UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelo de classificação de risco aplicado ao Seguro de Automóvel

Erika Novais Sales Correia

Trabalho de Conclusão de Curso - MBA em Ciência de Dados (CEMEAI)

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelo de classificação de risco aplicado ao Seguro de Automóvel

Erika Novais Sales Correia

USP - São Carlos

2021

ERIKA NOVAIS SALES CORREIA

Modelo de Classificação de risco aplicado ao Seguro de Automóvel

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciência de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Francisco Aparecido Rodrigues

USP - São Carlos

2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassie Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Correia, Erika Novais Sales
C824m Modelo de classificação de risco aplicado ao Seguro de Automóvel / Erika Novais Sales Correia; orientador Francisco Aparecido Rodrigues. -- São Carlos, 2021.
53 p.

Trabalho de conclusão de curso (MBA em Ciência de Dados) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Modelo Classificação. 2. Seguro Automóvel. 3. Gestão de Risco. 4. Subscrição de Seguro. I. Rodrigues, Francisco Aparecido , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

A minha querida filha Ana Luísa.

AGRADECIMENTOS

Agradeço a minha família e amigos que me apoiaram e incentivaram ao longo do curso e compreenderam minha ausência.

Agradeço, ainda, pela oportunidade de realizar esse curso, aos colegas, monitores e professores.

Agradeço especialmente ao meu orientador Prof. Francisco Rodrigues por todo apoio e paciência, seu direcionamento foi fundamental para conclusão desse trabalho.

“As grandes ideias surgem da observação dos pequenos detalhes”

Augusto Cury

RESUMO

CORREIA, E. N. S. **Modelo de classificação de risco aplicado ao Seguro de Automóvel**. 2021. 52 f. Trabalho de conclusão de curso (MBA em Ciência de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.

O objetivo fundamental de uma seguradora é a proteção de bens contra os riscos de perdas patrimoniais de um grupo de pessoas, para isso é fundamental que ocorra uma análise detalhada para prever possíveis intercorrências de um novo contrato de seguro tornando o trabalho de subscrição do risco umas das mais relevantes dentro de uma seguradora, pois ele tem a responsabilidade de aceitar e precificar o risco a fim de garantir as condições financeiras e garantia de proteção ao grupo de segurados. Nesse trabalho será explorada a aplicação de modelos de classificação e predição de risco para ocorrência de sinistro futuro para seguro de automóvel, com isso será possível determinar de forma mais assertiva a precificação e a subscrição, ou aceitação, de um determinado risco. Para isso foi realizado um estudo dos algoritmos KNN, Naive Bayes e Regressão Logística aplicando a base de sinistros de seguro de automóvel disponibilizado pela superintendência de Seguros Privados (SUSEP).

Palavras-chave: gestão de risco; modelo de classificação; seguros; subscrição.

ABSTRACT

CORREIA, E. N. S. **Risk classification model applied to Auto Insurance**. 2021. 52 f. Trabalho de conclusão de curso (MBA em Ciência de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.

The fundamental objective of an insurance company is the protection of assets against the risks of property losses of a group of people, for this it is essential that a detailed analysis takes place to predict possible complications of a new insurance contract, making the risk underwriting work one of the most relevant within an insurance company, as it has the responsibility to accept and price the risk in order to guarantee the financial conditions and guarantee of protection to the group of policyholders. In this job, the application of risk classification and prediction models will be explored for the occurrence of a future accident for auto insurance, with this it will be possible to determine in a more assertive way the pricing and the subscription, or acceptance, of a certain risk. For this, a study of the KNN, Naive Bayes and Logistic Regression algorithms was carried out applying the car insurance claims base provided by superintendent of private Insurance (SUSEP).

Keywords: risk management; classification model; insurance; underwriting.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SUSEP	–	Superintendência de Seguros Privados
CNSEG	–	Confederação Nacional das Empresas de Seguros Gerais
NOS	–	Número de sinistros ocorridos
NER	–	Número de expostos ao risco
F	–	Frequência de sinistros
MSO	–	Montante de sinistros ocorridos
VM	–	Valor médio do sinistro ou severidade
PE	–	Prêmio estatístico
PR	–	Prêmio de risco
KNN	–	K-Nearest Neighbor
ROC	–	*Receiver Operating Characteristic*)
AUC	–	*Area Under the Curve*
IS	–	Importância Segurada

SUMÁRIO

1 INTRODUÇÃO

1.1 Mercado Segurador

Conforme Vilanova (1969 *apud* FILHO, 2011, p.2) o seguro é uma operação que a partir do pagamento de uma pequena parte do objeto segurado (prêmio) uma pessoa (segurado ou beneficiário) no caso de ocorrência de um evento determinado (risco) receberá uma indenização do segurador que assume o risco para si.

