

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Programa de Pós-Graduação em Engenharia Elétrica

Leonardo Vieira da Costa

**Predição de Seguro de Automóveis no
Brasil Utilizando Inteligência
Computacional**

Belo Horizonte - Minas Gerais

Fevereiro, 2020

Leonardo Vieira da Costa

Predição de Seguro de Automóveis no Brasil Utilizando Inteligência Computacional

Versão Final

Dissertação de Mestrado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica.

Cristiano Leite de Castro

Belo Horizonte - Minas Gerais

Fevereiro, 2020

C837p

Costa, Leonardo Vieira da.

Predição de seguro de automóveis no Brasil utilizando inteligência computacional [recurso eletrônico] / Leonardo Vieira da.Costa. - 2020.
1 recurso online (91 f. : il., color.) : pdf.

Orientador: Cristiano Leite de Castro..

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Anexos: f. 74-82.

Bibliografia: f. 70-73.

Exigências do sistema: Adobe Acrobat Reader.

1. Engenharia elétrica - Teses. 2. Inteligência computacional - Teses.
3. Predição (Lógica) - Teses. 4. Seguro de automóveis - Teses. I. Castro, Cristiano Leite de. II. Universidade Federal de Minas Gerais. Escola de Engenharia. III. Título.

CDU: 621.3(043)

**"Predição de Seguro de Automóveis No Brasil Utilizando
Inteligência Computacional"**

LEONARDO VIEIRA DA COSTA

Dissertação de Mestrado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do grau de Mestre em Engenharia Elétrica.

Aprovada em 07 de fevereiro de 2020.

Por:



Prof. Dr. Cristiano Leite de Castro
(UFMG) - Orientador



Prof. Dr. Luiz Carlos Bambirra Torres
DECSI (UFOP)



Prof. Dr. Wilian Soares Lacerda
Depto. Automatica (UFLA)

Agradecimentos

Agradecer em primeiro lugar a **Deus** pelo dom da vida.

Aos meus pais pela dedicação e incentivo.

Ao meu orientador e amigo pelos ensinamentos recebidos.

A minha esposa pela paciência e suporte nos momentos difíceis.

A minha filha que nasceu no decorrer desse trabalho trazendo muita alegria.

“There are no easy methods to solve hard problems.” (René Descartes)

Resumo

O uso da inteligência computacional está cada dia mais presente em diversas situações do nosso cotidiano, incluindo a área financeira, com o uso de classificadores e regressores. Uma ferramenta de classificação e regressão eficiente é a Máquina de Vetor de Suporte, projetada para obter uma boa capacidade de generalização. Este trabalho propõe a predição de seguro de automóveis no Brasil utilizando inteligência computacional. O método proposto faz a coleta de dados através de um aplicativo desenvolvido para celulares, ranqueia em termo de importância as características inseridas no aplicativo, seleciona o número ideal de características, calcula o valor esperado do seguro utilizando o SVM e, por último, calcula os quantis desejados utilizando *bootstrap*. Nos experimentos foi constatado que o método performou melhor com as métricas adotadas do que o regressor utilizado como *benchmark*.

Palavras-chave: Seguro de automóveis, Seleção de Características, Máquinas de Vetores de Suporte, *Bootstrap*.

Abstract

The use of computational intelligence is increasingly present in various situations in our daily lives, including financially with the use of classification and regression. An efficient classification and regression tool is a Support Vector Machine, designed to obtain good generalizability. This dissertation proposes the forecast of auto insurance in Brazil by using computational intelligence. The proposed method collects data through an application developed for smartphones, ranking the characteristics inserted in terms of importance, selecting the ideal number of characteristics, calculating the expected value of insurance using the SVM regression, and finally calculating the desired quantiles using the bootstrap technique. In the experiments, it was found that the method performs better with the adopted metrics, than when the regression used as a benchmark.

Keywords: Auto Insurance, Feature Selection, Support Vector Machine, Bootstrap.

Lista de Ilustrações

Figura 1 – Organograma do Sistema Nacional de Seguros Privados	18
Figura 2 – Seleção de características utilizando <i>wrappers</i>	25
Figura 3 – Seleção de características utilizando métodos embutidos	25
Figura 4 – Seleção de características utilizando filtros	26
Figura 5 – Cruzamento com Variável Real	31
Figura 6 – Cruzamento com Variável Binária	31
Figura 7 – Mutaç�o com Vari�vel Real	32
Figura 8 – Aproxima�o da PF para o problema DTLZ1	34
Figura 9 – Aproxima�o da PF para o problema DTLZ2	35
Figura 10 – Perda suave de margem para um SVM linear	37
Figura 11 – Intervalo de Confian�a	38
Figura 12 – Exemplo dos quatro Quartis	40
Figura 13 – Metodologia utilizada para predic�o do valor do seguro de autom�veis utilizando t�cnicas de aprendizado de m�quina	43
Figura 14 – Erro Quadr�tico M�dio vs N�mero de Vari�veis	58
Figura 15 – Erro Quadr�tico M�dio vs N�mero de Vari�veis obtido com o NSGA II	59
Figura 16 – Gera�o de 1000 respostas do SVM para o mesmo perfil de usu�rio . .	60
Figura 17 – Exemplo do Resultado de um Perfil em Todas as Seguradoras	62
Figura 18 – Estrat�gia adotada para compara�o do m�todo proposto e o SVM . .	63
Figura 19 – M�todo Proposto (Vermelho) vs <i>SVM Pinbal Loss</i> (Verde) com a vari�vel preco por seguradora	65
Figura 20 – M�todo Proposto (Vermelho) X <i>SVM Pinbal Loss</i> (Verde) com a vari�vel datanascimento por seguradora	66
Figura 21 – M�todo Proposto (Vermelho) X <i>SVM Pinbal Loss</i> (Verde) com a vari�vel datacarteirahabilitacao por seguradora	67
Figura 22 – Login do Aplicativo.	76
Figura 23 – Informa�es do Ve�culo.	78
Figura 24 – Informa�es do Uso do Ve�culo.	79
Figura 25 – Cadastro e Informa�es do Condutor.	81
Figura 26 – Informa�es das Coberturas do Seguro.	82

Lista de Tabelas

Tabela 1 – Importância das características usando RFE	50
Tabela 2 – Importância das características usando SS	52
Tabela 3 – Importância das características usando AG	54
Tabela 4 – Importância das características usando RF	56
Tabela 5 – Somatório da importância das características	57
Tabela 6 – Exemplificação de uma solução obtida por seguradora	61
Tabela 7 – Método Proposto vs SVM com <i>Pinbal loss</i>	64

Lista de Algoritmos

1	Eliminação Recursiva de Características	28
2	Seleção <i>Forward Stepwise</i>	28
3	Pseudocódigo Algoritmo Evolucionário	30

Lista de Abreviaturas e Siglas

UFMG	Universidade Federal de Minas Gerais
SVM	Máquinas de Vetor de Suporte
IRB	Instituto de Resseguros do Brasil
FENASEG	Federação Nacional das Seguradoras
MSE	Erro Médio Quadrático
FS	Seleção de Características
RFE	<i>Recursive Feature Elimination</i>
EA	Algoritmo Evolucionário
MOEA	Algoritmo Evolucionário Multiobjetivo
VEGA	<i>Vector Evaluated Genetic Algorithm</i>
CD	Distância de Multidão
MP	Método Proposto
PL	Pinbal Loss
SUSEP	Superintendência de Seguros Privados
RCF-V	Responsabilidade Civil Facultativa de Veículos
APP	Acidentes Pessoais de Passageiros
SVR	<i>Support Vector Regression</i>
Conecta	Conecta Corretora de Seguros Ltda
CRLV	Certificado de Registro e Licenciamento de Veículo
Fipe	Fundação Instituto de Pesquisas Econômicas
CPF	Cadastro de Pessoa Física

CNH	Carteira Nacional de Habilitação
WAR	<i>Wrappers</i>
GA	Algoritmo Genético
RF	<i>Random Forest</i>
LDA	<i>Latent Dirichlet Allocation</i>
GLM	Modelos Lineares Generalizados
IC	Inteligência Computacional
IoT	Internet das Coisas
GTB	<i>Gradient Tree-Boosted</i>

Lista de Símbolos

α	Nível de significância
σ	Desvio padrão
μ	Média
δ	Fator da mutação polinomial
τ	Quantil

Sumário

1	Introdução	16
1.1	Breve Histórico sobre Seguro no Brasil	16
1.2	Motivação	20
1.3	Trabalhos Correlatos	20
1.4	Objetivo	22
1.5	Organização do texto	22
2	Fundamentação	24
2.1	Glossário do Seguro de Automóveis	24
2.2	Padronização dos Dados	24
2.3	Seleção das Características	24
2.3.1	Ranking das Características	26
2.3.2	Eliminação Recursiva de Características	27
2.3.3	Seleção Stepwise	27
2.3.4	Random Forest	28
2.3.5	Algoritmo Genético	29
2.3.5.1	Torneio	30
2.3.5.2	Cruzamento	30
2.3.5.3	Mutação	31
2.4	Otimização Multiobjetivo	32
2.5	Problemas de Classificação e Regressão	35
2.6	Máquina de Vetor de Suporte	36
2.7	Intervalo de Confiança	37
2.8	Bootstrap	38
2.8.1	Bootstrap com regressão	39
2.9	Quantil	39
2.10	Previsão com Intervalo	40
2.11	Erro Quadrático Médio	41
3	Metodologia	42
3.1	Coleta dos Dados	43
3.2	Pré-processamento das Variáveis	44
3.3	Ranqueamento e Seleção das Características	44
3.3.1	Definição da Quantidade Mínima de Características	45
3.4	Previsão do Valor e dos Quantis do Seguro do automóvel	45
4	Resultados	47
4.1	Resultado da Coleta de Dados	47
4.2	Resultado do Pré-processamento das Variáveis	47

4.3	Resultado do Ranqueamento e Seleção das Características	49
4.4	Resultado da Definição da Quantidade Mínima de Características	57
4.4.1	Resolução do Problema Multiobjetivo: MSE e Quantidade de Variáveis	59
4.5	Resultado da Previsão do Valor e dos Quantis do Seguro do Automóvel . .	59
4.6	Comparação com Benchmarking	60
5	Conclusão e Trabalhos Futuros	68
5.1	Conclusão	68
5.2	Trabalhos Futuros	69
	Referências	70
	Anexo A Glossário do Seguro de Automóveis	74
	Anexo B Aplicativo de Coleta de Dados	76

Capítulo 1

Introdução

1.1 Breve Histórico sobre Seguro no Brasil

O seguro está entre as atividades mais antigas regulamentadas no Brasil. Ela iniciou-se no século XVI com os jesuítas, com destaque para o Padre José de Anchieta, criador de formas de mutualismo ligadas à assistência. Mais tarde, no século XVIII, foram promulgadas as “Regulações da Casa de Seguros de Lisboa” e mantidas até a Proclamação da Independência. Neste mesmo século, os portos nacionais iniciaram a exploração de seguros marítimos intermediados pela Companhia de Seguros Boa Fé. Esta companhia era sediada na Bahia e foi considerada a primeira sociedade seguradora no Brasil [Alberti and Leopoldi, 2001].

A fiscalização deu-se início logo após a implantação dos seguros. A Procuradoria de Seguros das Províncias Imperiais foi instituída e tomou como base as leis portuguesas. No início do século XIX algumas seguradoras aprovaram seus estatutos possibilitando a atuação em outros ramos, como o seguro de vida. A criação do “Código Comercial Brasileiro” teve papel fundamental para o desenvolvimento do seguro, pois incentivou as companhias a abranger o seguro terrestre. A partir daí, as empresas estrangeiras, especialmente as de seguro, passaram a se interessar pelo mercado brasileiro. Como consequência, já no século XIX surgiram as primeiras sucursais de empresas seguradoras no Brasil com matriz no exterior [Alberti and Leopoldi, 2001].

O Decreto 4.270, de 10/12/1901, em conjunto com o “Regulamento Murtinho”, ditaram a regulamentação das companhias de seguro no Brasil. O enfoque foi para os seguros de vida, marítimos e terrestres. Tal regulamento ajudou diretamente na criação da Superintendência Geral de Seguros que no final do século XIX foi substituída por uma Inspeção de Seguros. Esta última estava ligada ao Ministério da Fazenda por meio do Decreto 5.072 [Galiza, 1997].

Após o mercado sinalizar a complexidade e diversidade dos negócios, o Estado

começou a intervir como normatizador e fiscalizador, requerendo um mecanismo de modulação de interesses. Tal mecanismo teve o objetivo de atender os interesses da nação e preservar o funcionamento das instituições do mercado assegurando o cumprimento das coberturas dos seguros pelos seus segurados.

O Instituto de Resseguros do Brasil (IRB) foi criado ainda no século XIX tornando-se o instrumento estatal de ordenação econômica. Ele tinha como objetivo a proteção do mercado brasileiro em contraposição à presença das companhias estrangeiras, a regulação do resseguro e o desenvolvimento das operações de seguro em geral. O modelo começou a se desgastar por ser monopolista e centralizador. Por fim, ele não conseguiu atender às exigências demandadas pelo mercado, chegando inclusive a dificultar a livre concorrência das empresas do setor [Alberti and Leopoldi, 2001].

O Decreto-Lei 73, datado em 1966, instituiu o Sistema Nacional de Seguros Privados. A Superintendência de Seguros Privados (SUSEP) foi criada e era responsável por controlar e fiscalizar o funcionamento das sociedades seguradoras e entidades abertas de previdência privada. O dever era apurar as responsabilidades e penalizar os corretores de seguro que atuassem de forma irresponsável. Desde então, o órgão assumiu a proteção dos interesses dos consumidores de seguros no Brasil. Com isto, Alberti and Leopoldi [2001] apontaram que o IRB passou a compartilhar algumas atribuições com a SUSEP que se sobreporiam em importantes temas.

Logo após a sua criação, a SUSEP instalou o sistema de audiência pública para a definição de medidas gerais e tomada de decisões em todos os segmentos. Nesse momento, passou-se então a dar autonomia a criação de novos produtos e a expansão do setor com a criação de novas empresas. As diretrizes deste sistema estão listadas abaixo. Desta forma, estavam definidas as condições gerais de liberdade e realismo contratual que incentivou o crescimento do mercado em um cenário de competitividade justa de concorrência.

- a) Alterar os critérios para aplicação de reservas técnicas em ativos mobiliários;
- b) Extinguir a exigência de carta-patente para o funcionamento das seguradoras;
- c) Promover a indexação dos contratos.

O texto constitucional de 1988 apresentava alguns avanços para as empresas seguradoras, uma vez que o mercado brasileiro era muito modesto mediante ao internacional. Faria [2007] destacou que nos termos do Artigo 21, item VIII, a capitalização e a previdência privada ultrapassaram o limite da seguridade e evoluíram para o de investidores institucionais, passando então a integrar o sistema financeiro nacional .

Em 1996 algumas importantes medidas de natureza legal e administrativas aconteceram na história do seguro do Brasil. O destaque foi para a liberação da entrada das empresas estrangeiras e a quebra do monopólio ressegurador do IRB. Essa medida foi a

resposta à consulta do Ministério da Fazenda sobre a possibilidade de autorização para o funcionamento dessas empresas nos ramos vida e previdência. O resultado foi a decisão pela inconstitucionalidade da Resolução do Conselho Nacional de Seguros Privados (CNSP), nº 14/86. Esta resolução impedia que empresas estrangeiras tivessem mais de 50% do capital ou um terço das ações das seguradoras brasileiras. Isso ocasionou de forma imediata a entrada de pelo menos vinte empresas estrangeiras no Brasil. Em paralelo, a quebra do monopólio ressegurador do IRB consta da Ementa nº 13 feita à Constituição Federal que recebeu também o apoio da Federação Nacional das Seguradoras (FENASEG) [Alberti and Leopoldi, 2001].

Dada a Lei nº 10.190/2001, atualmente é de competência do Governo Federal a formulação da política de seguros privado, estabelecimento de suas normas e fiscalização das operações no mercado nacional conforme o organograma representado na Figura 1 da Confederação Nacional das Empresas de Seguros Gerais (CNSEG) [CNSEG, 2019].



Figura 1 – Organograma do Sistema Nacional de Seguros Privados

Fonte: CNSEG [2019]

De acordo com a [CNSEG \[2019\]](#), a Carta de Brasília foi uma declaração de princípios que nortearam as atividades das seguradoras no ano de 2002. Tal documento foi a primeira manifestação consensual e conjunta que veio a público como plataforma de demandas e proposta ao Governo. Ela foi construída em torno de três princípios:

1. Compromisso com a economia de mercado e a livre competição;
2. Responsabilidade econômica e social do setor de seguros;
3. Opção pela modernidade que se baseia na experiência do próprio mercado.

A Carta de Brasília ainda destacou: necessidade da melhor divulgação da imagem pública do seguro, desregulamentação do setor, colaboração do Governo, desestatização do seguro de acidente de trabalho e ampliação na operação de seguro de saúde. Após a publicação deste documento, a IRB, SUSEP e a Secretaria de Política Econômica se juntaram para elaborar um Plano Diretor do Sistema de Seguros, Capitalização e Previdência Complementar, reafirmando a importância dos seguintes itens:

- a) Desregulamentação do setor;
- b) Política de liberação de tarifas;
- c) Controle de solvência das empresas;
- d) Abertura do setor ao capital estrangeiro;
- e) Redefinição do papel do corretor;
- f) Reestruturação do IRB;
- g) Retorno do seguro de acidente de trabalho ao setor privado;
- h) Regulamentação de novas modalidades de seguros.

Dentro desse cenário está o de seguro de automóveis. Ele teve início na década de 20, em conjunto com a montagem da Ford no estado de São Paulo. A partir daí, criou-se a necessidade de proteção de veículo automotores que virava uma realidade na vida do brasileiro.

A criação da Comissão de Desenvolvimento Industrial e a criação da subcomissão de jipes, tratores, caminhões e automóveis em 1957 impulsionou a utilização do seguro para proteção automotiva. Conforme artigo 12 da circular 256/04 [[SUSEP, 2019](#)]: “o objetivo do seguro é o de estabelecer compromisso assumido pela sociedade seguradora perante o segurado, quanto às coberturas oferecidas, especificando, com clareza, quais são os prejuízos indenizáveis”.

Diante do panorama apresentado, o seguro de automóveis constitui-se como uma contratação opcional com o objetivo de proteger veículos terrestres de propulsão a motor e que não andem sobre trilhos. Tal seguro oferece a garantia de indenização em caso de danos ao veículo e outros eventos que estiverem cobertos no contrato. Há a opção da

inclusão de danos com terceiros. A cobertura do veículo segurado tipo abrangente, por exemplo, garante indenização para danos parciais ou totais causados ao veículo que foram resultantes de colisão, incêndio, explosão, roubo ou furto.

1.2 Motivação

Com o uso da Internet das Coisas (*Internet of Things*, em inglês (IoT)) [Gubbi et al., 2013], é possível otimizar muitas tarefas que exigem grande esforço humano. No ramo de seguros de automóveis não é diferente. Atualmente existe um alto custo monetário e temporal em que as empresas de corretagem gastam para obter resultados das seguradoras. O custo é gerado pela contratação de profissionais para inserção dos dados nos sites das seguradoras ou pelo pagamento mensal a empresas desenvolvedoras de software que fazem a conexão em algumas seguradoras para obtenção desses valores.

