajetória, 69% do erro está dividido nas classes adjacentes. O HMM teve, em média, 11 pontos percentuais a menos no desempenho do ponto de vista da acurácia que as árvores.

Tabela 45 – Matriz de confusão do HMM do eixo z do quarto experimento.

a	b	c	d	e	Classificado/ Esperado	Acurácia
985	212	103	91	78	a = MuitoConfortavel	67%
42	81	168	951	109	d = RazoavelmenteDesconfortavel	70,4%

Fonte: Tabela gerada pela autora.

5.2.4.4 Metamodelo

A matriz de confusão do Metamodelo, representada pela Tabela 46, apresenta um desempenho de 84% de acurácia, em média. O erro é bem distribuído nas duas trajetórias, com exceção das classes extremas para a primeira trajetória, onde a classe “MuitoDesconfortavel” apresenta apenas 5,5% e comparação aos 15,6% apresentados na classe “MuitoConfortavel”.

Nesse contexto, o metamodelo claramente tem melhor desempenho que os modelos que o compõe, com 10 pontos percentuais, em média, de maior desempenho. Além disso, os comportamentos entre os modelos, conforme Figura 23, foram similar entre si, de forma que as árvores de decisão tiveram desempenho muito similar, e os

HMMs também se comportaram da mesma maneira. Ainda assim, o comportamento dos dois grupos é diverso o suficiente para que o desempenho do metamodelo seja melhor que o desempenho de cada modelos separadamente.

Tabela 46 – Matriz de confusão do meta modelo do quarto cenário.

a	b	c	d	e	Classificado/ Esperado	Acurácia
1249	85	71	82	12	a = MuitoConfortavel	85%
37	48	99	1113	54	d = RazoavelmenteDesconfortavel	82,4%

Fonte: Tabela gerada pela autora.

Finalmente, esse metamodelo apresenta diversidade média de 27%. Essa diversidade é composta por 26% do eixo x, 28% do eixo y e 27% do eixo z. Observa-se, novamente, que nas três duplas de avaliação a diversidade é equilibrada, mesmo assim o metamodelo tem desempenho superior ao dos modelos que o compõe.

5.2.4.5 Considerações sobre o quarto cenário

Esse cenário evidencia o melhor desempenho por parte do metamodelo em relação aos modelos que o compõe. Ainda que uma similaridade exista entre eles, eles são suficientemente diversos entre si para garantir um metamodelo tenha capacidade de classificar adequadamente as trajetórias dos ônibus.

A média de acurácia dos modelos é de 74%, mesmo com os HMMs tendo desempenho abaixo de 70%. A média de acerto do metamodelo, no entanto, é de 84%, demonstrando que ele tem capacidade de generalização para o cenário de avaliação em tela.

A dificuldade apresentada nesse cenário se repete a encontrada nos outros, e está relacionada a incapacidade dos modelos na diferenciação entre as classes próximas, onde todos os modelos apresentaram grande concentração dos erros nas classes adjacentes.

5.3 CONSIDERAÇÕES SOBRE OS RESULTADOS

O metamodelo e os modelos foram aplicados sobre um conjunto de dados de testes que não foi utilizado no treinamento. Esses dados foram selecionados por representarem avaliações sobre o mesmo trajeto do ônibus, porém, com anotações contrárias, mas em condições e horários similares.

As árvores de decisão possuem resultados satisfatórios, com capacidade de generalização para todos os cenários. Ainda assim, em aplicações no mundo geral os modelos podem enfrentar dificuldades, uma vez que existe chance razoável da classificação das instâncias acontecerem em classes próximas. Dessa maneira, aplicações

que necessitam de maior exatidão não devem se basear na classificação de um único modelo, mas sim num conjunto de modelo para tentar mitigar a ocorrência de erro de classificação.

Os modelos ocultos de Markov, por sua vez, são capazes de adicionar diversidade para a previsão do metamodelo. No entanto, apenas essa característica não é suficiente para justificar a utilização desse tipo de modelo. Ainda assim, fica demonstrado que o metamodelo precisa de alguma forma de diversidade.

Finalmente, o meta modelo tem desempenho menor que o esperado no terceiro cenário, em particular pela dificuldade de diversificação da classificação entre os modelos. Ainda assim, a utilização de um metamodelo se justificada para esse contexto, uma vez que, com exceção do terceiro cenário, o metamodelo tem desempenho superior ao apresentado individualmente por cada um dos modelos. Além disso, a necessidade de classificação de instâncias com base no conforto geral justifica a utilização de alguma forma de agrupamento/junção da classificação do conforto em cada um dos 3 eixos de aceleração.