Há dois pontos importantes nessa fase: tempo de resposta ao usuário final e geração de uma base de dados histórica confiável. O primeiro relaciona-se com a decisão na contratação do seguro. Caso essa informação não esteja disponível no tempo ideal, o usuário final pode optar em não contratar esse tipo de serviço. O outro é a carência de dados sobre seguro de automóveis no Brasil disponível para análise computacional. Uma elevada parte dos corretores ainda fazem todo o processo de avaliação de um seguro de forma manual, da captação dos dados do potencial cliente à geração da apólice de contratação.

Diante deste cenário, esse trabalho deslumbra uma oportunidade em criar uma solução que capture os dados sem a intervenção da figura do corretor. Essa captura leva em consideração a coerência do dado, como por exemplo, uma pessoa não poder digitar uma data de primeira habilitação de condutor com um período menor que dezoito anos da data de nascimento do mesmo, por se tratar de uma regra do departamento de trânsito brasileiro. Ainda neste sentido é de senso comum entre os corretores o grande número de perguntas que as seguradoras precisam para fazer uma cotação do seguro de automóvel. Desta forma a maioria dos corretores assumem algumas respostas, sem perguntar ao segurado final, o que cria uma base de dados não confiável e pode levar até ao não pagamento da apólice do seguro pela seguradora.

1.3 Trabalhos Correlatos

Na literatura foram encontrados poucos trabalhos com o objetivo de predição do valor do seguro de automóveis com foco no cliente. As pesquisas encontradas foram desenvolvidas em outros países, como França e Portugal, que possuem regulamentações diferentes da brasileira e/ou dão ênfase ao valor do cálculo de fraude do seguro de

automóveis ou ao cálculo do prêmio que deve ser pago pela seguradora ao cliente. Algumas delas são mencionadas a seguir.

[Krashennikova et al. \[2019\]](#) destacaram o preço como um importante critério no seguro de crédito, incluindo os de renovação de seguros. Porém, há dois objetivos conflitantes neste ponto: manter o cliente na sua carteira e aumentar a receita. Dessa forma, ajustar os preços de renovação torna-se um desafio complexo. [Krashennikova et al. \[2019\]](#) propuseram solucioná-lo como um problema de decisão sequencial sendo, desta forma, um processo de decisão de Markov. Eles analisaram duas estratégias para resolução desses modelos por meio de um algoritmo de aprendizagem: a primeira foi maximizar a análise da receita de forma a reter o cliente e a segunda foi maximizar a receita sujeita ao nível de retenção do cliente.

[David \[2015\]](#) apresentou uma visão geral dos Modelos Lineares Generalizados (*Generalized Linear Models*, em inglês, ou GLM) aplicados a uma carteira de seguros de automóveis francesa. A autora apresentou o cálculo do prêmio com base em métodos estatísticos incorporando as informações disponíveis sobre o risco aceito. Por meio do software SAS ela propôs uma avaliação mais precisa das tarifas atribuídas a cada cliente.

[Henckaerts et al. \[2019\]](#) destacaram que os preços de predição normalmente operam com estrutura de GLMs. Eles basearam em técnicas de aprendizado de máquina para desenvolver novos planos de tarifa levando em consideração a frequência e a gravidade das reivindicações. A função de perda foi adaptada para dar prioridade às características (*features*) dos dados. Outro ponto discutido na pesquisa foi a explicabilidade dos modelos para as partes interessadas. Por isso, eles focaram em modelos de árvore de decisão mais simples até os mais complexos. A validação cruzada foi utilizada para melhorar os parâmetros de cada método. Como resultado, modelos baseados em árvore superaram os GLMs clássicos, permitindo que a seguradora forme portfólios mais lucrativos e se proteja contra seleção de riscos adversos.

[Yang et al. \[2018\]](#) criaram uma alternativa ao método *Tweedie* GLM com o objetivo de prever os prêmios do valor do seguro. O método utiliza uma abordagem de probabilidade de perfil com *Gradient Tree-Boosted* (GTB) para estimar os parâmetros de índice e dispersão. Eles apontaram que o método implementado foi superior aos demais existentes na precisão das previsões de prêmios e pode ser uma solução para ao problema de seleção adversa.

[Wang and Xu \[2018\]](#) indicaram que a fraude de seguro de automóvel representa um valor elevado nos custos das companhias de seguros. Ela afeta diretamente as estratégias de preços e os benefícios a longo prazo tornando-se, portanto, um importante critério de avaliação das seguradoras em busca da redução de custos. Características como o horário da reclamação e a marca do veículo, pouco utilizadas em pesquisas anteriores, foram levadas em consideração na formulação do problema. Os autores propuseram um novo

modelo de aprendizado com *Latent Dirichlet Allocation* (LDA) para detecção de fraudes. Os resultados experimentais feitos em uma base de dados real indicaram superioridade da abordagem proposta em relação aos modelos clássicos existentes.

1.4 Objetivo

O objetivo geral deste trabalho é desenvolver uma solução capaz de efetuar previsões de seguro de automóveis no Brasil utilizando técnicas de Aprendizado de Máquina. A solução proposta deve coletar os dados em um aplicativo de celular e estimar o valor que o segurado deve pagar pelo seguro do seu veículo, além dos quantis que o valor do seguro está inserido.

Os objetivos específicos do trabalho podem ser descritos pelos seguintes tópicos:

- a) Coletar os dados para cotação de seguro utilizando um aplicativo disponibilizado nas lojas virtuais *Google Play* e *App Store*;
- b) Fazer a seleção das características mais importantes para o modelo de previsão. Além disso, observou-se que a maioria dos usuários normalmente não finaliza a cotação de seguro devido ao grande número de dados que ele precisa inserir. Assim, o número de variáveis foi reduzido de forma que mantivesse significativo para a construção do método computacional;
- c) Estimar o valor do seguro do automóvel utilizando técnicas de Aprendizado de Máquina. Atualmente cada seguradora tem seu cadastro personalizado para inserção dos dados. Esse modelo gera um trabalho considerável ao corretor, que na maioria das vezes não consegue solicitar uma proposta para cada seguradora em tempo hábil. Por isso, a solução proposta deve ser capaz de estimar esses valores sem a necessidade de consultar oficialmente a seguradora, agilizando o trabalho do corretor e, claro, o tempo de resposta ao cliente. Além disso, acredita-se que os resultados podem ajudar a diminuir o custo das corretoras, evitando que elas contratem profissionais para inserir as informações nos sites ou paguem pelo serviço de cotações em software que fazem a coleta de dados em diversas seguradoras;
- d) Por último, a solução deve estimar tanto o valor do seguro quanto o valor dos quantis que a cotação está inserida. Acredita-se que é importante que o usuário final saiba o intervalo de valores possível para seu seguro. Assim, ele pode optar pelo avanço na contratação do seguro e vai receber o valor final dado pela seguradora.

1.5 Organização do texto

O presente trabalho está organizado da seguinte forma: o capítulo 2 apresenta uma fundamentação teórica a respeito dos termos relacionados à contratação de um seguro,

bem como os principais conceitos de aprendizado de máquina que foram empregados ao longo do trabalho. O capítulo 3 detalha a metodologia da pesquisa, passando pela coleta de informações, pré-processamento, seleção das características, previsão do seguro e previsão dos valores de quantis. O capítulo 4 apresenta o resultado da solução proposta e algumas análises e comparações com métodos existentes. O capítulo 5 apresenta a conclusão do trabalho e propostas futuras de pesquisa.

Capítulo 2

Fundamentação Teórica

2.1 Glossário do Seguro de Automóveis

Para melhor entendimento das nomenclaturas utilizadas na área de seguros de automóveis é necessário o esclarecimento de alguns termos. Eles estão detalhados em [CNSEG \[2019\]](#) e foram descritos no anexo [A](#).

2.2 Padronização dos Dados

O processo de padronização dos dados consiste em uniformizar os valores de maneira que o processo de classificação ou regressão não sofra desvios resultantes de diferentes unidades de grandeza. Um dos métodos comumente utilizado na estatística baseia-se na subtração do valor da variável de um conjunto de dados da média desse conjunto [[Han et al., 2011](#)]. O resultado dessa subtração é dividido pelo desvio padrão do conjunto de dados conforme Equação (2.1).

$$Z = \frac{X - \mu}{\sigma} \quad (2.1)$$

onde Z é a variável padronizada, X representa uma variável com média μ e desvio padrão σ . A nova variável Z terá média zero e desvio padrão igual a 1. Basicamente esse método corresponde a deslocar a média de um conjunto de dados para a origem do sistema cartesiano. Por outro lado, a variabilidade avaliada pelo desvio padrão passa a corresponder ao valor 1, simplificando a comparação de conjuntos padronizados.

2.3 Seleção das Características

O objetivo da seleção de características (*Feature Selection*, em inglês - FS) é selecionar um subconjunto mínimo desses atributos que permita que um problema seja claramente definido. Ao escolher este subconjunto, os recursos irrelevantes e redundantes

são removidos de acordo com alguns critérios razoáveis para que a tarefa original possa ser alcançada tão bem, ou até com uma melhor eficiência [Roffo and Melzi, 2016]. As técnicas de FS podem ser divididas em três tipos: 1) *wrappers*; 2) Métodos embutidos (*embedded methods*); 3) Filtros (*filters*).

Os *wrappers*, representados na Figura 2, utilizam classificadores para escolher um determinado subconjunto de características. O algoritmo de aprendizado é utilizado para direcionar o processo de seleção e uma heurística é utilizada para executar a busca maximizando seu desempenho. Este método normalmente possui alto custo computacional e isso pode o tornar impeditivo. Apolloni et al. [2016], por exemplo, utilizaram esta técnica para reduzir o número de características de experimentos com micro-array que usualmente possuem amostras complexas e de alta dimensionalidade.

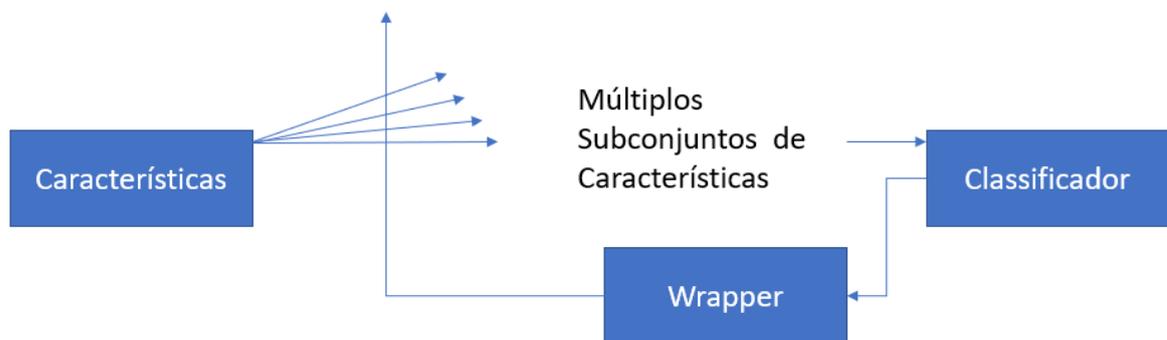


Figura 2 – Seleção de características utilizando *wrappers*

Os métodos embutidos, representados na Figura 3, inserem o processo de seleção na aprendizagem do classificador e do regressor. Assim, a seleção passa a fazer parte do algoritmo de aprendizado. Normalmente esses métodos são baseados em árvores de decisão (ou *Decision Trees*). Maldonado et al. [2014] propuseram um conjunto de métodos que selecionam os atributos relevantes utilizando métodos embutidos para a identificação da classe-alvo em um problema de classificação binária.

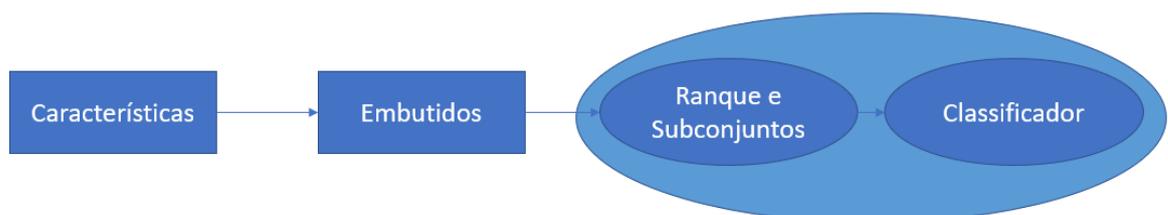


Figura 3 – Seleção de características utilizando métodos embutidos

Os filtros, representados na Figura 4, analisam as propriedades intrínsecas dos dados ignorando o classificador. Usualmente eles são robustos contra a superposição e privilegiam os conjuntos de atributos altamente relacionados com a saída desejada e pouco relacionados entre si. Por outro lado, eles fazem seleção de forma indireta, o que pode levar a piores resultados. [Abdoos et al. \[2016\]](#) utilizaram este método para eliminar recursos redundantes em problemas de qualidade de energia.



Figura 4 – Seleção de características utilizando filtros

2.3.1 Ranking das Características

No sentido de entender a importância das características para um modelo de IC, a metodologia proposta por [Cortez and Embrechts \[2013\]](#) sugeriu que uma entrada relevante (x_a) é capaz de produzir mudanças substanciais na saída ao variar seus níveis de entrada. Essa relevância de entrada é quantificada usando uma medida de sensibilidade. Tal modelo leva em consideração quatro medidas que são descritas abaixo:

1. Intervalo, dado pela Equação (2.2).

$$S_r = \max(\hat{y}_{a_j} : j \in \{1, \dots, L\}) - \min(\hat{y}_{a_j} : j \in \{1, \dots, L\}) \quad (2.2)$$

onde S_r é a diferença entre o maior e menor valor encontrado, levando em consideração o nível L ;

2. Gradiente, dado pela Equação (2.3).

$$S_g = \sum_{j=2}^L |\hat{y}_{a_j} - \hat{y}_{a_{j-1}}| / (L - 1) \quad (2.3)$$

onde S_g é o somatório da diferença entre o último e penúltimo valor encontrado, levando em consideração o nível L ;

3. Variância, dado pela Equação (2.4).

$$S_v = \sum_{j=2}^L (\hat{y}_{a_j} - \bar{y}_a)^2 / (L - 1) \quad (2.4)$$

onde S_v é o somatório da diferença entre o valor e a média dos valores encontrados ao quadrado, levando em consideração o nível L ;

4. Desvio Médio Absoluto, dado pela Equação (2.5).

$$S_d = \sum_{j=2}^L |\hat{y}_{a_j} - \tilde{y}_a| / L \quad (2.5)$$

onde S_d é o somatório da diferença entre a média e a mediana dos valores encontrados, levando em consideração o nível L ;

O gradiente é a única medida que depende da ordem das respostas de sensibilidade. Essa propriedade pode ser uma desvantagem ao abordar variáveis de entrada nominais, onde grandes diferenças podem levar a alterações de resposta de forma brusca. Para todas as medidas, quanto maior o valor, mais relevante é a variável de entrada. Ao final, o algoritmo proposto por Cortez and Embrechts [2013] soma as quatro medidas para determinar o valor de importância da variável. Esse valor é padronizado de maneira que o somatório da importância de todas as variáveis seja igual a 1.

2.3.2 Eliminação Recursiva de Características

Eliminação Recursiva de Características (*Recursive Feature Elimination*, em inglês - RFE) é um método *wrapper* de seleção *backward* (i.e. eliminação passo atrás) que começa de um modelo contendo todas as características e calcula a pontuação de importância de cada uma delas. As menos importantes são removidas. Refaz-se o modelo e as pontuações de importância são calculadas novamente [James et al., 2014]. O Algoritmo 1 contempla uma descrição detalhada deste método.

Seja S uma sequência de números ordenados que são valores candidatos para o número de características ($S_1 > S_2, \dots$). Em cada iteração de seleção de características, os principais preditores classificados de S_i são mantidos, o modelo é reajustado e o desempenho é avaliado. O valor de S_i com o melhor desempenho é determinado e os principais preditores de S_i são usados para ajustar o modelo final.

2.3.3 Seleção Stepwise

Nem sempre é possível utilizar todas as variáveis em modelos de classificação ou regressão devido ao alto custo computacional. Além disso, podem existir características no conjunto de dados que inserem ruídos ao modelo. Quanto maior o espaço de pesquisa, maior a chance de encontrar modelos que se adaptam bem aos dados de treinamento. Porém, eles podem não ter o mesmo comportamento nos dados futuros. Por esses motivos,

Algorithm 1 Eliminação Recursiva de Características

```

1: Treina o modelo usando a base de treinamento com todas as características
2: Calcula a performance do modelo
3: Calcula o ranking da variável
4: for Para cada tamanho de subconjunto  $S_i, i = 1 \dots S$  do
5:   Mantenha a variável  $S_i$  mais importante
6:   [Opcional]: Pré-processamento dos dados
7:   Treina o modelo usando a base de treinamento com as características  $S_i$ 
8:   Calcula a performance do modelo
9:   [Opcional]: Recalcula o ranking para cada predição
   end
10: Calcula a performance total de  $S_i$ 
11: Determina o correto número dos preditores
12: Usa o modelo correspondente do ótimo  $S_i$ 

```

a seleção *stepwise* (*Stepwise Selection*, em inglês - SS) pode ser uma alternativa atraente para melhorar a seleção de subconjuntos de características [James et al., 2014].

A seleção *Forward Stepwise* é uma alternativa eficiente no que diz respeito ao custo computacional. Tal método é um método *wrapper* e considera um conjunto menor de modelos. Inicia-se um modelo com nenhuma variável e as adiciona uma por vez até que todas estejam inseridas. A vantagem no processamento computacional da seleção *forward stepwise* sobre o melhor subconjunto é bem clara e está detalhada no Algoritmo 2.

Algorithm 2 Seleção *Forward Stepwise*

```

1: Inicia com o modelo  $M_0$  com zero variáveis
2: for  $k = 0, \dots, p - 1$  do
3:   Considere todos os modelos  $p - k$  que aumentam as variáveis em  $M_k$  com uma variável adicional.
4:   Escolha o melhor entre esses modelos  $p - k$ , e rotule de  $M_k + 1$ .
   Escolha aquele com menor erro.
   end
5: Selecione o melhor modelo com validação cruzada dentre  $M_0, \dots, M_p$ 

```

2.3.4 Random Forest

O método *Random Forest* (RF) consiste em um conjunto de árvores de decisão geradas dentro de um mesmo objeto. Cada objeto (conjunto de árvores) passa por um mecanismo de votação (*bagging*) que elege a classificação e regressão mais votada [Breiman, 2001]. Tal método é uma combinação de preditores de árvores, de modo que cada árvore depende dos valores de um vetor aleatório amostrado independentemente e com a mesma distribuição para todas as árvores na floresta. O método RF é um algoritmo de aprendizado supervisionado que usa aprendizado por conjuntos para classificação e regressão [James

et al., 2014]. Ele também pode ser utilizado como um método embutido de seleção de características [Menze et al., 2009].

Para ranquear a importância das características, o RF leva em consideração duas informações: *%IncMSE* e *IncNodePurity*. O *%IncMSE* é calculado levando em consideração os dados permutados. Para cada árvore o erro médio quadrático (*Mean Square Error*, em inglês - MSE) de previsão é registrado depois de permutar cada variável preditora. A diferença entre os nós é calculada sobre toda a árvore e normalizado pelo desvio padrão das diferenças. A importância da variável é diretamente proporcional a essa diferença. Já o *IncNodePurity* refere-se à diminuição total das impurezas no nó da variável em relação a média de toda a árvore. Tal impureza leva em consideração a soma residual dos quadrados.

2.3.5 Algoritmo Genético

De uma forma simplificada, a evolução biológica pode ser analisada como mudanças na forma e comportamento de organismos vivos (indivíduos) dentro de uma população entre gerações [Ridley, 2004]. A explicação do funcionamento da evolução biológica pela teoria da seleção natural é feita por Darwin [1859]. A teoria Darwiniana da evolução diz que a evolução biológica ocorre por meio de três princípios: 1) Variedade de características; 2) Transmissão de características; 3) Luta pela existência.

Desta forma, os indivíduos mais adaptados ao ambiente tendem a produzir mais descendentes que os outros indivíduos da população. A esse fenômeno dá-se o nome de seleção natural [Ridley, 2004]. As características desejáveis são então passadas para os descendentes e ao longo de várias gerações elas se tornam dominantes na população.

A evolução biológica pode ser observada como um processo de otimização no sentido de que ela leva os indivíduos de uma população a se tornarem cada vez mais adaptados ao seu ambiente [Engelbrecht, 2007]. Segundo Ridley [2004], a adaptação está relacionada às características que aumentam as chances de um indivíduo de sobreviver e reproduzir em um determinado ambiente. Desta forma, um Algoritmo Evolucionário (*Evolutionary Algorithm*, em inglês - EA), detalhado no Algoritmo 3, é qualquer algoritmo estocástico de busca inspirado nos conceitos da teoria de Darwin referente a evolução [Engelbrecht, 2007].

Os EAs inicialmente foram aplicados em problemas mono-objetivos, mas atualmente já são aplicados em inúmeros problemas. Uma grande motivação para usá-los está relacionada ao fato de que eles não necessitam de informação de gradiente, o que é importante no caso em que a função objetivo é não-diferenciável. Outro fator importante destacado por Adra [2007] é que eles são capazes de fornecer boas soluções em problemas nos quais os métodos tradicionais falham, como problemas que possuem descontinuidades,

Algorithm 3 Pseudocódigo Algoritmo Evolucionário

```

1: procedure
2:   INICIALIZE população com soluções candidatas aleatórias;
3:   AVALIE cada candidato;
4:   while condição de parada não é satisfeita do
5:     SELECIONE os pais;
6:     RECOMBINE pares de pais;
7:     MODIFIQUE a prole resultante;
8:     AVALIE novos candidatos;
9:     SELECIONE indivíduos para a próxima geração;
10:  end
11: end

```

multimodalidade e grandes espaços de buscas.

2.3.5.1 Torneio

O objetivo deste método é promover um torneio entre um grupo de n ($n \geq 2$) indivíduos aleatoriamente tomados da população. O indivíduo com melhor valor de aptidão entre o grupo é selecionado para fazer parte da geração da nova população, enquanto os demais indivíduos são descartados. O processo de seleção termina quando se realiza uma quantidade de torneios igual ao tamanho da população [Eiben et al., 2003].

2.3.5.2 Cruzamento

O cruzamento consiste em recombinar o material genético de dois indivíduos a fim de criar dois novos. Esse operador tem a função de extrair de diferentes indivíduos os melhores genes e recombiná-los para formar indivíduos com melhores características. A ideia do operador cruzamento é obter vantagem do material genético presente na população [Eiben et al., 2003].

Cruzamento com Variável Real:

No operador de cruzamento com variável real foi determinado o $kcross$ como um número aleatório que identifica a variável no intervalo de $[1, nvar]$, sendo $nvar$ o número de variáveis do problema. O coeficiente de multiplicação polarizado α_{pol} é escolhido em um intervalo $[0.5, 1]$. Da mesma forma foi definido um coeficiente de multiplicação linear α em um intervalo $[-0.1, 1.1]$. É importante também a definição de uma variável aleatória dir para indicar em qual direção será feito o cruzamento. Os conteúdos que não estão incluídos no intervalo definido de P1 e P2 é adicionado em F1 e F2 respectivamente. Na Figura 5 é ilustrado um exemplo de um cruzamento de variável real com $\alpha_{pol} = 0,9$, $\alpha_{pol} = 0,5$, $kcross = 5$ e $dir = 6$.

Cruzamento com Variável Binária:

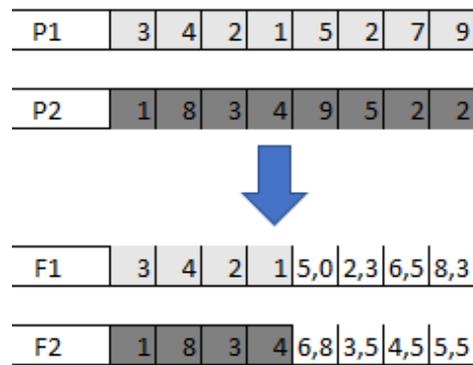


Figura 5 – Cruzamento com Variável Real

No operador de cruzamento com variável binária, uma posição x é escolhida aleatoriamente entre 0 e $(|P1|; |P2|)$, sendo $|P1|$ e $|P2|$ o número de variáveis do vetor do primeiro e do segundo pai, respectivamente. O conteúdo entre 0 e x em $P1$ e $P2$ é copiado diretamente para $F1$ e $F2$. Em seguida, os elementos contidos entre x e $|P1|$ são copiados para $F2$ e os elementos entre x e $|P2|$ são copiados para $F1$. A ilustração deste cruzamento pode ser verificada na Figura 6.

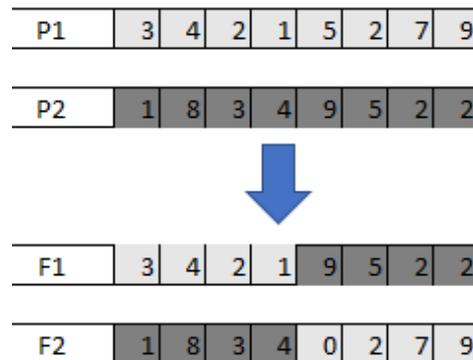


Figura 6 – Cruzamento com Variável Binária

2.3.5.3 Mutaç o

O operador de muta o   utilizado para introduzir e manter a diversidade na popula o, atrav s de perturba es nos indiv duos presentes na popula o. Desta forma, geram-se novas solu es e permite a entrada de novos elementos na popula o, aumentando, assim, a capacidade explorat ria do EA. A muta o   aplicada aos indiv duos com uma probabilidade definida pela taxa de muta o. Normalmente, esse valor   pequeno, evitando que o algoritmo se torne um m todo de busca totalmente aleat ria [Eiben et al., 2003].

Na mutação das variáveis com codificação real, os operadores são implementados como uma perturbação aplicada com uma probabilidade para cada indivíduo da população. A perturbação pode ser gerada com diferentes distribuições de probabilidade. Dentre as mutações mais comuns estão: não uniforme, polinomial, gaussiana, caótica e *cauchy*.

De forma semelhante ao cruzamento com variáveis reais, definiu-se o *kmut* como a variável alvo da mutação, *dir* indica em qual direção será feita a mutação, *range* a amplitude da faixa definida pelos limites máximo e mínimo de cada variável e β um número aleatório com distribuição uniforme entre $[-1, 1]$. Estas informações estão ilustradas na Figura 7.

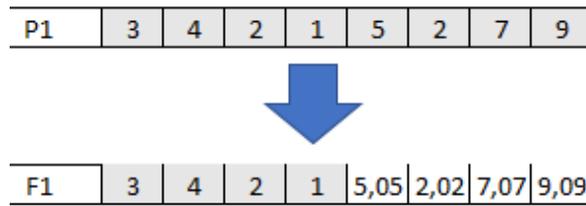


Figura 7 – Mutação com Variável Real

A Equação (2.6) representa o procedimento a ser aplicado pela mutação polinomial.

$$v_k = \pi(b_k - a_k)\delta_k \quad (2.6)$$

onde

$$\delta_k = \begin{cases} 2u_k^{\frac{1}{n_m+1}} - 1 & \text{se } u_k \leq 0.5 \\ 1 - [2(1 - u_k)]^{\frac{1}{n_m+1}} & \text{se } u_k > 0.5 \end{cases} \quad (2.7)$$

sendo que quanto maior o n_m , maior a probabilidade de se produzir uma perturbação de natureza pequena, e, portanto, maior a probabilidade de gerar uma solução aproximada da solução inicial.

2.4 Otimização Multiobjetivo

A otimização multiobjetivo (*Multiobjective Optimization*, em inglês - MOO) consiste na otimização de dois ou mais objetivos, usualmente conflitantes entre si [Meneghini et al., 2020]. A solução para problemas MOO consiste em um conjunto de soluções representando os melhores compromissos possíveis entre os objetivos. Levando em consideração a dominância de Pareto, eles representam as soluções Pareto-ótimas. Para a solução de problemas MOO, incluindo com muitos objetivos (acima de 3), ver [Meneghini et al., 2020].

Uma técnica tradicional para lidar com problema MOO é transformar o problema multiobjetivo em um problema mono-objetivo [Adra, 2007]. O problema convertido é

então solucionado por alguma técnica convencional de otimização mono-objetivo. Como resultado dessa transformação, uma única solução é encontrada sobre a fronteira Pareto (*Pareto Front*, em inglês - PF) em cada execução [Engelbrecht, 2007]. Um ponto negativo é que para gerar uma boa aproximação da PF esses algoritmos precisam ser executados várias vezes usando diferentes parâmetros para que boas soluções sejam encontrados [Adra, 2007]. Outros problemas dos métodos tradicionais são: a) alguns algoritmos são sensíveis a forma da PF [Meneghini et al., 2020]; b) eles podem requerer informação de gradiente; c) eles podem envolver a agregação de objetivos em escalas diferentes.

Os algoritmos genéticos (*Genetic Algorithm*, em inglês - AG) multiobjetivo permitem aproveitar a natureza do problema em sua própria solução ao invés de convertê-lo em um problema mono-objetivo. Desta forma, evita-se o problema de se agregar objetivos em escalas diferentes, comum nas técnicas tradicionais. Outra vantagem é que por lidar com múltiplas soluções simultaneamente, o processo de busca ocorre com cooperação e sinergia. Além disso, são capazes de trabalhar com grandes espaços de buscas, multimodalidade e não usarem informação de gradiente [Engelbrecht, 2007].

Levando em consideração as características descritas acima, esses algoritmos são conhecidos na literatura como Algoritmos Evolucionários Multiobjetivo (*Multiobjective Evolutionary Algorithm*, em inglês - MOEA) [Coello et al., 2007]. O algoritmo *Vector Evaluated Genetic Algorithm* (VEGA), proposto por Schaffer [1984], é normalmente referido como a primeira extensão de um EA para tratar problemas MOO.

A próxima estratégia investigada na literatura foi o ordenamento lexicográfico com AG [Coello et al., 2006]. O objetivo considerado mais importante era o primeiro a ser otimizado. Em seguida, o segundo objetivo mais importante é otimizado, mas sujeito à condição de não piorar o valor obtido no primeiro objetivo. Esse procedimento é repetido até o último objetivo. Porém, a desvantagem dessa técnica é que ela necessita que os objetivos sejam ordenados por importância e exige, portanto, conhecimento do problema por parte do decisor.

O grande avanço no desenvolvimento dos MOEAs ocorreu a partir da definição dada por Coello et al. [2006] de incorporar dominância Pareto na atribuição de aptidão para as soluções da população. A sugestão é baseada no conceito de solução localmente não-dominada. Tal definição é diz que uma solução $a \in A$ (vetor de espaço objetivo) é dita ser localmente não dominada com respeito ao conjunto A de vetores objetivos, se e somente se $\nexists x \in A$, tal que $x \prec a$.

Para a classe de problemas MOO, outros algoritmos, como o NSGA [Srinivas and Deb, 1994], utiliza o mecanismo de ordenamento por não-dominância para atribuir valores de aptidões para as soluções da população. Diferentemente do primeiro, o NSGA-II [Deb et al., 2002] apresenta um novo mecanismo de estimação de densidade denominado de distância de multidão (*Crowding Distance* - CD). A vantagem do CD é que ele não requer

nenhum parâmetro do usuário. Geometricamente, a CD de multidão de uma solução i corresponde ao semi-perímetro de um hiper-retângulo conectando as soluções dos vizinhos imediatos daquela solução.

Algoritmos baseados em pontos de referência incorporam um conjunto de pontos com o objetivo de direcionar a busca do algoritmo para diversidade e convergência das soluções, como o NSGA-III [Li et al., 2014]. Tal algoritmo leva em consideração o NSGA-II, substituindo a métrica CD por uma baseada em pontos de referência.

Outro algoritmo, MOEA/D, transforma um problema com dois ou mais objetivos em vários subproblemas mono-objetivo, através de uma ponderação baseada em vetores de peso. Tal metodologia tem como vantagem o custo computacional por não necessitar classificar os indivíduos em fronteiras de soluções não dominadas [Zhang and Li, 2007]. Em contrapartida, o algoritmo MOEA/DD utiliza-se dos pontos de referência introduzidas com o NSGA-III em conjunto com a ideia de decomposição do MOEA/D, reunindo simultaneamente dominância e decomposição [Li et al., 2014].

Com o objetivo de avaliar os algoritmos em problemas com muitos objetivos várias funções de teste estão disponíveis na literatura [Meneghini et al., 2020]. Dentre elas, uma das mais conhecidas é a DTLZ Deb et al. [2005]. Uma completa revisão destas e outras funções, tanto para problemas MOO quanto para muitos objetivos, está disponível em Meneghini et al. [2020].

As Figuras 8 e 9 mostram a aproximação do algoritmo NSGA-II com a PF dos problemas DTLZ1 e DTLZ2, respectivamente.

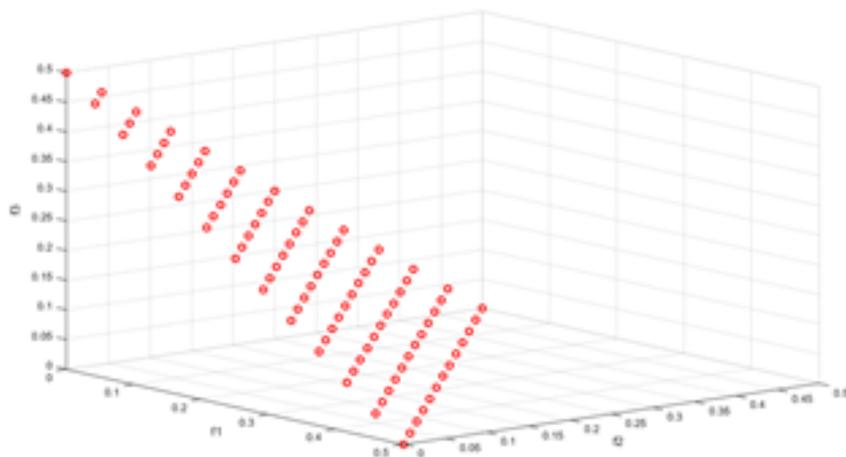


Figura 8 – Aproximação da PF para o problema DTLZ1

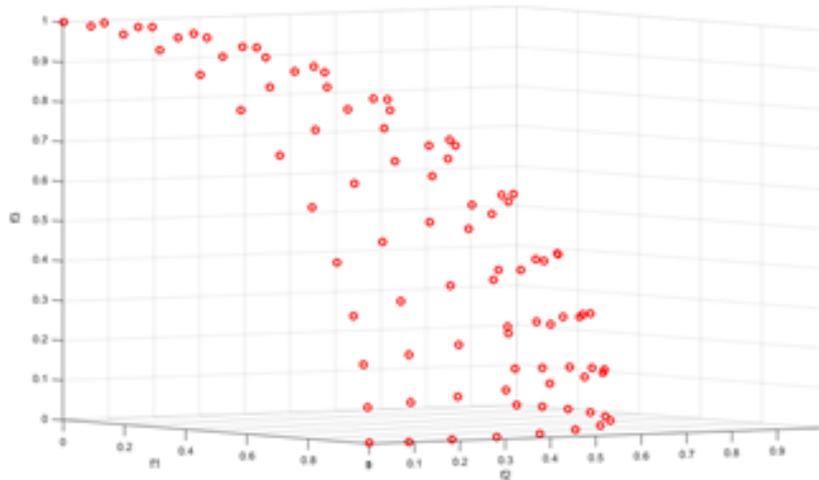


Figura 9 – Aproximação da PF para o problema DTLZ2

2.5 Problemas de Classificação e Regressão

As variáveis de um problema são as informações de entrada e podem ser classificadas como quantitativas e qualitativas. Variáveis quantitativas assumem valores numéricos, enquanto variáveis qualitativas são definidas por categorias, ou seja, representam uma classificação. As variáveis podem incluir, por exemplo, o sexo de uma pessoa (masculino ou feminino), a marca do veículo (A, B ou C), preço do veículo, ou se a pessoa possui garagem em casa para estacionar o veículo (0-possui, 1-Não Possui). De modo geral, classifica-se problemas com resposta quantitativa como problemas de regressão e resposta qualitativa como problemas de classificação.

A regressão linear é uma metodologia simples para resolver problemas com resposta quantitativa [James et al., 2014]. A técnica visa prever uma resposta quantitativa \hat{y} com base em uma única variável preditora x , como descrito pela Equação (2.8).

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (2.8)$$

onde \hat{y} é a resposta esperada, $\hat{\beta}_0$ e $\hat{\beta}_1$ são duas constantes desconhecidas que representam os termos de interceptação e inclinação no modelo linear e x a variável preditora. Porém, a maioria dos problemas reais leva em consideração mais de uma variável preditora e por isso não é possível utilizar esta técnica. Neste sentido, a regressão linear múltipla se propõe a trabalhar com problemas de regressão para múltiplas variáveis predictoras p [James et al., 2014]. Ela é descrita na equação (2.9).

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p. \quad (2.9)$$

onde \hat{y} é a predição da resposta, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ são as estimativas do coeficiente de

regressão e x_p a p -ésima preditora.

Uma limitação na aplicação de regressão linear é que, em geral, não há um processo de conversão de uma variável de resposta qualitativa com mais de dois níveis em uma resposta quantitativa. Por isso, a literatura indica o uso de outras técnicas, como a Regressão de Vetores de Suporte [James et al., 2014].

2.6 Máquina de Vetor de Suporte

Um classificador muito conhecido é Máquina de Vetor de Suporte (*Support Vector Machine*, em inglês - SVM) desenvolvida por Boser et al. [1992]. A SVM constrói um hiperplano ou conjunto de hiperplanos em um espaço de alta dimensão que pode ser usado para classificação, regressão ou outras tarefas. Uma boa separação é obtida pelo hiperplano que tem a maior distância até o ponto de dados de treinamento mais próximo de qualquer classe, chamada como margem funcional. Em geral, quanto maior a margem, menor o erro de generalização do classificador com base na teoria de aprendizado estatístico. A sua fórmula resulta em um problema convexo que pode ser resolvido por programação quadrática demonstrada em Vapnik [1999], conforme Equação (2.10).

$$\begin{aligned} \max_{\alpha} f(\alpha) &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j k(\vec{x}_i, \vec{x}_j) \\ \text{sujeito a } 0 &\leq \alpha_i \leq C \text{ e } \sum_{i=1}^n \alpha_i y_i = 0 \end{aligned} \quad (2.10)$$

onde $k(\vec{x}_i, \vec{x}_j) = \exp\left(-\frac{\|\vec{x}_i - \vec{x}_j\|^2}{2\sigma^2}\right)$

onde α_i são os coeficientes de Lagrange, C é o parâmetro de regularização e K é a função de *Kernel*. A escolha do *Kernel* e dos parâmetros influenciam diretamente na qualidade do modelo [Scholkopf and Smola, 2004].