6 CONCLUSÕES

Esse trabalho teve como objetivo construir um modelo de classificação de conforto de passageiros de transporte coletivo utilizando dados de trajetórias dos ônibus. Para tanto, dispõe de um conjunto de dados coletados em ônibus na cidade de Joinville e de um conjunto de respostas a um questionário realizado com passageiros que utilizaram o ônibus durante o período de coleta dos dados.

A literatura apresenta diferentes abordagens para a análise do conforto de passageiros de transporte coletivo, no entanto boa parte dessas abordagens conta com a opinião de especialistas que pode não ser generalizável para diferentes regiões e culturas. Além disso, o ponto de vista do passageiro dificilmente é levado em consideração, uma vez que coletar seu *feedback* e associá-lo com a viagem em si não é tarefa trivial.

Dessa maneira, a generalização de um modelo criado por técnicas de aprendizagem de máquina pode ser uma solução para a junção de dados de sensores com a percepção de conforto de passageiros para a análise de forma automática, sem a necessidade da presença de um especialista durante todo o processo. No entanto, a literatura suporta que o conforto é um conceito imensurável e pessoal, além de ser afetado por variados e diferentes aspectos tanto da viagem do passageiro quanto da pessoa em si.

Durante a implementação do trabalho em tela diferentes etapas foram vencidas até a criação dos modelos de classificação. Os dados de sensores foram obtidos por meio de um dispositivo instalado em um dos veículos que circulam na cidade de Joinville, enquanto que os dados de conforto foram coletados por meio de questionários preenchidos com a percepção dos passageiros.

Uma vez de posse de todos esses dados, eles foram processados a fim de retirar os pontos pouco significativos das trajetórias dos veículos, bem como a realização do consenso entre as percepções dos passageiros em cada uma das viagens. Finalmente, os dados de sensores foram anotados com os dados de percepção dos passageiros. Essa anotação leva em consideração o cálculo de médias e desvio padrão dos dados provindos dos sensores de acelerômetro e suas equivalências na escala de percepção, de 1 a 5. Os dados de acelerômetro foram eleitos como ponto de apoio para a anotação, uma vez que a literatura, de um modo geral, apresenta as acelerações triaxiais como as principais influências no conforto de passageiros.

Os modelos de classificação de dados foram selecionados com o objetivo de generalizar os resultados ao mesmo tempo que as diferentes características de cada

instância nas trajetórias fosse valorizada. Dessa maneira, um modelo de árvore de classificação foi selecionado, uma vez que esse tipo de classificador se preocupa com os aspectos de maior significância para o conjunto de dados, ignorando quaisquer outros. Por outro lado, um modelo oculto de Markov foi selecionado objetivando a valorização das influências entre os diferentes aspectos que compõem uma trajetória e que podem afetar o conforto dos passageiros.

Os submodelos de árvores de decisão e modelos ocultos de Markov foram treinados para cada um dos eixos de aceleração, e um meta modelo foi utilizado para unificar seus resultados. A árvore final foi construída através do empilhamentos dos submodelos, ou seja, os resultados das classificações dos modelos foram utilizados para o treinamento do meta modelo. Como os modelos ocultos de Markov podem se tornar muito especializados, a árvore foi selecionada para que o conforto geral fosse generalizável.

De um modo geral, os resultados apresentados nesse trabalho são promissores e apresentam uma visão diferente para esse problema discutido na literatura. As árvores de decisão apresentadas demonstram algumas das influências e interferências que os diferentes atributos podem causar no conforto de passageiros, mas não são suficientes para construir um modelo de classificação de conforto em transporte público, tanto por não fornecerem informações suficientes para isso, quanto pelo fato de que os modelos foram divididos em diferentes árvores que não consideram o conforto geral da viagem, apenas o conforto em cada eixo avaliado. Os modelos ocultos de Markov, por sua vez, apresentaram acurácia inferior às árvores de decisão, mas ainda assim são adequados para o contexto analisado, pois eles complementam os resultados obtidos pelas árvores.

Os resultados do meta modelo foram superiores aos dos modelos que o compõe em até 10 pontos percentuais, demonstrando que a escolha pela sua utilização foi adequada e agregou valor à solução. Por outro lado, existem situações em que o meta modelo tem desempenho inferior ao dos modelos base, possivelmente pela falta de diversidade entre eles.

De um modo geral, esse modelo alcançou seu objetivo de combinar os dados de percepção provindos dos passageiros com os dados de telemetria colhidos por sensores, para utilizá-los no treinamento de modelos de aprendizagem de máquina. Os modelos gerados são, em última instância, generalizáveis e atingem satisfatórios índices de acurácia com relação à percepções de conforto dos passageiros, sejam elas particularizadas por eixo de aceleração ou no contexto da viagem como um todo.