A SVM pode ser utilizada tanto para classificação quanto para regressão, mantendo todos os principais recursos inclusive o de margem máxima que caracterizam o algoritmo.

A Regressão de Vetores de Suporte (*Support Vector Regression*, em inglês - SVR), por sua vez, faz uso do conceito do SVM para classificação, sendo que o resultado passa a ser um número real. Assim, o SVR torna-se mais complexo para fazer a predição, contendo então infinitas possibilidades. Para o caso de regressão, uma margem de tolerância (*epsilon*) é definida em aproximação ao SVM. De igual modo ao SVM, o SVR busca otimizar os limites de generalização dos dados para regressão. Tal algoritmo se baseia na definição da função de perda ignorando os erros que estão localizados a uma certa distância do valor

esperado [Smola and Schölkopf, 2004]. Existem também algumas modificações na função objetivo, conforme apresentado na Equação (2.11).

$$\begin{aligned} \text{Minimize : } & \frac{1}{2} |w|^2 + C \sum_{i=1}^N (\xi_i + \xi_i^*) \\ \text{sujeito a : } & \begin{cases} y_i - w \cdot \phi(x_i) - b \leq \varepsilon + \xi_i \\ w \cdot \phi(x_i) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \quad (2.11)$$

onde a constante $C \geq 0$ pondera os erros da função ($\xi_i, \xi_i^* \geq 0$) em função da tolerância ε descrito pela função de perda $|\xi|_\varepsilon$ chamada ξ -insensível ilustrado na Figura 10.

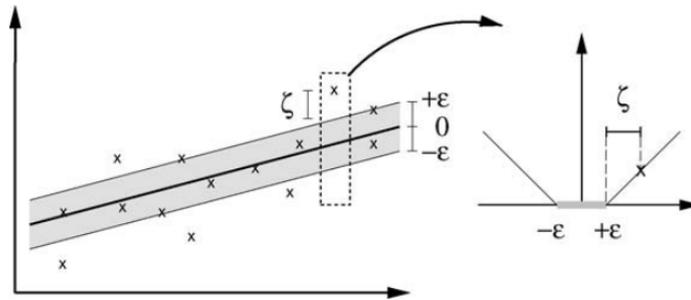


Figura 10 – Perda suave de margem para um SVM linear

Fonte: Smola and Schölkopf [2004]

2.7 Intervalo de Confiança

Um intervalo de confiança de 95% é definido como um intervalo de valores que, com 95% de probabilidade, conterà o verdadeiro valor desconhecido do parâmetro. O intervalo é definido em termos de limites inferior e superior calculados a partir da amostra de dados [Casella et al., 2013]. Uma importante grandeza para cálculo do intervalo de confiança é a média da população definida pela Equação (2.12).

$$\bar{X} = \frac{1}{n} \sum X_i \quad (2.12)$$

A média é a base para iniciar o processo de definição do intervalo de confiança. Levando em consideração o Teorema do Limite Central (TLC), temos a formulação da

Equação (2.13).

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \quad (2.13)$$

Tendo com premissa uma probabilidade $1 - \alpha$, calcula-se o quantil $Z_{\alpha/2}$ conforme Equação (2.14).

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha \quad (2.14)$$

A Figura 11 ilustra o quantil $Z_{\alpha/2} = 2$ para o nível de confiança de $95\% = (1 - \alpha)$. A quantidade $1 - \alpha$ é chamada de coeficiente de confiança e α é o nível de significância.

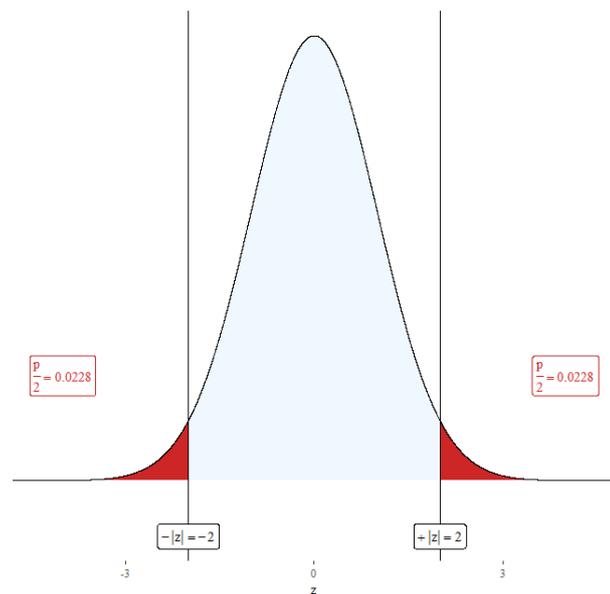


Figura 11 – Intevalo de Confiança

2.8 Bootstrap

A técnica de *Bootstrap* é uma técnica de reamostragem muito simples [Efron, 1979]. Ela pode ser utilizada para aproximar distribuição na amostra de um dado estatístico, para aproximar o viés ou a variância em dados estatísticos e para construir intervalos de confiança. No caso de possuir um conjunto de dados com poucas amostras, a técnica permite que você faça a reamostragem de forma que algumas apareçam mais de uma vez e outras sejam excluídas, ou seja, caracteriza-se como uma amostragem com substituição. No caso de o *bootstrap* ser utilizado para o cálculo do intervalo de confiança, o método executa o cálculo da média das amostras várias vezes, sendo que uma vez para cada reamostragem

dos seus dados. Posteriormente é obtido o intervalo de confiança levando em consideração o limite superior e inferior da estimativa. Desta forma, utiliza-se uma função chamada de quantil para extrair o intervalo desejado [DiCiccio and Efron, 1996].

2.8.1 Bootstrap com regressão

Com o objetivo de estimar intervalos de confiança nos parâmetros de regressão como alternativa à técnica ANOVA foi utilizado o *bootstrap*. Basicamente são definidas duas formas para esse método: amostra com substituição, onde algumas amostras não aparecem nos dados finais e outras aparecem mais de uma vez; e cálculo dos resíduos do modelo de regressão. Em ambos os casos a randomização é realizada muitas vezes, o modelo é ajustado, os parâmetros estimados e o intervalo de confiança é obtido por meio dos quantis da distribuição.

O TLC indica que a distribuição das estatísticas de um valor considerado grande do número de amostras aleatórios e independentes tem a tendência a uma distribuição normal. Isso torna-se verdadeiro mesmo levando em consideração que a população amostrada não tenha uma distribuição normal [Weisstein, 2011].

2.9 Quantil

Na maioria das vezes o ponto central de uma distribuição não descreve como os dados estão distribuídos. Possuir a mesma média não quer dizer que os conjuntos de dados são semelhantes. Neste sentido, outras medidas são utilizadas para descreverem melhor a dispersão dos dados ao redor da média. A amplitude interquartil destaca-se como uma medida de variabilidade. Ela baseia-se na divisão do conjunto de dados em quartis de quatro partes iguais, sendo que o primeiro quartil ($Q1$) está no intervalo de [0%, 25%], segundo quartil ($Q2$) no intervalo [25%, 50%], terceiro quartil ($Q3$) no intervalo [50%, 75%] e por último o quarto quartil ($Q4$) no intervalo [75%, 100%], conforme ilustrado na Figura 12.

Outro conceito importante é a divisão do conjunto de dados em percentuais. De maneira geral denomina-se “percentil” ou “quantil”, sendo que este último é quando se trata de distribuição de probabilidade [Hyndman and Fan, 1996]. O quantil de uma distribuição é definido na Equação (2.15)

$$Q(p) = F^{-1}(p) = \inf \{x : F(x) \geq p\}, 0 < p < 1, \quad (2.15)$$

onde $F(x)$ é a função de distribuição. Como regra, os quantis de uma amostra fornecem estimadores não-paramétricos que representam sua população tendo como base um conjunto de observações independentes $\{X_1, \dots, X_n\}$ da distribuição F . Seja $\{X_{(1)}, \dots, X_{(n)}\}$

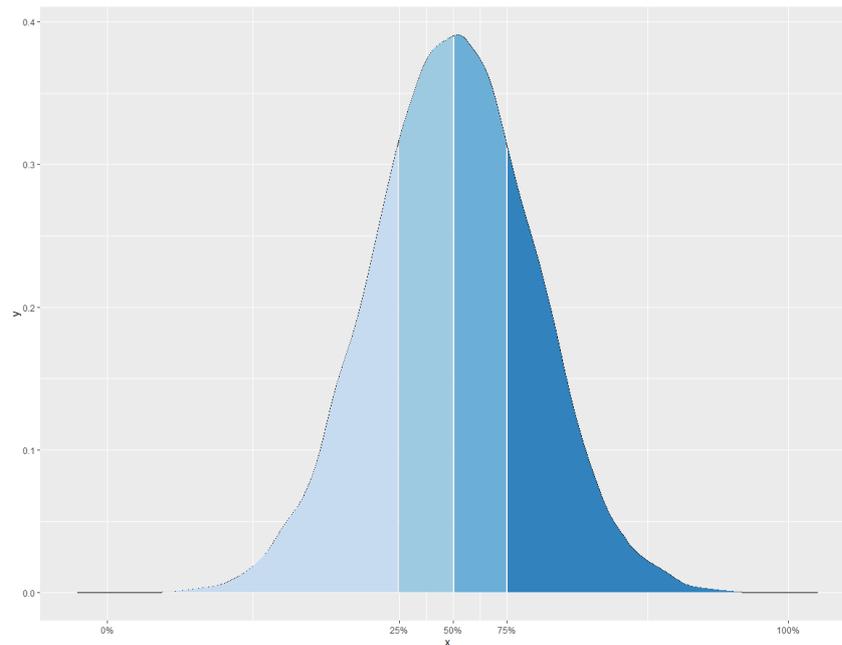


Figura 12 – Exemplo dos quatro Quartis

representa as estatísticas de ordem de $\{X_1, \dots, X_n\}$ e $\hat{Q}_i(p)$ representa a definição i -ésima de quantil de uma amostra.

2.10 Previsão com Intervalo

Devido a importância da previsão na área do aprendizado de máquina, várias evoluções do tema têm sido apresentadas na literatura. Uma importante evolução é a previsão com intervalo [Hansen, 2006]. No primeiro momento as previsões de intervalo podem ser comparadas com intervalos de confiança, porque elas compartilham a mesma estrutura, mas são coisas um pouco diferentes. Os intervalos de confiança lidam com estimativas fixas, por outro lado a previsão de intervalo são estimativas do valor futuro. Por semelhança, ambas são definidas em relação a um valor desconhecido de y com um intervalo $I = [a, b]$ com o nível de confiança α para conter o valor real de y . Para Chatfield [1993] a previsão de intervalo é uma estimativa de um valor futuro não conhecido, desta forma considerado como uma variável aleatória no momento em que a previsão é feita. O nível de confiança $\alpha \in [0, 1]$ utilizado pelo intervalo de confiança é então uma maneira de determinar um intervalo interquartil simétrico $[\alpha, 1 - \alpha]$ para algum valor previsto. Se uma distribuição de probabilidade $F : U \rightarrow [0, 1]$ encontra-se a probabilidade $F(x) = P(X \leq x)$, a função de quantil $Q : [0, 1] \rightarrow U$ executa o processo oposto: $Q(\tau) = \text{Min}_x \{X \in U \mid \tau < f(x)\}$, onde $\tau \in [0, 1]$ é um Quantil.

Chatfield [2001] propôs também um método simples para criar intervalos de previsão de nível α para modelos genéricos de previsão. Steinwart et al. [2011] fizeram uso da função *Pinbal Loss* para criar uma medida geral para a aproximação do quantil. A *Pinbal loss* é uma função de custo, sendo utilizada na otimização de modelos de predição orientados à quantis, conforme equação (2.16).

$$L_t(y, z) = \begin{cases} (y - z)\tau & \text{if } y \geq z \\ (z - y)(1 - \tau) & \text{if } z < y \end{cases} \quad (2.16)$$

onde L_τ é a função de perda *Pinbal loss*, τ o quantil alvo, y o valor real e z a previsão quantílica.

2.11 Erro Quadrático Médio

Para avaliar o desempenho de um método de aprendizagem estatístico de um conjunto de dados é preciso alguma metodologia para medir o quão próximas às previsões vão estar dos dados observados. Ou seja, é necessário quantificar a extensão para o qual o valor de resposta previsto para uma determinada observação está próximo do verdadeiro valor para essa observação. No ajuste de regressão, uma medida comumente utilizada é o MSE dado pela Equação (2.17),

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 \quad (2.17)$$

onde $\hat{f}(x_i)$ é a predição que \hat{f} dá para a i -enésima observação [James et al., 2014]. O MSE será pequeno se as respostas previstas estiverem muito próximas das respostas verdadeiras, e será grande se as observações forem substancialmente diferentes entre o previsto e o verdadeiro.

Capítulo 3

Metodologia

Esse capítulo descreve a metodologia que foi utilizada para predição dos valores do seguro de automóveis, juntamente com as predições dos intervalos quantílicos.

As etapas da metodologia representadas na Figura 13 são:

1. Coleta de dados utilizando o aplicativo de celular;
2. Pré-processamento dos dados;
3. Ordenação das características pela sua importância utilizando um comitê de métodos:
 - a) Seleção *stepwise*;
 - b) Eliminação de características recursivas;
 - c) Algoritmo Genético;
 - d) Algoritmo *Random Forest*;
4. Seleção do número mínimo de características a serem utilizadas;
5. Treinamento e ajuste dos parâmetros de um comitê de modelos SVM;
6. Geração dos quantis para cada perfil com os dados *bootstrap*.

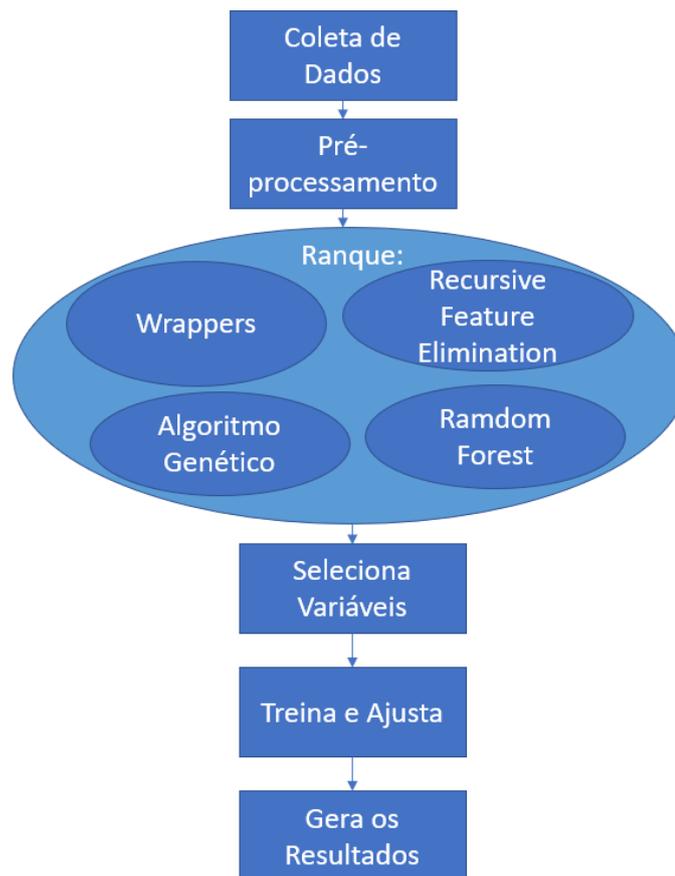


Figura 13 – Metodologia utilizada para predição do valor do seguro de automóveis utilizando técnicas de aprendizado de máquina

3.1 Coleta dos Dados

O primeiro passo para efetuar a predição do valor do seguro de automóveis utilizando aprendizado de máquina foi a coleta dos dados. A qualidade neste processo é fundamental para que a previsão seja bem sucedida. Para isso foi desenvolvido pela empresa Conecta Corretora de Seguros Ltda (Conecta) um aplicativo capaz de coletar os dados e gerar uma proposta formal de seguros para o segurado. A Conecta é uma corretora de seguros com registro na SUSEP e, portanto, credenciada a oferecer serviço para que o usuário faça a contratação de seguro através de sua operação. O aplicativo gerado pela Conecta e detalhado no Anexo B foi desenvolvido e disponibilizado nas lojas *Google Play* e *App Store*.

O usuário é responsável por fazer a inserção das informações no aplicativo descrito no Anexo B para fazer a cotação do seguro. Quando ele opta por fazer a cotação, o aplicativo faz a conexão ao servidor usando *webservice* para repassar todas as informações coletadas pelo aplicativo. O servidor, de posse dessas informações, faz uma conexão em

todas as seguradoras cadastradas no sistema da Conecta para retornar ao usuário final as informações dos preços de seguro para aquelas informações inseridas.

Na maioria das vezes essa conexão não é estabelecida com todas as seguradoras, uma vez que elas possuem um sistema instável de conexão e muitas vezes acontece uma concorrência de conexão com outros corretores. Além dessa instabilidade existe ainda uma cobrança adicional por *software* especializados em retornar essa informação, o que muitas vezes inviabiliza ou onera em muito o resultado.

3.2 Pré-processamento das Variáveis

O pré-processamento é uma fase importante no processo de classificação e regressão. Foram detectados vários problemas nesta fase, como por exemplo, o problema de ausência de dados (*missing values*). Tal problema é muito comum em diversos campos de pesquisa e deve ser solucionado antes de qualquer tipo de modelagem. Como a maioria dos procedimentos de análise não foram projetados para lidar com dados incompletos, os pesquisadores frequentemente recorrem a procedimentos de edição, excluindo casos incompletos ou substituindo os valores faltantes por meio de amostras, para dar uma aparência de completude [Yu et al., 2013].

Outro ponto importante é a categorização de algumas variáveis, que sem tratamento podem não ajudar no modelo de predição. A categorização das variáveis preditoras quantitativas do modelo baseia-se em transformar uma variável quantitativa em uma variável qualitativa ordinal com k níveis. Tal transformação implica que para cada variável originalmente quantitativa são incluídas $k - 1$ variáveis indicadoras que irão participar do modelo de regressão logística.

Por fim, para a implementação do modelo de regressão logística, foi utilizado o processo de padronização dos dados descrito anteriormente na subseção 2.2.

3.3 Ranqueamento e Seleção das Características

A seleção de características é outro processo importante no aprendizado de máquina. Muitas características podem estar correlacionadas e em determinadas situações representam ruído para o modelo. Além disso, um ponto importante para o processo de contratação e simulação de seguro é diminuir ao máximo o número de variáveis que o usuário precisa inserir no aplicativo. Essa redução resulta em maior satisfação do usuário, além de ter uma maior probabilidade de o usuário chegar ao final do processo.

Para determinar as características mais importantes foi utilizado um comitê de métodos de seleção de características para escolher a quantidade e quais seriam as características que deveriam ser informadas pelo usuário final. Foram utilizados quatro

métodos e ao final foi feito o ranqueamento das características pela soma ponderada dos pesos que cada característica possui para cada método. Tal comitê foi formado pelos métodos SS, RFE, AG e RF para cada seguradora. Com o objetivo de não priorizar nenhum dos métodos, esse trabalho optou por contabilizar o mesmo peso no voto de cada método. O somatório de todas as características normalizado é igual a 1. Isso resulta em um ranqueamento uniformizado entre todos os métodos.

A base de dados foi dividida em 70% como base de treinamento e 30% para a base de teste.

Desta forma, ao final é possível gerar um ranking de importância de característica partindo da maior para a menor. Esse ranqueamento será utilizado posteriormente para a decisão da quantidade mínima de características que serão aplicadas ao modelo.

3.3.1 Definição da Quantidade Mínima de Características

Uma vez feito o ranqueamento da importância das características, o próximo passo foi definir a quantidade mínima de características que participarão do modelo de aprendizado de máquina. Tal número é importante para satisfação do usuário final e sua interação com o uso da ferramenta. A necessidade de muitas inserções pode levar o usuário a desistir do processo e não obter o valor do seguro do automóvel. Por outro lado, é preciso coletar dados suficientes para um bom desempenho do modelo computacional.

Neste sentido, esse trabalho utilizou o somatório do MSE do conjunto de todas as seguradoras para comparar a variação desse erro em relação a inserção de variáveis. As variáveis foram inseridas no modelo por ordem de importância, da maior para a menor, conforme resultado do comitê de métodos descrito anteriormente.

3.4 Previsão do Valor e dos Quantis do Seguro do automóvel

Uma vez definido a quantidade e quais são as variáveis que participam do modelo regressor, o próximo passo foi a definição do modelo a ser utilizado. Neste trabalho foi utilizado um modelo baseado em SVM, o qual possui controle de capacidade de generalização embutido em seu processo de treinamento. Dessa forma, os parâmetros de entrada do modelo são as características de entrada, dez replicações utilizando o *bootstrap*, 9/10 como tamanho da amostra para cada *bootstrap*, 2/3 como parte dos dados para treinamento para amostragem fixas e por fim o número de 10 validações cruzadas. Para a predição quantílica foram gerados 1000 modelos via *bootstrap*. Cada processo de geração via *bootstrap* resulta em um conjunto de dados diferente para o treinamento da SVM. A partir desses modelos, foi estimado para cada amostra do conjunto de teste 1000 estimativas de

predição do valor do seguro por seguradora, o que nos permite gerar uma distribuição de estimativas, definindo assim um intervalo de predição com grau de confiança utilizado nesse trabalho de 90%.

Capítulo 4

Resultados

4.1 Resultado da Coleta de Dados

Neste trabalho foram utilizados os dados coletados pelo aplicativo detalhado no Anexo B durante o ano de 2019. Foram obtidas 2.174 requisições de cotação de diferentes perfis em quinze seguradoras, totalizando 22.534 amostras. Todavia, vale ressaltar que nem sempre todas as quinze seguradoras conseguiram informar o valor ao usuário requisitante. Por isso, não foram gerados um total de 32.610 amostras (15 seguradoras x 2.174 requisições).

Por questões de marca, optou-se por não divulgar os nomes das seguradoras. Para que a metodologia empregada possa ser replicada, utilizou-se números de 1 a 15 para representá-las.

Algumas características coletadas na inserção dos dados não fazem parte diretamente da composição do preço final do seguro. Um exemplo disso é o número de telefone do solicitante, utilizado apenas para que a corretora e/ou seguradora possa entrar em contato em eventuais problemas. Assim, o total de características (*features*) utilizadas neste trabalho foram 24 e elas são apresentadas a seguir.

4.2 Resultado do Pré-processamento das Variáveis

Algumas variáveis passaram pelo processo de categorização após a coleta de dados. A primeira delas foi a variável FIPE, onde foram utilizados os três primeiros dígitos que representam o fabricante do veículo. A segunda o CEP, utilizando os dois primeiros dígitos que representam a região e sub-região onde o seguro foi cotado. A terceira e última foi o CPF, que significa a região fiscal onde o CPF foi emitido. Já as variáveis do tipo data foram transformadas em datas do horário *Unix*, definido como o número de segundos passados desde o *epoch* (01/01/1970 00:00:00 UTC).

As demais características, com suas siglas e categorização, estão listadas abaixo por questões de organização e melhor entendimento:

- a) danos: 0 para Básico, 1 para Intermediário e 2 para Avançado;
- b) franquia: valor em reais da franquia;
- c) bonus: valor numérico entre 0 e 10;
- d) sinistro: valor numérico com a quantidade de sinistros;
- e) guincho: 0 para ter guincho e 1 para não ter;
- f) vidros: 0 para Básico, 1 para Intermediário e 2 para Avançado;
- g) carroreserva: 0 para não ter carro reserva e 1 para ter carro reserva;
- h) tipocarroreserva: 0 para Básico, 1 para Intermediário e 2 para Avançado;
- i) sexo: 0 para Masculino e 1 para Feminino;
- j) cep: dois primeiros dígitos que representam a região do CEP;
- k) datanascimentocondutor: data de nascimento do condutor em formato numérico;
- l) datacarteirahabilitacao: data da primeira habilitação do condutor convertido para valor numérico;
- m) anomodelo: valor numérico do ano do modelo do carro.
- n) fipec: valor numérico dos três primeiros dígitos da FIPE que representa o fabricante;
- o) preco: valor em reais do preço do casco do veículo;
- p) cpf: valor numérico de 0 a 9 que representa a região do CPF da residência do condutor;
- q) question10: 1 - Sim, sem portão eletrônico na residência, 2 - Sim, com portão eletrônico e 3 - Não tem garagem na residência;
- r) question11: 1 - Sim, portão eletrônico no trabalho, 2 - Não, sem portão eletrônico no trabalho, 3 - Não utiliza para trabalhar, 4 - Não trabalha;
- s) question12: 1 - Sim, portão eletrônico no Estudo/Curso, 2 - Não, sem portão eletrônico no Estudo/Curso, 3 - Não utiliza para Estudo/Curso, 4 - Não Estudo/Curso;
- t) question13: 1 - Locomoção diária, 2 - Para uso comercial, 3 - Locomoção diária e uso comercial;
- u) question14: Valor numérico com a quantidade de veículos na residência;
- v) question15: Distância do trabalho: 1 - Não informado, 2 - Até 10 km, 3 - Até 20 km, 4 - Até 30 km, 5 - Até 40 km, 6 - Acima de 40 km;
- w) question16: Valor numérico com a km Mensal;
- x) question18: Residem com o principal condutor pessoas na faixa etária de: 1 - Não residem, 2 - Entre 18 e 25 anos e não utilizam o veículo, 3 - Entre 18 e 25 anos e utilizam o veículo.

Uma vez que todas as variáveis foram armazenadas numericamente em sistema decimal, utilizou-se a técnica de padronização de dados apresentada na subseção 2.2.

4.3 Resultado do Ranqueamento e Seleção das Características

Após o pré-processamento procedeu-se o ranqueamento e seleção das características. O primeiro método utilizado foi o RFE. A Tabela 1 sumariza o peso de cada característica por seguradoras. Ao final é exibido uma coluna “SOMA” com o somatório dos pesos das características de todas as seguradoras. Elas estão ordenadas da mais importante para a menos importante, segundo o RFE.

O ranqueamento das características ficou definido da sequência do maior para o menor como: “Preco, bonus, anomodelo, datanascimentocondutor, fipe, datacarteirahabilitacao, question14, sexo, question13 , question15, sinistro, danos, cep, vidros, guincho, carroreserva, question11, question10, franquia, question18 , question16, cpf, tipocarroreserva, question12”.

Tabela 1 – Importância das características usando RFE

Características	Seguradoras															Soma
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
danos	0,01	0,07	0,05	0,05	0,02	0,00	-	0,01	0,05	0,03	0,03	0,01	0,03	0,03	0,02	0,40
franquia	0,01	0,04	0,01	0,01	0,02	0,03	-	0,01	-	0,02	0,01	0,05	0,03	-	0,02	0,26
bonus	0,07	0,05	0,02	0,06	0,08	0,09	0,06	0,11	0,13	0,08	0,05	0,08	0,07	0,07	0,06	1,09
sinistro	0,04	0,04	0,05	0,02	0,06	-	-	-	-	-	0,05	0,08	0,04	-	0,05	0,43
guincho	0,03	0,05	0,04	0,03	0,01	0,01	0,06	0,01	-	0,04	0,01	0,02	0,01	0,01	0,03	0,36
vidros	0,03	0,02	0,03	0,02	0,06	-	0,08	-	-	0,00	0,02	0,02	0,02	-	0,07	0,38
carroreserva	0,01	0,07	0,07	0,07	0,02	-	0,01	-	-	-	0,02	-	0,00	0,02	0,02	0,33
tipocarroreserva	0,00	0,01	0,02	0,01	0,01	0,03	0,06	-	-	-	0,01	-	0,01	0,01	0,01	0,18
sexo	0,02	0,03	0,03	0,04	0,04	0,03	0,04	0,04	0,06	0,04	0,03	0,03	0,02	0,03	0,04	0,51
cep	0,01	0,05	0,04	0,03	0,03	0,03	0,05	0,04	-	-	0,01	0,05	0,03	-	0,03	0,40
datanascimentocondutor	0,03	0,02	0,03	0,01	0,07	0,05	0,07	0,04	0,10	0,14	0,02	0,15	0,03	0,07	0,09	0,92
datacarteirahabilitacao	0,05	0,04	0,04	0,06	0,03	0,02	0,07	0,05	0,11	0,03	0,03	0,01	0,04	0,04	0,03	0,65
anomodelo	0,10	0,02	0,03	0,02	0,06	0,07	0,13	0,11	0,10	0,11	0,07	0,09	0,06	0,08	0,03	1,08
fipe	0,06	0,04	0,02	0,04	0,08	0,15	0,05	0,06	-	0,04	0,08	-	0,09	0,10	0,08	0,90
preco	0,29	0,20	0,22	0,25	0,23	0,28	0,23	0,32	0,44	0,30	0,31	0,31	0,30	0,33	0,20	4,22
cpf	0,02	0,02	0,00	0,01	0,02	0,04	-	-	-	-	0,01	0,03	0,01	0,04	0,01	0,21
question10	0,01	0,01	0,02	0,03	0,01	0,04	0,01	0,02	-	0,04	0,01	0,01	0,00	0,03	0,04	0,27
question11	0,03	0,03	0,03	0,01	0,04	-	0,01	-	-	0,06	0,02	-	0,01	0,04	0,01	0,29
question12	0,01	0,01	0,02	0,01	0,01	-	-	0,05	-	0,02	0,01	-	0,00	-	0,02	0,15
question13	0,02	0,04	0,04	0,04	0,02	-	-	0,07	-	-	0,04	0,03	0,09	-	0,05	0,44
question14	0,03	0,04	0,04	0,06	0,04	0,10	-	0,05	-	0,03	0,07	-	0,06	0,05	0,05	0,62
question15	0,04	0,05	0,07	0,06	0,03	0,03	0,03	-	-	0,02	0,03	0,02	0,01	0,05	0,01	0,44
question16	0,03	0,02	0,04	0,02	0,03	-	-	-	-	-	0,03	-	0,03	-	0,03	0,22
question18	0,05	0,01	0,04	0,02	0,00	-	0,03	0,01	0,01	-	0,03	-	0,01	-	0,02	0,24

O segundo método utilizado foi a SS para as características do modelo e posteriormente o método *Importance* do SVM para ranqueá-las. A Tabela 2 sumariza o peso de cada característica por seguradoras usando SS. Elas estão ordenadas da mais importante para a menos importante, segundo o SS.

Para o método SS o ranqueamento das características ficou definido da sequência do maior para o menor como: “Preco, anomodelo, bonus, datanascimentocondutor, danos, datacarteirahabilitacao, question15, sexo, cep, vidros, cpf, fipe, sinistro, tipocarroreserva, question10, question11, question16, franquia, guincho, question13, carroreserva, question14, question12, question18”.

Tabela 2 – Importância das características usando SS

Características	Seguradoras															Soma
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
danos	-	0,12	0,06	0,15	-	-	0,01	-	0,06	0,05	-	0,01	-	-	0,06	0,53
franquia	-	0,06	-	-	-	-	-	-	-	-	0,02	0,05	-	-	0,01	0,15
bonus	0,04	0,10	0,05	0,08	0,06	0,09	0,11	0,09	0,11	0,09	0,06	0,09	0,12	0,05	0,07	1,21
sinistro	0,02	-	-	-	0,10	-	-	0,08	-	-	-	0,04	-	0,07	-	0,31
guincho	-	0,05	-	-	0,01	-	0,05	0,00	-	-	-	-	-	0,02	-	0,13
vidros	-	0,06	0,01	0,05	-	-	0,05	-	0,01	-	-	0,08	0,02	-	0,06	0,35
carroreserva	-	-	-	-	-	-	0,08	-	-	0,01	-	-	0,01	-	-	0,10
tipocarroreserva	-	-	0,09	0,02	-	0,04	0,11	-	0,02	-	-	-	-	-	0,02	0,30
sexo	0,04	0,05	0,07	0,02	0,04	0,04	-	0,03	-	-	0,05	0,04	0,01	0,03	0,03	0,44
cep	-	-	-	0,02	0,03	0,01	0,07	-	0,01	0,05	0,02	0,03	0,05	0,06	-	0,36
datanascimentocondutor	0,04	0,10	0,05	0,07	0,03	0,09	-	0,07	0,07	0,15	0,03	0,16	0,06	0,07	0,06	1,05
datacarteirahabilitacao	0,04	0,08	0,09	0,07	0,02	0,05	-	-	0,03	0,03	0,03	-	-	-	0,05	0,50
anomodelo	0,16	0,08	0,11	-	0,16	0,11	0,15	0,11	0,14	0,14	0,13	0,07	0,14	0,08	0,15	1,72
fipe	-	-	-	-	-	-	0,03	0,03	-	-	0,14	-	-	0,13	-	0,32
preco	0,48	0,29	0,34	0,41	0,45	0,46	0,27	0,49	0,39	0,31	0,41	0,34	0,45	0,39	0,33	5,81
cpf	-	-	0,06	0,02	-	-	0,05	0,01	-	0,07	-	0,04	0,03	0,06	-	0,34
question10	0,01	0,01	-	-	0,03	0,03	-	-	-	-	0,01	-	0,03	-	0,10	0,22
question11	0,04	-	0,03	-	-	-	-	0,01	0,05	0,05	0,02	-	-	-	-	0,20
question12	-	0,01	-	0,03	-	0,01	0,03	-	-	-	-	-	-	-	-	0,07
question13	0,04	-	-	-	-	-	-	-	-	-	-	0,05	-	0,03	-	0,11
question14	-	-	0,03	-	0,03	-	-	-	-	0,03	-	-	-	-	-	0,08
question15	0,04	-	-	0,06	0,03	0,02	-	0,03	0,08	0,02	0,07	-	0,04	0,02	0,06	0,48
question16	0,05	-	-	-	-	0,04	-	0,06	-	-	-	-	0,03	-	-	0,19
question18	-	-	-	-	-	-	-	-	0,02	-	-	-	-	-	-	0,02

O terceiro método utilizado foi o AG para definir as características do modelo e o *Importance* do SVM para ranqueamento em ordem de importância. A função objetivo buscou minimizar o MSE. O AG utilizou uma codificação binária para o vetor de posições das características e 80 indivíduos para o tamanho da população. As condições de parada foram um número máximo de 10 iterações sem melhoria do MSE ou o número máximo de 200 iterações. Na Tabela 3 é possível verificar o peso de cada característica por seguradoras. Elas estão ordenadas da mais importante para a menos importante.

Para o método utilizando o AG o ranqueamento das características ficou definido da sequência do maior para o menor como: “Preco, anomodelo, bonus, datanascimentocondutor, datacarteirahabilitacao, fipe, cep, vidros, question14, sexo, carroreserva, question10, cpf, question15, question13, question11, guincho, question12, sinistro, franquia, danos, tipocarroreserva, question18, question16” .

Tabela 3 – Importância das características usando AG

Características	Seguradoras															Soma
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
danos	0,01	-	-	-	-	-	0,03	-	0,05	-	-	0,03	-	-	0,03	0,15
franquia	0,02	-	-	0,04	-	-	-	-	-	-	0,01	0,05	0,02	-	0,01	0,15
bonus	0,07	0,08	0,03	0,07	0,09	0,11	0,08	0,09	0,10	0,09	0,04	0,14	0,05	-	0,08	1,13
sinistro	0,03	-	-	-	-	-	-	-	-	-	-	0,06	-	0,07	-	0,17
guincho	-	0,02	0,02	0,02	0,02	0,04	0,06	-	-	0,02	-	-	0,01	0,01	0,04	0,25
vidros	0,03	0,06	0,06	0,06	-	-	0,09	-	-	-	-	0,07	-	-	0,07	0,43
carroreserva	0,00	0,11	0,08	0,09	0,01	-	-	-	-	-	-	0,02	0,03	0,02	0,01	0,39
tipocarroreserva	-	-	-	-	-	-	0,08	-	-	0,04	-	-	-	-	-	0,11
sexo	0,04	0,04	0,04	0,04	-	-	-	0,04	0,05	-	0,03	0,03	-	0,02	0,06	0,39
cep	0,03	0,03	0,01	0,03	0,04	0,04	0,03	0,05	0,01	0,01	-	0,08	0,07	-	0,02	0,46
datanascimentocondutor	0,04	0,07	0,03	0,00	0,05	0,05	-	0,07	0,07	0,15	0,07	-	0,03	0,11	0,13	0,88
datacarteirahabilitacao	0,03	0,08	0,08	0,06	-	0,03	-	0,07	0,07	-	0,03	-	0,04	0,05	0,02	0,57
anomodelo	0,18	0,09	0,06	-	0,14	0,11	0,21	0,13	0,14	0,16	0,12	0,06	0,11	0,07	0,10	1,69
fipe	-	-	0,04	0,08	-	-	-	-	0,06	-	0,13	-	0,11	0,15	-	0,56
preco	0,39	0,22	0,31	0,26	0,36	0,49	0,35	0,44	0,36	0,36	0,44	0,34	0,45	0,43	0,28	5,46
cpf	0,04	0,02	0,03	-	0,06	-	0,03	-	-	0,05	0,02	-	0,04	-	0,02	0,30
question10	-	0,01	0,07	0,06	0,04	0,04	0,03	-	0,02	-	-	-	0,03	-	0,05	0,36
question11	0,02	0,05	-	0,03	0,05	0,02	0,01	-	-	0,05	-	-	0,02	-	0,00	0,26
question12	0,01	0,04	-	0,02	-	0,02	-	0,05	0,06	0,03	0,01	-	-	-	-	0,23
question13	-	-	-	-	0,06	-	-	0,07	-	-	-	0,08	-	0,03	0,04	0,28
question14	-	0,03	0,09	0,10	-	0,04	-	-	-	0,03	0,09	0,02	-	-	-	0,40
question15	0,06	0,05	-	-	0,09	-	-	-	-	0,01	0,02	-	-	0,03	0,04	0,30
question16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0,00
question18	-	-	0,06	0,04	-	-	-	-	-	-	-	-	-	0,01	-	0,11

O quarto e último método utilizado foi o RF. A Tabela 4 sumariza a importância e o peso de cada característica pelo método.

Para o método RF o ranqueamento das características ficou definido da sequência do maior para o menor como: “Preco, bonus, datanascimentocondutor, fipe, anomodelo, cep, datacarteirahabilitacao, question16, cpf, question12, vidros, sexo, question15, question14, question11, question10, guincho, tipocarroreserva, carroreserva, danos, franquia, sinistro, question18, question13”.