6.1 TRABALHOS FUTUROS

Esses resultados são promissores, mas poderiam ser mais apurados. Para tanto, é interessante realizar experimentos com outras técnicas de anotação de dados, talvez com menor importância para os eixos de aceleração, ou ainda, com o mesmo enfoque apresentado para as acelerações combinando-as mais especificamente com características particulares dos dados. Por exemplo, na anotação de eixo x de aceleração considerar esse eixo e também a velocidade no momento da anotação, além das outras acelerações com menor grau de importância para a anotação dos outros eixos.

A exploração de outras formas de coleta de dados qualitativos pode, também, ser benéfica para o resultado. A coleta em tempo real e em mais de um momento durante a viagem pode, também, aumentar a sensibilidade para os diferentes aspectos cinéticos além do eixos de aceleração. Dessa maneira, os dados poderiam ser anotados em tempo real, durante o instante em que o passageiro sentiu maior conforto ou desconforto. Ainda com relação à percepção, incluir outros aspectos de conforto em um questionário simples pode, também, auxiliar no enriquecimento do modelo como um todo, com perguntas relacionadas à percepção com relação à velocidade, tempo de viagem, entre outros aspectos.

A coleta de mais percepções, em viagem mais diversificadas, com mais de um veículo certamente pode contribuir para a melhora do desempenho do modelo, uma vez que dados com maior variação de características oferecem mais oportunidades de aprendizado para os modelos. Dessa maneira, é possível que a combinação de diferentes técnicas de coletas com diferentes veículos e em diferentes locais e momentos do tempo possam contribuir para o desempenho no modelo, em particular na generalização para contextos variados daqueles estudados no trabalho em tela, como a possibilidade de aplicação em outras localidades.

A utilização de técnicas de consenso mais sofisticadas, ou, ainda que considerem mais de uma classificação para a mesma instância pode, também, contribuir para a melhora do modelo. Essa dupla classificação pode ser utilizada para alguma técnica de votação entre modelos, de maneira que em tempo de execução do treinamento o próprio algoritmo escolha a melhor classificação possível para a instância. Assim o meta modelo pode empregar técnicas que tenham diferentes opções para maximizar a acurácia na classificação do modelo.

A utilização do HMM demonstrou ser pouco eficiente para a criação do meta modelo, ainda que tenha sido importante para o resultado como um todo por meio do aumento da diversidade. Dessa maneira, seria interessante utilizar outras técnicas de classificação de dados que tenham implementação mais simples e que sejam

mais maleáveis e menos sensíveis. Técnicas de *clustering*, por exemplo, com base na combinação das diferentes características, em duplas, trios ou conjuntos de maior tamanho podem auxiliar no aumento da diversidade do metamodelo, oferecendo diferentes possibilidades para um mesmo dado de maneira que o meta modelo tenha maior variedade para maximizar a classificação.

Similarmente ao HMM, as redes Bayesianas se preocupam com as relações probabilísticas entre causa e efeito. Dessa maneira, sendo o conforto ou falta dele o efeito, as características cinéticas seriam a causa desse conforto. É possível que a utilização dessas redes tanto combinadas com as árvores quanto com os HMMs, quanto somente com as árvores pode ser uma fonte de diversidade para o meta modelo e, por consequência, melhorar o desempenho do modelo.

É possível, ainda que esse modelo seja replicado com resultados similares ou melhores com a utilização apenas das árvores de classificação, mas com a aplicação de alguma técnica de diversificação dos conjuntos de treinamento, como as que foram discutidas nas técnicas de *bagging* e *boosting*. Dessa maneira, um meta modelo heterogêneo, com características dos três tipos de combinação de submodelos pode ter resultados com melhor generalização.

A aplicação de um algoritmo de *random forest*, por exemplo, para o treinamento de cada modelo pode, também, auxiliar na melhora de desempenho do modelo. Esse algoritmo consiste na combinação de diversas árvores de decisão no mesmo conjunto de dados e da escolha, entre elas, do melhor resultado possível, com foco na acurácia. Dessa maneira o algoritmo tem características de um meta modelo, como o *bagging* e pode, para cada modelo, melhorar os resultados individuais que, por sua vez, podem melhorar o resultado do meta modelo.

REFERÊNCIAS

- AGGARWAL, C. C. **Data classification: algorithms and applications.** [S.I.]: CRC press, 2014.
- BAYER, F. R. **Uma abordagem para identificação de perfis de motoristas do transporte coletivo com foco na minimização do consumo de combustível.** 2017. Monografia (Bacharel em Ciência da Computação), UDESC (Universidade do Estado de Santa Catarina), Joinville, Brasil.
- BEURIER, G. Analysis of the Discomfort Feeling of Standing Bus Passengers on the TEOR T1 Rouen Bus Lane. **Procedia - Social and Behavioral Sciences**, v. 48, p. 425–434, 2012. ISSN 18770428. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S1877042812027589>>.
- BILMES, J. A. et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. **International Computer Science Institute**, v. 4, n. 510, p. 126, 1998.
- BINDER, J.; MURPHY, K.; RUSSELL, S. Space-efficient inference in dynamic probabilistic networks. **Bclr**, Citeseer, v. 1, p. t1, 1997.
- BOGORNY, V.; BRAZ, F. Introdução a trajetórias de objetos móveis: conceitos, armazenamento e análise de dados. **Univille**, 2012.
- BOUTEN, C. V. et al. A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. **IEEE Transactions on Biomedical Engineering**, IEEE, v. 44, n. 3, p. 136–147, 1997.
- BUIZZA, R.; MILLEER, M.; PALMER, T. N. Stochastic representation of model uncertainties in the ecmwf ensemble prediction system. **Quarterly Journal of the Royal Meteorological Society**, Wiley Online Library, v. 125, n. 560, p. 2887–2908, 1999.
- CARDOSO, V. E. et al. A discussion about thermal comfort evaluation in a bus terminal. **Energy and Buildings**, Elsevier B.V., v. 168, p. 86–96, 2018. ISSN 03787788. Disponível em: <<https://doi.org/10.1016/j.enbuild.2018.03.013>>.
- CASCIOLI, V.; HEUSCH, A. I.; MCCARTHY, P. W. Does prolonged sitting with limited legroom affect the flexibility of a healthy subject and their perception of discomfort? **International Journal of Industrial Ergonomics**, Elsevier Ltd, v. 41, n. 5, p. 471–480, 2011. ISSN 01698141. Disponível em: <<http://dx.doi.org/10.1016/j.ergon.2011.06.003>>.
- CASTELLANOS, J. C.; FRUETT, F. Embedded system to evaluate the passenger comfort in public transportation based on dynamical vehicle behavior with user's feedback. **Measurement**, Elsevier, v. 47, p. 442–451, 2014.
- CHOWDHURY, S. Users' willingness to ride an integrated public-transport service: A literature review. **Transport Policy**, Elsevier, v. 48, p. 183–195, 2016. ISSN 1879310X. Disponível em: <<http://dx.doi.org/10.1016/j.tranpol.2016.03.007>>.

CNT. **Anuario CNT do transporte – estatísticas consolidadas 2017.** CNT, 2017. Disponível em: <<http://anuariodotransporte.cnt.org.br/2017/Rodoviario/1-4-2-1-1-/Frota>>.

COHN, N. Tomtom traffic index: Toward a global measure. **ITS France, Paris**, 2014.

De Oña, J. et al. Perceived service quality in bus transit service: A structural equation approach. **Transport Policy**, Elsevier, v. 29, p. 219–226, 2013. ISSN 0967070X. Disponível em: <<http://dx.doi.org/10.1016/j.tranpol.2013.07.001>>.

DEB, S.; Ali Ahmed, M. Determining the service quality of the city bus service based on users' perceptions and expectations. **Travel Behaviour and Society**, Elsevier, v. 12, n. February, p. 1–10, 2018. ISSN 2214367X. Disponível em: <<https://doi.org/10.1016/j.tbs.2018.02.008>>.

DIETTERICH, T. G. Ensemble methods in machine learning. In: SPRINGER. **International workshop on multiple classifier systems**. [S.I.], 2000. p. 1–15.

DIETTERICH, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. **Machine learning**, Springer, v. 40, n. 2, p. 139–157, 2000.

DOSIO, A. Projections of climate change indices of temperature and precipitation from an ensemble of bias-adjusted high-resolution euro-cordex regional climate models. **Journal of Geophysical Research: Atmospheres**, Wiley Online Library, v. 121, n. 10, p. 5488–5511, 2016.

DRUCKER, H. et al. Boosting and other ensemble methods. **Neural Computation**, MIT Press, v. 6, n. 6, p. 1289–1301, 1994.

EBOLI, L.; MAZZULLA, G.; PUNGILLO, G. Measuring bus comfort levels by using acceleration instantaneous values. **Transportation research procedia**, Elsevier, v. 18, p. 27–34, 2016.

EDEN, A. H. Three paradigms of computer science. **Minds and machines**, Springer, v. 17, n. 2, p. 135–167, 2007.