Tabela 4 – Importância das características usando RF

Características	Seguradoras															Soma
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
danos	0,02	0,03	0,03	0,03	0,02	0,01	0,02	0,02	0,02	0,01	0,02	0,02	0,01	0,02	0,02	0,31
franquia	0,01	0,02	0,02	0,02	0,01	0,03	0,01	0,01	0,02	0,03	0,02	0,04	0,01	0,01	0,02	0,28
bonus	0,06	0,07	0,07	0,08	0,10	0,13	0,06	0,13	0,13	0,10	0,10	0,08	0,11	0,08	0,08	1,37
sinistro	0,01	0,02	0,03	0,01	0,04	0,01	0,01	0,02	0,01	0,02	0,02	0,03	0,02	0,01	0,02	0,27
guincho	0,03	0,03	0,03	0,02	0,02	0,01	0,02	0,03	0,01	0,04	0,02	0,03	0,02	0,01	0,02	0,35
vidros	0,03	0,03	0,04	0,04	0,02	0,03	0,06	0,02	0,02	0,04	0,02	0,03	0,02	0,03	0,03	0,45
carroreserva	0,02	0,04	0,04	0,03	0,02	0,02	0,03	0,01	0,02	0,02	0,02	0,02	0,02	0,00	0,02	0,33
tipocarroreserva	0,02	0,03	0,04	0,04	0,02	0,02	0,04	0,02	0,02	0,02	0,02	0,02	0,02	0,01	0,01	0,35
sexo	0,04	0,02	0,03	0,02	0,03	0,04	0,03	0,04	0,03	0,03	0,03	0,01	0,01	0,04	0,03	0,44
cep	0,05	0,06	0,05	0,05	0,05	0,04	0,07	0,06	0,05	0,06	0,06	0,08	0,06	0,03	0,05	0,83
datanascimentocondutor	0,07	0,08	0,09	0,08	0,08	0,09	0,08	0,05	0,08	0,08	0,08	0,05	0,07	0,08	0,10	1,15
datacarteirahabilitacao	0,06	0,07	0,06	0,06	0,04	0,04	0,04	0,05	0,05	0,04	0,04	0,04	0,04	0,07	0,06	0,75
anomodelo	0,08	0,05	0,05	0,05	0,06	0,07	0,06	0,06	0,07	0,05	0,06	0,06	0,07	0,06	0,05	0,91
fipe	0,06	0,04	0,05	0,04	0,06	0,08	0,05	0,06	0,05	0,06	0,06	0,07	0,07	0,12	0,05	0,92
preco	0,20	0,14	0,12	0,15	0,19	0,18	0,10	0,18	0,18	0,15	0,18	0,16	0,20	0,23	0,13	2,49
cpf	0,03	0,05	0,04	0,04	0,05	0,03	0,05	0,04	0,04	0,04	0,05	0,04	0,05	0,04	0,05	0,64
question10	0,03	0,02	0,02	0,03	0,02	0,02	0,02	0,02	0,03	0,02	0,03	0,03	0,02	0,03	0,03	0,35
question11	0,03	0,03	0,03	0,03	0,02	0,02	0,03	0,02	0,02	0,03	0,03	0,02	0,02	0,02	0,03	0,40
question12	0,02	0,04	0,04	0,04	0,03	0,03	0,04	0,03	0,03	0,03	0,03	0,03	0,02	0,03	0,03	0,47
question13	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,02	0,00	0,01	0,01	0,04	0,03	0,00	0,01	0,16
question14	0,02	0,03	0,03	0,03	0,02	0,02	0,03	0,02	0,03	0,03	0,03	0,02	0,03	0,01	0,03	0,40
question15	0,03	0,03	0,03	0,03	0,02	0,02	0,03	0,03	0,03	0,03	0,03	0,04	0,03	0,02	0,03	0,43
question16	0,03	0,05	0,05	0,05	0,03	0,04	0,11	0,04	0,03	0,04	0,04	0,05	0,04	0,02	0,06	0,70
question18	0,03	0,02	0,02	0,03	0,02	0,01	0,01	0,01	0,02	0,01	0,01	0,00	0,02	0,01	0,02	0,25

Uma vez ordenada todas as características pela sua importância, foi utilizado um comitê entre os métodos utilizados anteriormente para definir o somatório das importâncias de todas as características. Na tabela 5 é possível verificar os pesos de cada método e ao final a coluna “SOMA” exibe o somatório dos pesos dos métodos. O ranqueamento do comitê ficou definido da sequência do maior para o menor como:

“Preco, anomodelo, bonus, datanascimentocondutor, fipe, datacarteirahabilitacao, cep, sexo, question15, vidros, question14, cpf, danos, question10, sinistro, question11, carroreserva, question16, guincho, question13, tipocarroreserva, question12, franquia, question18”.

Tabela 5 – Somatório da importância das características

	Características	SS	RFE	RF	GA	Soma
1	preco	5,81	4,22	2,49	5,46	17,99
2	anomodelo	1,72	1,08	0,91	1,69	5,40
3	bonus	1,21	1,09	1,37	1,13	4,80
4	datanascimentocondutor	1,05	0,92	1,15	0,88	3,99
5	fipe	0,32	0,90	0,92	0,56	2,71
6	datacarteirahabilitacao	0,50	0,65	0,75	0,57	2,47
7	cep	0,36	0,40	0,83	0,46	2,05
8	sexo	0,44	0,51	0,44	0,39	1,78
9	question15	0,48	0,44	0,43	0,30	1,65
10	vidros	0,35	0,38	0,45	0,43	1,61
11	question14	0,08	0,62	0,40	0,40	1,51
12	cpf	0,34	0,21	0,64	0,30	1,50
13	danos	0,53	0,40	0,31	0,15	1,39
14	question10	0,22	0,27	0,35	0,36	1,19
15	sinistro	0,31	0,43	0,27	0,17	1,18
16	question11	0,20	0,29	0,40	0,26	1,15
17	carroreserva	0,10	0,33	0,33	0,39	1,15
18	question16	0,19	0,22	0,70	0,00	1,11
19	guincho	0,13	0,36	0,35	0,25	1,08
20	question13	0,11	0,44	0,16	0,28	1,00
21	tipocarroreserva	0,30	0,18	0,35	0,11	0,94
22	question12	0,07	0,15	0,47	0,23	0,91
23	franquia	0,15	0,26	0,28	0,15	0,83
24	question18	0,02	0,24	0,25	0,11	0,61

4.4 Resultado da Definição da Quantidade Mínima de Características

Para definir a quantidade mínima de características, inseriu-se ao modelo proposto as características por ordem de importância de acordo com seu impacto no MSE. Percebe-se

que iniciando com a característica “*preço*” e incluindo as características “*anomodelo*, *bonus* e *datanascimentocondutor*” reduz-se o MSE. Ao inserir a quinta característica, “*fipe*”, há um pequeno aumento no erro indicando um momento de parada. Na sexta característica, “*datacarteirahabilitacao*”, a curva do MSE volta a reduzir de forma considerável. De forma semelhante, o mesmo ocorre com a sétima característica que é “*CEP*”. Por outro lado, a inclusão da oitava característica, “*sexo*”, gerou um acréscimo do MSE. Neste ponto percebeu-se um “joelho” na curva, caracterizando o critério de parada com a escolha das sete características. A Figura 14 ilustra este caso.

Portanto, definiu-se sete variáveis para a utilização do modelo. Este número leva em conta a minimização do RMSE e a importância das variáveis de acordo com os critérios antes adotados.

As sete variáveis são listadas abaixo:

- a) preço;
- b) anomodelo;
- c) bonus;
- d) datanascimentocondutor;
- e) fipe;
- f) datacarteirahabilitacao;
- g) cep.

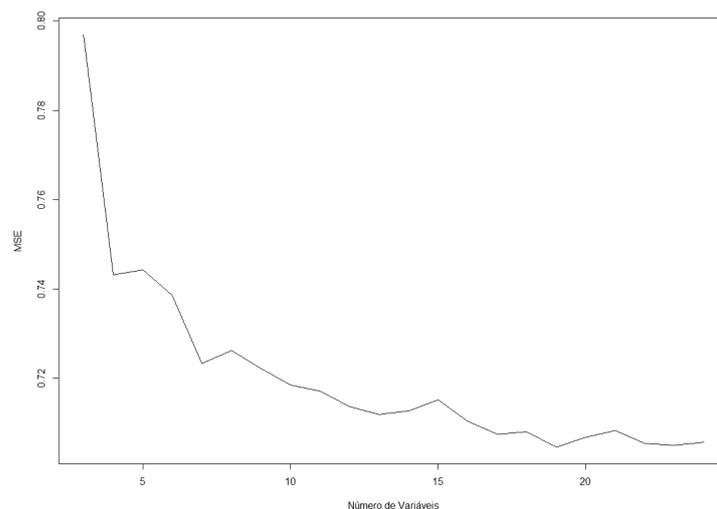


Figura 14 – Erro Quadrático Médio vs Número de Variáveis

4.4.1 Resolução do Problema Multiobjetivo: MSE e Quantidade de Variáveis

A fim de validar a escolha do número de variáveis escolhida, utilizou-se o algoritmo NSGA-II partindo da premissa que os dois objetivos MSE e quantidade de variáveis são conflitantes. Contudo, esta premissa nem sempre é verdadeira, uma vez que é possível encontrar situações onde diminuir o número de variáveis pode melhorar o MSE e, conseqüentemente, o modelo.

De toda forma, na maioria das vezes essa relação é dada como conflitante. Por isso, esse trabalho avaliou os dois objetivos com o NSGA-II. Utilizou-se uma população de 100 indivíduos e 100 gerações para gerar a PF exibida na Figura 15. Neste caso, as sete variáveis aparecem também na PF ótima, o que permite reafirmar a boa solução obtida. Existem também outros pontos na PF com números maiores de variáveis, porém contradiz a ideia adotada no trabalho de reduzir o número de variáveis.

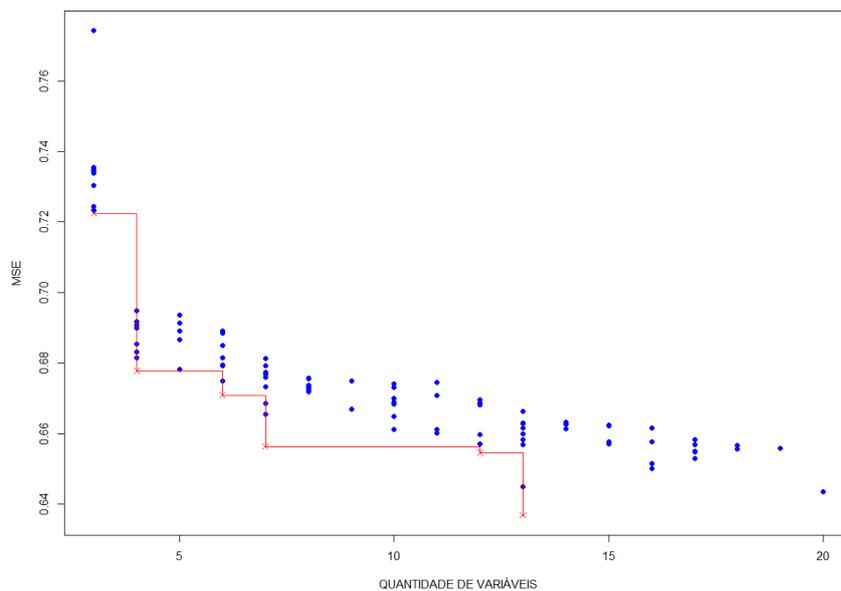


Figura 15 – Erro Quadrático Médio vs Número de Variáveis obtido com o NSGA II

4.5 Resultado da Previsão do Valor e dos Quantis do Seguro do Automóvel

O resultado da geração via *bootstrap* de 1000 valores de seguros esperados para um único perfil de dados inseridos no aplicativo é apresentado na Figura 16. Essa distribuição se assemelha muito com a uma distribuição normal, permitindo utilizar técnicas estatísticas como intervalo de confiança.

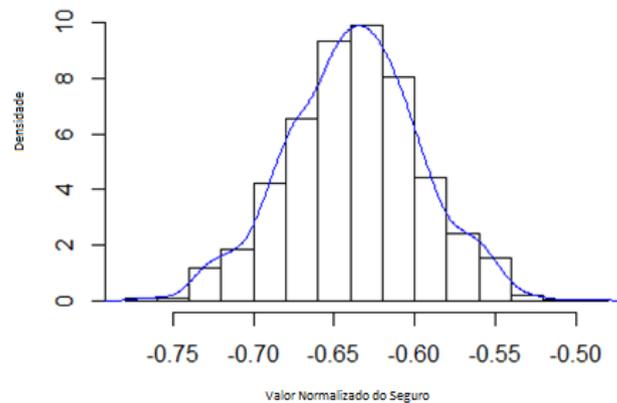


Figura 16 – Geração de 1000 respostas do SVM para o mesmo perfil de usuário

Para exemplificar uma solução, utilizou-se um perfil real cadastrado no aplicativo. O valor das variáveis estão descritos abaixo:

- a) preco: 55551;
- b) anomodelo: 2014;
- c) bonus: 4;
- d) datanascimento: 293155200000 (17/04/1979);
- e) fipec: 014 (Honda);
- f) datacarteirahabilitacao: 8.62272e+11 (29/04/1997)
- g) cep: 36.

O valor estimado do seguro para estes dados e os respectivos quantis podem ser visualizados na Figura 17. Os quantis estão representados pelos pontos azuis, o valor estimado em verde e o valor real em vermelho, quando o mesmo encontra-se dentro da região do gráfico plotado. Estes dados estão sumarizados na Tabela 6.

4.6 Comparação com Benchmarking

Para avaliar as soluções obtidas com a abordagem proposta, fez-se a comparação do mesmo com a SVM tomando como base a função de custo desenvolvida por [Steinwart et al. \[2011\]](#). A base de dados foi dividida em 70% treino e 30% teste. As 24 variáveis descritas em 3.2 foram utilizadas como entrada (*input*) e como saída (*output*) foi gerada a predição do valor do seguro e os quantis. Em ambas as abordagens utilizou-se validação cruzada para encontrar os melhores parâmetros, de acordo com a Figura 18.

Tabela 6 – Exemplificação de uma solução obtida por seguradora

	Valor do Quantil 1	Valor do Quantil 2	Valor Predição do Seguro	Valor Real do Seguro
Seguradora 1	R\$ 2.667,16	R\$ 3.581,28	R\$ 3.195,87	R\$ 2.719,29
Seguradora 2	R\$ 2.391,61	R\$ 4.381,97	R\$ 3.340,51	R\$ 2.509,18
Seguradora 3	R\$ 2.557,27	R\$ 4.403,27	R\$ 3.562,19	R\$ 2.952,64
Seguradora 4	R\$ 2.981,94	R\$ 7.792,64	R\$ 5.033,09	R\$ 3.411,30
Seguradora 5	R\$ 3.222,17	R\$ 4.174,36	R\$ 3.481,41	R\$ 3.696,32
Seguradora 6	R\$ 1.996,34	R\$ 2.460,28	R\$ 2.316,98	R\$ 2.455,35
Seguradora 7	R\$ 1.945,32	R\$ 2.491,39	R\$ 2.105,13	R\$ 2.192,64
Seguradora 8	R\$ 3.078,70	R\$ 4.121,71	R\$ 3.576,26	R\$ 3.110,38
Seguradora 9	R\$ 1.938,99	R\$ 2.835,16	R\$ 2.172,30	R\$ 2.717,13
Seguradora 10	R\$ 1.973,53	R\$ 2.464,86	R\$ 2.124,85	R\$ 2.174,64
Seguradora 11	R\$ 2.037,07	R\$ 2.734,32	R\$ 2.361,82	R\$ 2.181,90
Seguradora 12	R\$ 2.238,77	R\$ 4.311,98	R\$ 3.270,63	R\$ 3.466,07
Seguradora 13	R\$ 2.705,19	R\$ 3.657,17	R\$ 3.164,85	R\$ 3.535,22
Seguradora 14	R\$ 2.054,89	R\$ 4.377,24	R\$ 2.662,11	R\$ 2.083,18
Seguradora 15	R\$ 1.184,40	R\$ 2.624,94	R\$ 1.473,80	R\$ 2.018,88

Toda a implementação foi realizada no software *R Team* [2013]. Utilizou-se os seguintes pacotes com os algoritmos já implementados: *emphcaret*, *funModeling*, *tidyverse*, *KERE*, *rminer*, *e1071*, *xlsx* e *qrsvm*. Três indicadores foram utilizados como métricas de comparação na base de teste:

1. Variância dos métodos frente ao valor estimado, Equação (4.1):

$$\text{var1}(X) = E((X - \bar{Y}))^2 \quad (4.1)$$

onde *var1* representa a diferença dos valores encontrados pela função dos dois quantis em relação ao valor estimado calculado pelo preditor.

2. Variância dos métodos frente ao valor Real, Equação (4.2):

$$\text{var2}(X) = E((X - Y))^2 \quad (4.2)$$

onde a *var2* é a diferença dos valores encontrados pela função dos dois quantis em relação ao valor real.

3. Percentual do valor estimado encontrado dentro dos quantis com valores de 0,05 e 0,95. O resultado representa o percentual dos valores de predição de seguro que estão inseridos nos intervalos dos valores de predição dos quantis.

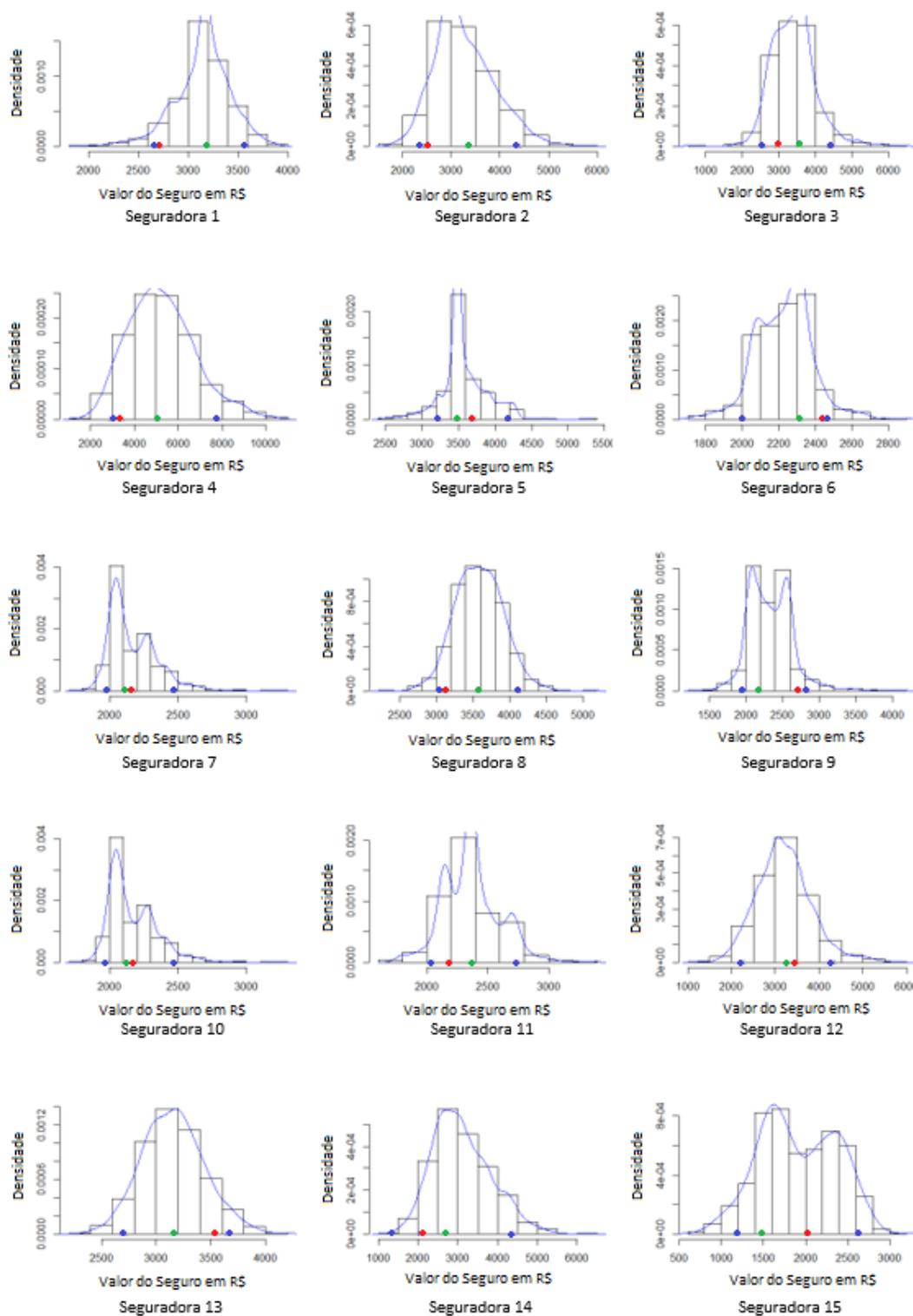


Figura 17 – Exemplo do Resultado de um Perfil em Todas as Seguradoras

Por meio da Tabela 7 é possível verificar que o método proposto apresenta os menores valores e, portanto, é melhor em relação a variância do valor estimado em todas as seguradoras. O mesmo ocorre para o percentual do valor estimado. Para o quesito de

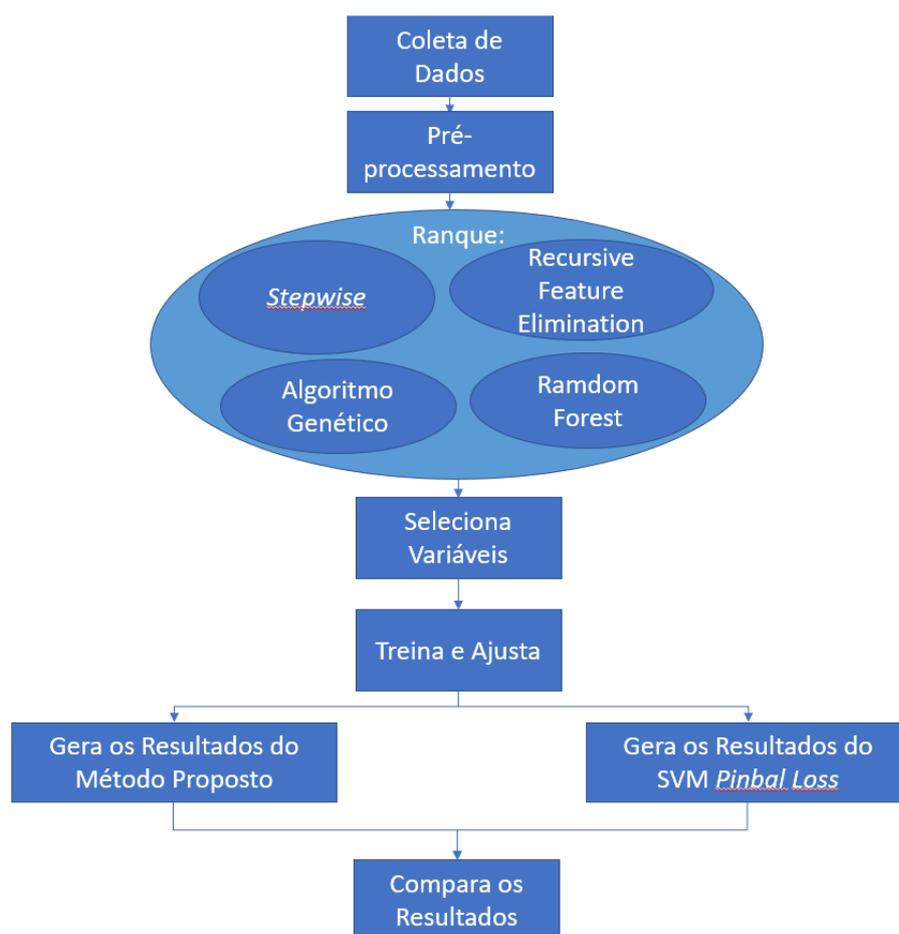


Figura 18 – Estratégia adotada para comparação do método proposto e o SVM

variância frente ao valor real, o método proposto é superior ao método SVM com *Pinbal loss* para quase todas as seguradoras, com exceção da de número 10, conforme ilustrado nas Figuras 19, 20 e 21. Foi plotada a variação do valor do seguro com os quantis do método proposto versus o método SVM com *Pinbal loss* em relação ao preço, data de nascimento e data da primeira habilitação do condutor, respectivamente. O resultado do método proposto está representado pela linha vermelha e o SVM em verde.

Tabela 7 – Método Proposto vs SVM com *Pinbal loss*

	Variância MP em relação ao Valor Estimado	Variância SVM (PL) em relação ao Valor Estimado	Variância MP em relação ao Valor Real	Variância SVM (PL) em relação ao Valor Real	% MP	% SVM (PL)
Seguradora 1	188,411	362,5986	403,3522	474,7792	100%	68%
Seguradora 2	250,3944	661,8627	674,043	788,9681	100%	69%
Seguradora 3	226,67	585,901	520,1666	668,2412	100%	75%
Seguradora 4	259,525	608,8412	658,6183	746,9538	100%	69%
Seguradora 5	135,8247	732,8523	354,4862	790,2249	100%	87%
Seguradora 6	180,1498	457,3811	430,2859	549,7789	100%	71%
Seguradora 7	114,1727	475,5761	338,8769	532,0221	100%	94%
Seguradora 8	114,1727	475,5761	338,8769	532,0221	100%	94%
Seguradora 9	208,8781	440,5739	456,8744	554,6974	100%	71%
Seguradora 10	109,0249	173,5145	217,9146	215,7444	100%	61%
Seguradora 11	236,3717	624,4141	613,2388	765,6924	100%	72%
Seguradora 12	110,687	378,1353	245,8038	456,7267	100%	90%
Seguradora 13	159,7168	386,6707	365,209	432,5563	100%	61%
Seguradora 14	87,68632	155,5392	175,8428	209,2958	100%	50%
Seguradora 15	124,1368	263,0547	294,7067	341,8291	100%	65%

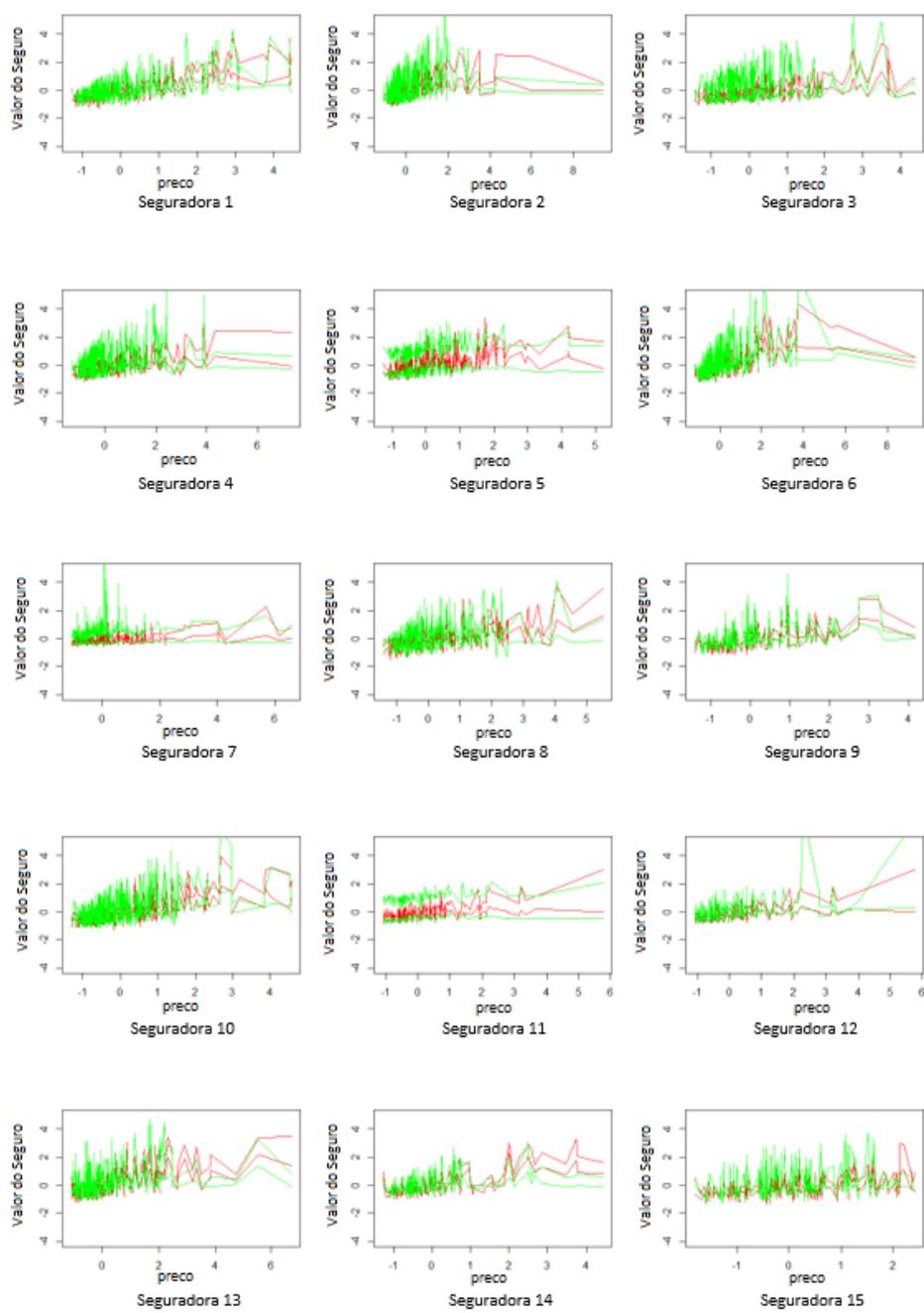


Figura 19 – Método Proposto (Vermelho) vs *SVM Pinbal Loss* (Verde) com a variável preço por seguradora

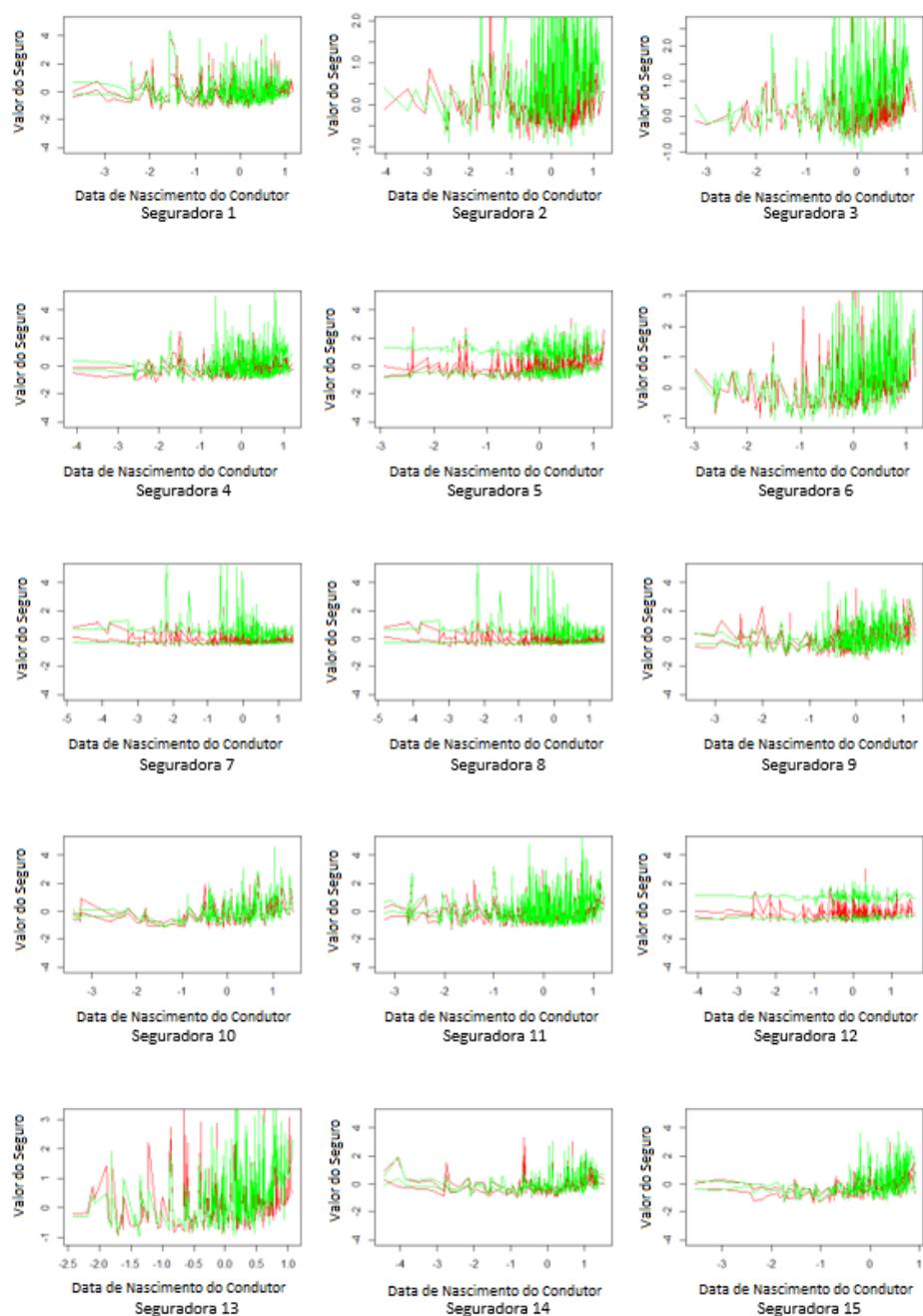


Figura 20 – Método Proposto (Vermelho) X *SVM Pinbal Loss* (Verde) com a variável datanascimento por seguradora

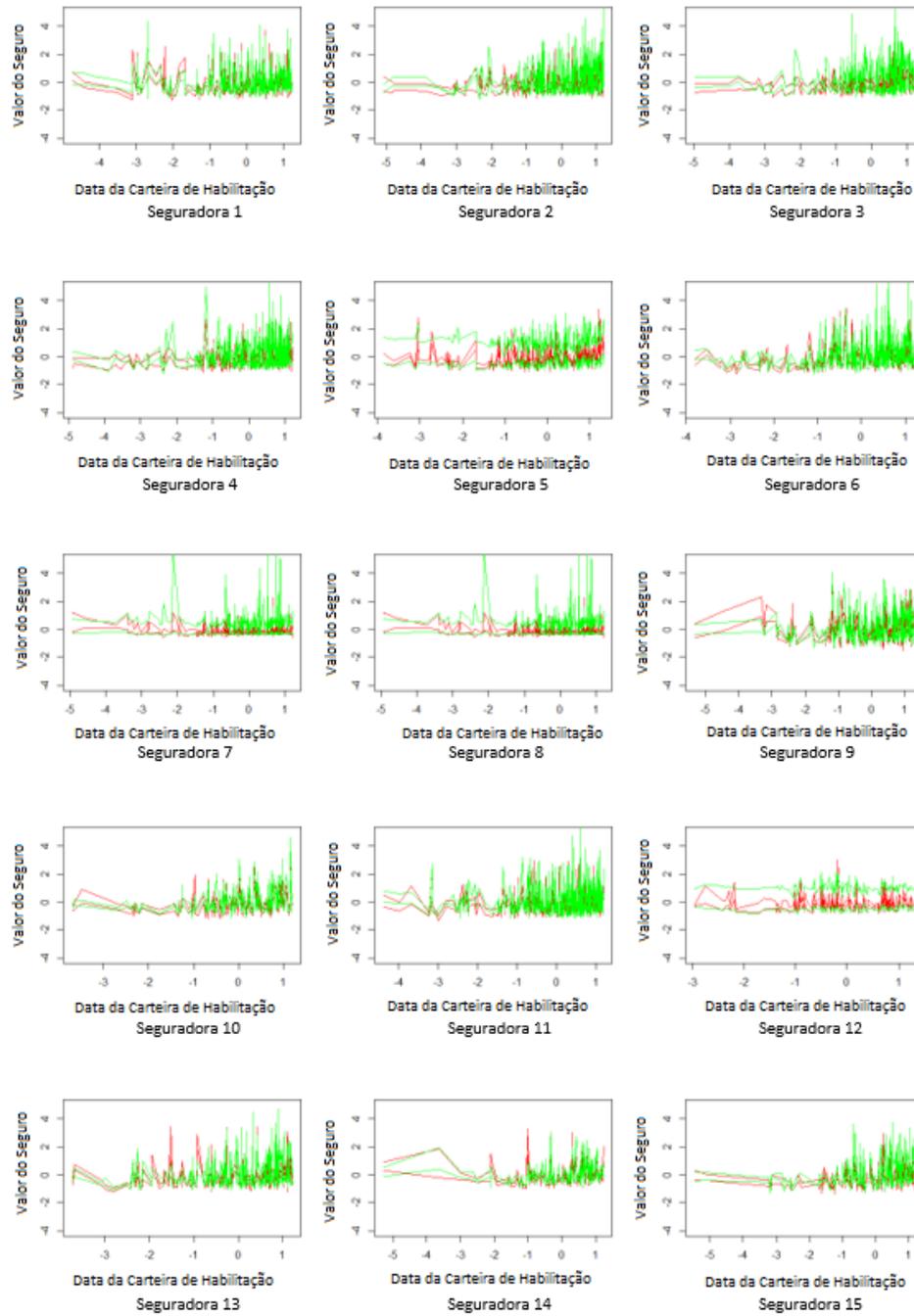


Figura 21 – Método Proposto (Vermelho) X *SVM Pinbal Loss* (Verde) com a variável *datacarteirahabilitacao* por seguradora

Capítulo 5

Conclusão e Trabalhos Futuros

5.1 Conclusão

Este trabalho propôs a predição do valor do seguro de automóveis no Brasil utilizando técnicas de inteligência computacional. O método proposto faz a coleta de dados do perfil do segurado através de um aplicativo desenvolvido para celulares disponibilizado para *Android* e *iOS*. O sistema pré-processa os dados por meio de padronização. A importância das características do modelo é ranqueada por métodos *wrappers*, embutidos e filtros. O número ideal de características é selecionado com base no MSE da regressão de todas as seguradoras. O valor estimado do preço do seguro é feito utilizando o SVM. Ao final, calcula-se os quantis desejados por perfil do segurado utilizando *bootstrap*.

Os resultados alcançados nos dados coletados pelo aplicativo mostraram-se superiores ao SVM com a função de custo *Pinbal loss*. Os critérios utilizados para esta avaliação foram a variância em relação ao valor estimado e ao valor real e o percentual do valor estimado encontrado dentro dos quantis. Durante os experimentos, o método foi superior para todas as seguradoras nos critérios variância em relação ao valor estimado e no percentual do valor estimado encontrado dentro dos quantis definidos. Em relação a variância comparada ao valor real, a abordagem mostrou-se superior ao SVM com *Pinbal loss* para quase todas as seguradoras, com exceção de uma delas.