ELLIOTT, L. T.; TEH, Y. W. et al. A nonparametric hmm for genetic imputation and coalescent inference. **Electronic Journal of Statistics**, The Institute of Mathematical Statistics and the Bernoulli Society, v. 10, n. 2, p. 3425–3451, 2016.

ERGONOMICS of the thermal environment Evaluation of thermal environments in vehicles Part 3 Evaluation of thermal comfort using human subjects. Geneva, CH, 2006. v. 1.

ERIKSSON, P.; FRIBERG, O. Ride comfort optimization of a city bus. **Structural and Multidisciplinary Optimization**, v. 20, n. 1, p. 67–75, 2000. ISSN 1615147X.

ESSADEQ, I.; DUBAIL, E.; JEANNIERE, E. Modelling Passenger Congestion in Transit System -Benchmark and Three Case Studies. **Transportation Research Procedia**, v. 14, n. 2, p. 1792–1801, 2016. ISSN 23521465.

- FELLESSON, M.; FRIMAN, M. Perceived satisfaction with public transport service in nine european cities. In: **Journal of the Transportation Research Forum**. [S.I.: s.n.], 2012. v. 47, n. 3.
- FREYSS, M.; GIESEN, R.; MUÑOZ, J. C. Continuous approximation for skip-stop operation in rail transit. **Transportation Research Part C: Emerging Technologies**, Elsevier Ltd, v. 36, p. 419–433, 2013. ISSN 0968090X. Disponível em: <<http://dx.doi.org/10.1016/j.trc.2013.07.004>>.
- GALAR, M. et al. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, IEEE, v. 42, n. 4, p. 463–484, 2011.
- GARDINER, C. W. et al. **Handbook of stochastic methods**. [S.I.]: springer Berlin, 1985. v. 3.
- GIANNOTTI, F. et al. Trajectory pattern mining. In: ACM. **Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.I.], 2007. p. 330–339.
- GOUDA, K.; ZAKI, M. J. Efficiently mining maximal frequent itemsets. In: IEEE. **Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on**. [S.I.], 2001. p. 163–170.
- HADDAD, K. E. et al. Speech-laughs: an hmm-based approach for amused speech synthesis. In: IEEE. **Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on**. [S.I.], 2015. p. 4939–4943.
- HAGERTY, M. R. Testing maslow's hierarchy of needs: National quality-of-life across time. **Social Indicators Research**, Springer, v. 46, n. 3, p. 249–271, 1999.
- HARTLE, J. B. **Gravity: An introduction to Einstein's general relativity**. [S.I.]: AAPT, 2003.
- HASSAN, M. N.; HAWAS, Y. E.; AHMED, K. A multi-dimensional framework for evaluating the transit service performance. **Transportation Research Part A: Policy and Practice**, Elsevier, v. 50, p. 47–61, 2013.
- HASTIE, T. et al. The elements of statistical learning: data mining, inference and prediction. **The Mathematical Intelligencer**, Springer, v. 27, n. 2, p. 83–85, 2005.
- HERINGER, R. Desigualdades raciais no brasil: síntese de indicadores e desafios no campo das políticas públicas. **Cadernos de Saude publica**, SciELO Public Health, v. 18, p. S57–S65, 2002.
- HINTON, G. E.; SEJNOWSKI, T. J.; POGGIO, T. A. **Unsupervised learning: foundations of neural computation**. [S.I.]: MIT press, 1999.
- HOBEROCK, L. L. A Survey of Longitudinal Acceleration Comfort Studies in Ground Transportation Vehicles. **Journal of Dynamic Systems, Measurement, and Control**, v. 99, n. 2, p. 76, 1977. ISSN 00220434.
- IMRE, S.; ÇELEBI, D. Measuring Comfort in Public Transport: A case study for Istanbul. **Transportation Research Procedia**, v. 25, p. 2445–2453, 2017. ISSN 23521465.

INNAMAA, S.; PENTTINEN, M. Impacts of a green-driving application in city buses on fuel consumption, speeding and passenger comfort. **IET Intelligent Transport Systems**, IET, v. 8, n. 5, p. 435–444, 2013.

KERR, J.; DUNCAN, S.; SCHIPPERJIN, J. Using global positioning systems in health research: a practical approach to data collection and processing. **American journal of preventive medicine**, Elsevier, v. 41, n. 5, p. 532–540, 2011.

KOLCABA, K. Y.; KOLCABA, R. J. An analysis of the concept of comfort. **Journal of advanced nursing**, Wiley Online Library, v. 16, n. 11, p. 1301–1310, 1991.