A metodologia apresentada neste trabalho pode influenciar diretamente na maneira atual como o seguro de automóveis funciona no Brasil. O modelo atual desse mercado mostra-se muito ultrapassado e com quase toda sua cadeia de processos feita de forma manual, o que resulta em um negócio oneroso e lento. Com o uso da abordagem proposta pode-se melhorar o modelo atual trazendo agilidade com a automação de várias partes do processo, do momento da cotação à contratação da apólice do seguro.

5.2 Trabalhos Futuros

O modelo proposto faz o cálculo do valor do seguro e dos quantis do segurado levando em consideração uma seguradora por vez, fazendo esse processo para as quinze seguradoras. Para cada seguradora o modelo gerado leva em consideração apenas os dados da seguradora em questão e não utiliza vários outros dados das outras seguradoras. Sugere-se em trabalhos futuros a utilização dos dados das outras seguradoras, por meio de aprendizado semi-supervisionado, que alguns dados poderiam ficar sem rótulo para melhor estimar o valor do seguro do automóvel por seguradora.

Além do valor do seguro, um outro ponto importante para a decisão do usuário final na escolha do seguro do automóvel é o valor da franquia. A franquia é a participação obrigatória do segurado na reparação de seu veículo no caso de perda parcial. Uma outra sugestão de trabalho futuro é a geração de uma predição do valor da franquia de forma muito semelhante a realizada neste trabalho com o valor do seguro. Desta forma, o usuário final recebe o valor do seguro e o valor da franquia do automóvel para uma possível decisão de contratação do serviço.

Referências

- A. A. Abdoos, P. K. Mianaei, and M. R. Ghadikolaei. Combined vmd-svm based feature selection method for classification of power quality events. *Applied Soft Computing*, 38: 637–646, 2016.
- S. F. Adra. *Improving convergence, diversity and pertinency in multiobjective optimisation*. PhD thesis, University of Sheffield, Department of Automatic Control and Systems Engineering, 2007.
- V. Alberti and M. A. P. Leopoldi. *Entre a solidariedade e o risco: história do seguro privado no Brasil*. Fundação Escola Nacional de Seguros, 2001.
- J. Apolloni, G. Leguizamón, and E. Alba. Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. *Applied Soft Computing*, 38:922–932, 2016.
- B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ACM Press:pp. 144–152, 1992.
- L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- G. Casella, S. Fienberg, and I. Olkin. Springer texts in statistics. *Springer*, 2013.
- C. Chatfield. Calculating interval forecasts. *Journal of Business & Economic Statistics*, 11(2):121–135, 1993.
- C. Chatfield. Prediction intervals for time-series forecasting. In *Principles of forecasting*, pages 475–494. Springer, 2001.
- CNSEG. Confederação nacional das empresas de seguros gerais, previdência privada e vida, saúde suplementar e capitalização-cnseg, 2019. URL <http://cnseg.org.br/>.
- C. Coello, S. De Computación, and C. Zacetenco. Twenty years of evolutionary multi-objective optimization: A historical view of the field. *IEEE computational intelligence magazine*, 1(1):28–36, 2006.

- C. A. C. Coello, G. B. Lamont, D. A. Van Veldhuizen, et al. *Evolutionary algorithms for solving multi-objective problems*, volume 5. Springer, 2007.
- P. Cortez and M. J. Embrechts. Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, 225:1–17, 2013.
- C. Darwin. *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*. J. Murray, 1859.
- M. David. Auto insurance premium calculation using generalized linear models. *Procedia Economics and Finance*, 20:147–156, 2015.
- K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002.
- K. Deb, L. Thiele, M. Laumanns, and E. Zitzler. Scalable test problems for evolutionary multiobjective optimization. In *Evolutionary multiobjective optimization*, pages 105–145. Springer, 2005.
- T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statistical science*, pages 189–212, 1996.
- B. Efron. Bootstrap methods: another look at the jackknife *annals of statistics* 7: 1–26. [View Article PubMed/NCBI Google Scholar](#), 1979.
- A. E. Eiben, J. E. Smith, et al. *Introduction to evolutionary computing*, volume 53. Springer, 2003.
- A. P. Engelbrecht. *Computational intelligence: an introduction*. John Wiley & Sons, 2007.
- I. Faria. Estrutura competitiva do mercado segurador brasileiro. *Revista Brasileira de Seguros*, 2(4):17–41, 2007.
- F. Galiza. Economia e seguro: uma introdução. *Rio de Janeiro: Funenseg*, 1997.
- J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami. Internet of things (iot): A vision, architectural elements, and future directions. *Future generation computer systems*, 29(7):1645–1660, 2013.
- J. Han, M. Kamber, and J. Pei. Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, pages 83–124, 2011.
- B. E. Hansen. Interval forecasts and parameter uncertainty. *Journal of Econometrics*, 135(1-2):377–398, 2006.
- R. Henckaerts, M.-P. Côté, K. Antonio, and R. Verbelen. Boosting insights in insurance tariff plans with tree-based machine learning. *arXiv preprint arXiv:1904.10890*, 2019.

- R. J. Hyndman and Y. Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365, 1996.
- G. James, D. Witten, T. Hastie, and R. Tibshirani. An introduction to statistical learning: with applications in r. *Springer Publishing Company, Incorporated*, 2014.
- E. Krasheninnikova, J. García, R. Maestre, and F. Fernández. Reinforcement learning for pricing strategy optimization in the insurance industry. *Engineering Applications of Artificial Intelligence*, 80:8–19, 2019.
- K. Li, K. Deb, Q. Zhang, and S. Kwong. An evolutionary many-objective optimization algorithm based on dominance and decomposition. *IEEE Transactions on Evolutionary Computation*, 19(5):694–716, 2014.
- S. Maldonado, R. Weber, and F. Famili. Feature selection for high-dimensional class-imbalanced data sets using support vector machines. *Information Sciences*, 286:228–246, 2014.
- I. R. Meneghini, M. A. Alves, A. Gaspar-Cunha, and F. G. Guimarães. Scalable and customizable benchmark problems for many-objective optimization. *Applied Soft Computing*, pages 106–139, 2020.
- B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, 10(1):213, 2009.
- M. Ridley. *Evolution*, malden ma, 2004.
- G. Roffo and S. Melzi. Features selection via eigenvector centrality. *Proceedings of new frontiers in mining complex patterns (NFMCP 2016)*, 2016.
- J. Schaffer. Multiple objective optimization with vector evaluated. *Genetic Algorithms, Ph. D. Dissertation, Vanderbilt University*, 1984.
- B. Scholkopf and A. Smola. Learning with kernels, support vector machines, regularization, optimization and beyond. *The MIT Press*, 2004.
- A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.
- N. Srinivas and K. Deb. Multiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary computation*, 2(3):221–248, 1994.
- I. Steinwart, A. Christmann, et al. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

- SUSEP. Autoseg. disponível em:< <http://www.susep.gov.br>> , superintendência, de seguros privados. *Acesso em*, 12, 2019.
- R. C. Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*, URL <http://www.R-project.org/>, 2013.
- V. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks* 10, no. 5:988–999, 1999.
- Y. Wang and W. Xu. Leveraging deep learning with lda-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, 105:87–95, 2018.
- E. W. Weisstein. Central limit theorem. *From MathWorld-A Wolfram Web Resource*, 2011.
- Y. Yang, W. Qian, and H. Zou. Insurance premium prediction via gradient tree-boosted tweedie compound poisson models. *Journal of Business & Economic Statistics*, 36(3): 456–470, 2018.
- Q. Yu, Y. Miche, E. Eirola, M. Van Heeswijk, E. SéVerin, and A. Lendasse. Regularized extreme learning machine for regression with missing data. *Neurocomputing*, 102:45–51, 2013.
- Q. Zhang and H. Li. Moea/d: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on evolutionary computation*, 11(6):712–731, 2007.

Anexo A

Glossário do Seguro de Automóveis

1. Segurado: Pessoa física ou jurídica que contrata um seguro em benefício próprio ou de terceiros.
2. Responsabilidade Civil Facultativa de Veículos (RCF-V): Trata-se de uma cobertura opcional com garantia de pagamento de indenização até o limite previsto na apólice. Tal cobertura é acionada no caso de ocorrência de danos materiais e/ou corporais causados a terceiros pelo veículo segurado.
3. Acidentes Pessoais de Passageiros (APP): Trata-se de uma cobertura opcional contratada em conjunto com uma cobertura básica de seguro de automóveis. Tal cobertura permitirá o pagamento de indenização ao passageiro do veículo (ou a seus beneficiários) que se tornaram vítima de acidente de trânsito com o veículo segurado.
4. Assistência 24 horas: Trata-se de uma cobertura opcional muito comum que pode ser encontrado tanto em planos mais simples como em planos mais complexos. Os planos básicos oferecem nesse modelo os serviços como reboque do veículo e chaveiro. Por outro lado, os planos mais complexos oferecem outros serviços como reboque do veículo com quilometragem ilimitada.
5. Vidros, Lanternas, Faróis e Retrovisores: Trata-se de uma cobertura opcional de vidros, lanternas, faróis e retrovisores com indenização no caso de sinistro exclusivo. Desta forma o segurado poderá optar pela troca do item danificado em um prestador de serviço ou então se quiser receber o reembolso levando em consideração o valor previsto na condição geral do seguro.
6. Carro Reserva: Trata-se de uma cobertura opcional de fornecimento de um carro reserva ao segurado quando o veículo dele estiver sendo reparado em razão de um sinistro. Vale lembrar que o veículo fornecido será o definido previamente na cobertura contratada.

7. Lucros Cessantes: Trata-se de uma cobertura opcional e destina-se ao segurado que utiliza o veículo com atividade profissional. Neste caso o segurado receberá a indenização se o veículo estiver indisponível para realizar o trabalho dele.
8. Perda Parcial: Trata-se das condições do contrato para o possível reparo de danos parciais causados ao veículo com sinistro, oferecendo a possibilidade de ser reparado em oficinas de livre escolha ou em oficinas pertencentes à rede específica do produto.
9. Indenização Integral: Trata-se de casos onde os prejuízos resultantes de um mesmo sinistro atingirem ou ultrapassarem 75% do valor contratado. Neste caso o seguro permite duas modalidades de pagamento: pelo valor de mercado referenciado ou pelo valor determinado no momento da contratação do seguro.
10. Franquia: Trata-se da parcela de participação do segurado nos prejuízos que ocorreram dos danos parciais do veículo. O valor definido para pagamento do segurado é fixo e deve ser definido no momento da contratação do seguro.
11. Bônus: Trata-se do possível desconto que o segurado poderá receber no momento da renovação de um seguro anual. Esse desconto leva em consideração o histórico de sinistro do segurado.
12. Prêmio: Trata-se do valor que o segurado paga a empresa contratada para exercer o direito ao seguro.
13. Seguro: Trata-se de um contrato onde uma das partes se obriga a indenizar a outra pela ocorrência de alguns eventos previstos em condições contratuais.
14. Sinistro: Trata-se da ocorrência do risco coberto, definido durante o período que rege o contrato.

Anexo B

Aplicativo de Coleta de Dados

Como tela principal o usuário faz um cadastro com usuário e senha, o que lhe permite o acesso às funcionalidades do aplicativo. A Figura 22 mostra a tela de login do aplicativo e como o usuário deve se cadastrar caso esteja efetuando o primeiro acesso.



Figura 22 – Login do Aplicativo.

Ao efetuar o acesso, o usuário tem a opção de fazer uma cotação de seguro de automóveis através do botão com símbolo "+" na tela principal. Ao clicar nessa opção, o aplicativo solicita a inserção de várias informações para que o futuro segurado informe com veracidade. Essa inserção é feita com validação em todos os campos, evitando assim que ao final obtenha-se informações com campos não preenchidos ou dados inválidos. O preenchimento é dividido em quatro páginas, começando com as informações do veículo.

Na primeira página o usuário informa os dados do veículo como:

1. Fabricante: Informação encontrada no CRLV;
2. Modelo: Informação encontrada no CRLV;
3. Ano de Fabricação: Informação encontrada no CRLV;
4. Ano Modelo: Informação encontrada no CRLV;
5. Zero km: é considerado veículo zero km no seguro é necessário que seja apresentado a nota fiscal emitida pelo distribuidor ou pela concessionária autorizada dentro de até 24h após a data de retirada do veículo (essa data consta na nota fiscal).
6. Categoria de uso: A categoria de uso corresponde à atividade fim em que o veículo é utilizado.
7. Financiado/Alienado: Informação encontrada no CRLV;

Todos esses campos podem ser visualizados na Figura 23. Foi gerado uma rotina de importação da tabela da Fundação Instituto de Pesquisas Econômicas (Fipe) evitando assim que o usuário digite o modelo do seu carro, o que melhora em muito a qualidade da inserção da informação.

Uma vez preenchido a página com as informações do veículo, o usuário é direcionado para a segunda página do aplicativo para inserção das informações de uso do veículo como:

1. Início da Vigência: Data do início do seguro do veículo;
2. Renovação: Se o usuário atualmente possui um seguro para o veículo em questão, então isso significa uma renovação. Caso contrário utilize a opção NÃO para renovação;
3. Classe de Bônus: É um indicador de experiência do seguro, que é considerado na composição do preço da renovação da apólice. Anualmente, o bônus aumenta na renovação da apólice.
4. Quantidade de Sinistro: Sinistro de veículo é a ocorrência de um acidente ou evento inesperado e/ou não premeditado com o veículo. Favor preencher a quantidade de Sinistro do seguro que está renovando.

15:18 74%

Informações do Veículo 1 DE 4

Fabricante (i)

Modelo (i)

Fabricação (i) Modelo (i)

Combustível (i) Zero Km? (i)

Código Fipe Preço Médio

Categoria de Uso (i)

Financiado/Alienado? (i)

Chassi remarcado? (i)

Veículo blindado? (i)

Figura 23 – Informações do Veículo.

5. Garagem na Residência: Possui garagem na residência do segurado.
6. Garagem no Trabalho: Possui garagem no local de trabalho do segurado.
7. Garagem no Estudo/Curso: Possui garagem no local de Estudo/Curso do segurado.
8. Utilização do veículo: O segurado utiliza o veículo para uso pessoal ou comercial.
9. Quantidade de Veículos na Residência: Quantos veículos o usuário possui na residência.
10. km Mensal: Quantos quilômetros o segurado percorre em média por mês com o Veículo que será segurado.

As informações de coleta de preenchimento do uso do veículo podem ser visualizadas na Figura 24.

15:18 74%

Informações de Uso 2 DE 4

Início Vigência

Início Vigência

Este veículo já possui seguro?

CEP Pernoite

Garagem na residência

Garagem no trabalho

Garagem no Estudo/Curso

Utilização do veículo

Quantidade de veículos na residencia

Figura 24 – Informações do Uso do Veículo.

Finalizado a inserção dos dados de preenchimento do uso do veículo, o usuário é direcionado para a página de preenchimento das informações dos Condutores. Neste momento ele pode selecionar um condutor pré-cadastrado ou adicionar um novo conforme Figura 25. Ao optar pelo cadastro de um novo condutor o usuário deverá preencher os itens:

1. Nome Completo;
2. E-mail;
3. Cadastro de Pessoa Física (CPF);
4. Número do telefone Celular;
5. Data de Nascimento do Condutor;

6. Sexo do Condutor;
7. Estado Civil;
8. Tipo de Residência;
9. Profissão;
10. Número da Carteira Nacional de Habilitação (CNH);
11. Data da primeira Habilitação.

Preenchido os dados do condutor ou selecionado um condutor pré-cadastrado, o aplicativo retorna para a página de informações de Condutores para o preenchimento dos itens:

1. Titular da Proposta;
2. Condutor Principal;
3. Relação com o Titular da Proposta;
4. Residem com o Condutor principal pessoas de outra faixa etária.

Para finalizar a parte da inserção de dados da cotação, o usuário final é direcionado à última página desse processo, Figura 26, que contempla as inserções relacionadas às coberturas contratadas como:

1. Coberturas adicionais: O usuário pode, ainda, contratar coberturas adicionais ao seguro do seu automóvel. Para cada uma das coberturas adicionais, é cobrado um prêmio e definido um limite máximo de indenização.

Básico: Sem cobertura adicional de vidros, carro reserva e assistência 24 horas.

Intermediário: Vidros - Reparo ou reposição dos vidros laterais, traseiro e para-brisa, em caso de quebra, trinca ou qualquer dano. Carro Reserva – Veículo básico no período 7 de dias. Assistência 24 horas – Quilômetros (km) do reboque limitado a 200 km.

Avançado: Vidros - Reparo ou reposição dos vidros laterais, traseiro, para-brisa, faróis, lanternas e retrovisores, em caso de quebra, trinca ou qualquer dano. Carro Reserva – Veículo executivo no período de 30 dias. Assistência 24 horas – Quilômetros (km) do reboque ilimitado

2. Danos: Ao contratar esse seguro, o usuário passa a ter o direito ao reembolso – até o limite determinado na sua apólice – de indenizações que seja obrigado a

Figura 25 – Cadastro e Informações do Condutor.

(a) Cadastro de Condutor

(b) Informações do Condutor

Nome Completo

Email

CPF Celular

Nascimento Sexo

Estado Civil Tipo de Residência

Profissão

Número CNH 1ª Habilitação

SALVAR

Selecione o titular da proposta

Selecione o condutor principal

Relação com o titular da proposta

Residem com o condutor principal pessoas na faixa etária de (considere sempre o mais jovem)

Selecione...

pagar, judicial ou extrajudicialmente, por ter provocado danos Morais, Materiais e Corporais.

Básico: Na opção de Plano Básico o usuário terá o direito a reembolso de (R\$ 50.000,00), (R\$ 50.000,00) e (R\$ 5.000,00) para danos Morais, Materiais e Corporais respectivamente

Intermediário: Na opção de Plano Básico o usuário terá o direito a reembolso de (R\$ 100.000,00), (R\$ 100.000,00) e (R\$ 10.000,00) para danos Morais, Materiais e Corporais respectivamente.

Avançado: Na opção de Plano Básico o usuário terá o direito a reembolso de (R\$ 200.000,00), R\$ (200.000,00) e (R\$ 10.000,00) para danos Morais, Materiais e Corporais respectivamente.

3. Franquia: A franquia do seguro tem relação inversa com o preço do seguro. Ou seja: se a franquia cai, o preço aumenta. Se a franquia aumenta, o preço cai. Isso

15:18 74%

Tipos de Coberturas 4 DE 4

Reboque/Assistência 24h
Selecione...

Vidros
Selecione...

Carro Reserva (Diárias)
Selecione...

Danos *i*

Morais, Materiais e Corporais

Básico Intermediário Avançado

Franquia *i*

As opções de franquia são:

Normal Reduzida

COTAR

Figura 26 – Informações das Coberturas do Seguro.

acontece porque a franquia é a participação obrigatória do segurado na reparação de seu veículo no caso de perda parcial. Se ele opta por diminuir essa participação, a seguradora procurará compensar aumentando o preço do seguro. A franquia reduzida, como o próprio nome diz, é menor que a franquia normal. No geral ela encarece o preço do seguro, mas tem a contrapartida positiva de, caso o usuário precise acionar o seguro (por exemplo, por causa de uma colisão) o gasto ser menor. A escolha entre a franquia normal e a franquia reduzida dependerá de como o usuário avalia determinadas circunstâncias.