KÖLSCH, M.; BEALL, A. C.; TURK, M. An objective measure for postural comfort. In: SAGE PUBLICATIONS SAGE CA: LOS ANGELES, CA. **Proceedings of the Human Factors and Ergonomics Society Annual Meeting**. [S.I.], 2003. v. 47, n. 4, p. 725–728.

LE, T. P. L.; TRINH, T. A. Encouraging public transport use to reduce traffic congestion and air pollutant: a case study of ho chi minh city, vietnam. **Procedia Engineering**, Elsevier, v. 142, p. 236–243, 2016.

MALHOTRA, N. K. et al. Dimensions of service quality in developed and developing economies: multi-country cross-cultural comparisons. **International marketing review**, Emerald Group Publishing Limited, v. 22, n. 3, p. 256–278, 2005.

MALVERN, L. E. **Introduction to the Mechanics of a Continuous Medium**. [S.I.]: s.n., 1969.

MARCONI, M. d. A.; LAKATOS, E. M. **Fundamentos de metodologia científica**. [S.I.]: 5. ed.-São Paulo: Atlas, 2003.

MARÍN, L.; QUEIROZ, M. S. A atualidade dos acidentes de trânsito na era da velocidade: uma visão geral. **Cadernos de Saúde Pública**, SciELO Public Health, v. 16, p. 7–21, 2000.

MECHANICAL vibration– Description and determination of seated postures with reference to whole-body vibration. Geneva, CH, 2012. v. 1.

MECHANICAL vibration and shock – Evaluation of human exposure to whole-body vibration – Part 1: General requirement. Geneva, CH, 1995. v. 2.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. **Ain Shams Engineering Journal**, Elsevier, v. 5, n. 4, p. 1093–1113, 2014.

MELLO, R. d. S. et al. Master: A multiple aspect view on trajectories. **Transactions in GIS**, Wiley Online Library, v. 23, n. 4, p. 805–822, 2019.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, v. 1, n. 1, p. 32, 2003.

MONZÓN, A.; HERNÁNDEZ, S.; Di Ciommo, F. Efficient Urban Interchanges: The City-HUB Model. **Transportation Research Procedia**, Elsevier B.V., v. 14, p. 1124–1133, 2016. ISSN 23521465. Disponível em: <<http://dx.doi.org/10.1016/j.trpro.2016.05.183>>.

- MORIN, E.; MOIGNE, J.-L. L. A inteligência da complexidade. Ed. Fundação Peirópolis, 2000.
- NAGY, P.; NÉMETH, G. Improving hmm speech synthesis of interrogative sentences by pitch track transformations. **Speech Communication**, Elsevier, v. 82, p. 97–112, 2016.
- NIELSEN, S. V.; SIMONSEN, S.; HOBOLTH, A. Inferring population genetic parameters: Particle filtering, hmm, ripley's k-function or runs of homozygosity? In: SPRINGER. **International Workshop on Algorithms in Bioinformatics**. [S.I.], 2016. p. 234–245.
- NOOR, H. M.; NASRUDIN, N.; FOO, J. Determinants of Customer Satisfaction of Service Quality: City Bus Service in Kota Kinabalu, Malaysia. **Procedia - Social and Behavioral Sciences**, Elsevier B.V., v. 153, p. 595–605, 2014. ISSN 18770428. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S1877042814055347>>.
- OBORNE, D. J.; CLARKE, M. J. Questionnaire surveys of passenger comfort. **Applied Ergonomics**, v. 6, n. 2, p. 97–103, 1975. ISSN 00036870.
- OPITZ, D.; MACLIN, R. Popular ensemble methods: An empirical study. **Journal of artificial intelligence research**, v. 11, p. 169–198, 1999.
- ORTIZ, M. A.; KURVERS, S. R.; BLUYSSSEN, P. M. A review of comfort, health, and energy use: Understanding daily energy use and wellbeing for the development of a new approach to study comfort. **Energy and Buildings**, Elsevier, v. 152, p. 323–335, 2017.
- PALA, U.; OZ, H. R. An investigation of thermal comfort inside a bus during heating period within a climatic chamber. **Applied Ergonomics**, Elsevier Ltd, v. 48, p. 164–176, 2015. ISSN 18729126. Disponível em: <<http://dx.doi.org/10.1016/j.apergo.2014.11.014>>.
- PALMA, A. de; KILANI, M.; PROOST, S. Discomfort in mass transit and its implication for scheduling and pricing. **Transportation Research Part B: Methodological**, v. 71, p. 1–18, 2015. ISSN 01912615.
- PARENT, C. et al. Semantic trajectories modeling and analysis. **ACM Computing Surveys (CSUR)**, ACM, v. 45, n. 4, p. 42, 2013.
- PARK, J. et al. Reduced ride comfort caused by beating idle vibrations in passenger vehicles. **International Journal of Industrial Ergonomics**, Elsevier, v. 57, p. 74–79, 2017.
- PIMENTA, A. M.; ASSUNÇÃO, A. Á. Thermal discomfort and hypertension in bus drivers and chargers in the metropolitan region of Belo Horizonte, Brazil. **Applied Ergonomics**, v. 47, p. 236–241, 2015. ISSN 18729126.
- PINEAU, C. The psychological meaning of comfort. **Applied Psychology**, Wiley Online Library, v. 31, n. 2, p. 271–282, 1982.
- POLIKAR, R. Ensemble based systems in decision making. **IEEE Circuits and systems magazine**, IEEE, v. 6, n. 3, p. 21–45, 2006.

- QUINLAN, R. J. C4. 5: Programs for machine learning. Morgan Kaufmann, 1993.
- RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. **Proceedings of the IEEE**, ieee, v. 77, n. 2, p. 257–286, 1989.
- RABINER, L. R.; JUANG, B.-H. An introduction to hidden markov models. **ieee assp magazine**, Citeseer, v. 3, n. 1, p. 4–16, 1986.
- RAILWAY applications – Heating, ventilation and air conditioning systems for rolling stock – Part 2: Thermal comfort. Geneva, CH, 2020. v. 1.
- RAYANA, S.; ZHONG, W.; AKOGLU, L. Sequential ensemble learning for outlier detection: A bias-variance perspective. In: IEEE. **2016 IEEE 16th International Conference on Data Mining (ICDM)**. [S.I.], 2016. p. 1167–1172.
- REZENDE, S. O. **Sistemas inteligentes: fundamentos e aplicações**. [S.I.]: Editora Manole Ltda, 2003.
- RICHARDSON, R. J. **Pós-Graduação-Metodologia-Pesquisa Social: Métodos e Técnicas-Métodos Quantitativos e Qualitativos-Capítulo 5**. [S.I.]: Editora ATLAS SA-2015-São Paulo, 2017.
- ROTMANS, J.; KEMP, R.; ASSELT, M. V. More evolution than revolution: transition management in public policy. **foresight**, MCB UP Ltd, v. 3, n. 1, p. 15–31, 2001.
- RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. [S.I.]: Malaysia; Pearson Education Limited,, 2016.
- SCHRANK, D.; EISELE, B.; LOMAX, T. Tti's 2012 urban mobility report. **Texas A&M Transportation Institute. The Texas A&M University System**, p. 4, 2012.
- SCHREIBER, J. Pomegranate: fast and flexible probabilistic modeling in python. **Journal of Machine Learning Research**, v. 18, n. 164, p. 1–6, 2018.
- SCHUBERT, P. et al. Biomechanical demand analysis of older passengers in a standing position during bus transport. **Journal of Transport and Health**, Elsevier Ltd, v. 4, p. 226–236, 2017. ISSN 22141405. Disponível em: <<http://dx.doi.org/10.1016/j.jth.2016.12.002>>.
- SEKULIĆ, D. et al. Analysis of vibration effects on the comfort of intercity bus users by oscillatory model with ten degrees of freedom. **Applied Mathematical Modelling**, v. 37, n. 18-19, p. 8629–8644, 2013. ISSN 0307904X.
- SEKULIĆ, D. et al. Definition and determination of the bus oscillatory comfort zones. **International Journal of Industrial Ergonomics**, v. 53, p. 328–339, 2016. ISSN 18728219.
- SEKULIĆ, D. et al. Analysis of bus users' vibration exposure time. **International Journal of Industrial Ergonomics**, v. 65, p. 26–35, 2018. ISSN 18728219.
- SEPUD. **FUNDACAO IPPUJ: Joinville Cidade em Dados 2016**. [S.I.]: Prefeitura Municipal de Joinville, 2016.

SEREDYNSKI, M. et al. Improving bus ride comfort using GLOSA-based dynamic speed optimisation. **Proceedings of the International Parallel and Distributed Processing Symposium, IPDPS**, p. 457–463, 2014. ISSN 23321237.

SHAM, R.; SAMSUDIN, N.; RAHMAN, K. Managing Public Transport Service Operation in Reducing Travel Fear Factor. **Procedia - Social and Behavioral Sciences**, Elsevier B.V., v. 101, p. 338–344, 2013. ISSN 18770428. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S1877042813021034>>.

SHEK, K. W.; CHAN, W. T. Combined comfort model of thermal comfort and air quality on buses in Hong Kong. **Science of the Total Environment**, v. 389, n. 2-3, p. 277–282, 2008. ISSN 00489697.

SILANON, K.; SUVONVORN, N. Hand motion analysis for thai alphabet recognition using hmm. **International Journal of Information and Electronics Engineering**, IACSIT Press, v. 1, n. 1, p. 65, 2011.

SIMONYI, E.; FAZEKAS, Z.; GÁSPÁR, P. Smartphone application for assessing various aspects of Urban public transport. **Transportation Research Procedia**, Elsevier B.V., v. 3, n. July, p. 185–194, 2014. ISSN 23521465. Disponível em: <<http://dx.doi.org/10.1016/j.trpro.2014.10.104>>.

SIMPSON, B. J.; SIMPSON, B. **Urban public transport today**. [S.I.]: Routledge, 2003.

SKALAK, D. B. et al. The sources of increased accuracy for two proposed boosting algorithms. In: CITESEER. **Proc. American Association for Artificial Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop**. [S.I.], 1996. v. 1129, p. 1133.

SOHONI, A. V.; THOMAS, M.; RAO, K. V. Mode shift behavior of commuters due to the introduction of new rail transit mode. **Transportation Research Procedia**, Elsevier B.V., v. 25, p. 2607–2622, 2017. ISSN 23521465. Disponível em: <<http://dx.doi.org/10.1016/j.trpro.2017.05.311>>.

STANDARDIZATION, I. O. for. **Standards**. 2018. Disponível em: <<https://www.iso.org/standards.html>>.

TENNØY, A. Why we fail to reduce urban road traffic volumes: Does it matter how planners frame the problem? **Transport Policy**, Elsevier, v. 17, n. 4, p. 216–223, 2010.

TIRACHINI, A. et al. Estimation of crowding discomfort in public transport: Results from Santiago de Chile. **Transportation Research Part A: Policy and Practice**, Elsevier Ltd, v. 103, p. 311–326, 2017. ISSN 09658564. Disponível em: <<http://dx.doi.org/10.1016/j.tra.2017.06.008>>.

TRAFFIC and travel information - Messages via media independent stationary dissemination systems - Graphic data dictionary for pre-trip and in-trip information dissemination systems. Geneva, CH, 2008. v. 2000.

TRIPP, D. Pesquisa-ação: uma introdução metodológica. **Educação e pesquisa**, Sci-ELO Brasil, v. 31, n. 3, 2005.

- VAMOS, I. R. C. **Pesquisa de percepção**. 2015. Disponível em: <<http://riocomovamos.org.br/site/visao-do-carioca/>>.
- VELT, K. B.; DAANEN, H. A. Optimal bus temperature for thermal comfort during a cool day. **Applied Ergonomics**, Elsevier Ltd, v. 62, p. 72–76, 2017. ISSN 18729126. Disponível em: <<http://dx.doi.org/10.1016/j.apergo.2017.02.014>>.
- VITERBI, A. J. A personal history of the viterbi algorithm. **IEEE Signal Processing Magazine**, IEEE, v. 23, n. 4, p. 120–142, 2006.
- WALD, R. M. **General relativity**. [S.I.]: University of Chicago press, 2010.
- WAZLAWICK, R. **Metodologia de pesquisa para ciência da computação**. [S.I.]: Elsevier Brasil, 2017. v. 2.
- WHORISKEY, K. et al. A hidden markov movement model for rapidly identifying behavioral states from animal tracks. **Ecology and evolution**, Wiley Online Library, v. 7, n. 7, p. 2112–2121, 2017.
- WITTEN, I. H. et al. **Data Mining: Practical machine learning tools and techniques**. [S.I.]: Morgan Kaufmann, 2016.
- XING, W. et al. Temporal predication of dropouts in moocs: Reaching the low hanging fruit through stacking generalization. **Computers in human behavior**, Elsevier, v. 58, p. 119–129, 2016.
- ZEEMAN, A. S.; BOOYSEN, M. J. Combining speed and acceleration to detect reckless driving in the informal public transport industry. **IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC**, n. Itsc, p. 756–761, 2013.
- ZHANG, K.; ZHOU, K.; ZHANG, F. Evaluating bus transit performance of chinese cities: developing an overall bus comfort model. **Transportation Research Part A: Policy and Practice**, Elsevier, v. 69, p. 105–112, 2014.
- ZHU, X. J. **Semi-supervised learning literature survey**. [S.I.], 2005.
- ZUCCHINI, W.; MACDONALD, I. L.; LANGROCK, R. **Hidden Markov models for time series: an introduction using R**. [S.I.]: Chapman and Hall/CRC, 2017.