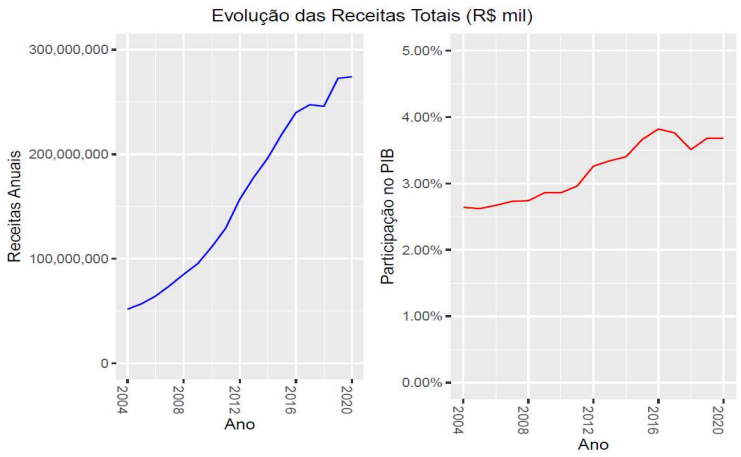
Para CALDEIRA (1997)

> Podemos, muito sinteticamente, afirmar que a operação de seguro privado é aquela na qual um grupo de pessoas (mutualismo) tenta se prevenir de um evento danoso futuro (incerteza) através de pequenas contribuições de todos os componentes do grupo (previdência). Nesse sentido, a finalidade do seguro é transferir o risco do segurado para a seguradora, assim, assumiria o risco, prestando a segurança de que, ocorrendo determinado evento, suas consequências seriam compensadas, economicamente, para o segurado. (CALDEIRA, 1997, p.13)

As empresas de seguros exercem um papel social importante, administrando riscos de pessoas e empresas, realizando investimentos no mercado financeiro seguindo as regras de um órgão regulador das atividades de Seguro (FILHO, 2011, p.6).

Em relação ao PIB brasileiro esse mercado cresceu significativamente, conforme dados extraídos da Superintendência de Seguros Privados (Susep), saltando de 2,6% em 2003 para o patamar de 3,7% em 2020, um dado muito expressivo uma vez que o PIB brasileiro se manteve crescente na maior parte do intervalo.

Figura 1- Evolução das Receitas Anuais e Participação no PIB



Fonte: SUSEP (2021).

1.1.1 Risco

Risco é todo evento incerto, porém possível de acontecer (Standerski, Kravec, 1979 *apud* FILHO, 2011).

Segundo Vilanova (1969 *apud* FILHO, 1969, p. 15) as características de risco segurável são:

a) afetar por igual a todos os componentes do grupo, podendo atingir a alguns, mas não a todo simultaneamente;
b) existir homogeneidade dos componentes do grupo, que deve ser o mais numeroso possível;
c) ocasionar uma necessidade econômica em sua realização;
d) ressarcir tão somente os prejuízos sofridos, não devendo constituir-se em lucro;
e) possibilitar, estatisticamente, basear-se em experiência passada, para deduzir leis que permitam prevês, em casos futuros da mesma natureza, iguais situações, desde que persistam as mesmas condições e circunstâncias;
f) existir independência na realização dos acontecimentos e essa realização deve ocasionar necessidade econômica, jurídica e efetivamente ressarcível.

Mendes (1977 *apud* FILHO, 2011) ainda menciona as condições de segurabilidade de um risco como ser possível, futuro, incerto, independente da vontade das partes, causar prejuízo de natureza econômica e ser quantitativamente mensurável.

1.1.2 Sinistro e indenização

Para Filho (2011) sinistro é a ocorrência de um risco segurado, ou seja, trata-se da efetivação de um evento coberto pelo seguro.

A indenização é o pagamento que a seguradora faz ao segurado em decorrência do sinistro (FILHO, 2011, p.8)

1.1.3 Prêmio de risco

Prêmio de risco é o valor pago na contratação do seguro, ou seja, o valor que a seguradora recebe do segurado para cobrir o risco de um determinado bem (FILHO, 2011, p.9).

Santos (1959 *apud* FILHO,2011) distingue a composição do prêmio em prêmio puro, resultante de dados estatísticos; e o carregamento, correspondente aos gastos gerais da seguradora.

Para Silva (1999 *apud* FILHO,2011) o preço deve ser formado por

- Valor esperado do sinistro;
- Despesas de comercialização;
- Despesas administrativas;
- Lucro esperado;
- Impostos;
- Oscilação do risco.

Conforme FILHO (2011) o Prêmio de risco pode ser obtido considerando os parâmetros abaixo:

NSO = Número de sinistros ocorridos
NER = Número de expostos ao risco
F = Frequência de sinistros
MSO = Montante de sinistros ocorridos
VM = Valor médio do sinistro ou severidade
PE = Prêmio estatístico ou *PR* = Prêmio de risco

Onde,

$$F = \frac{NSO}{NER} \qquad VM = \frac{MSO}{NSO}$$

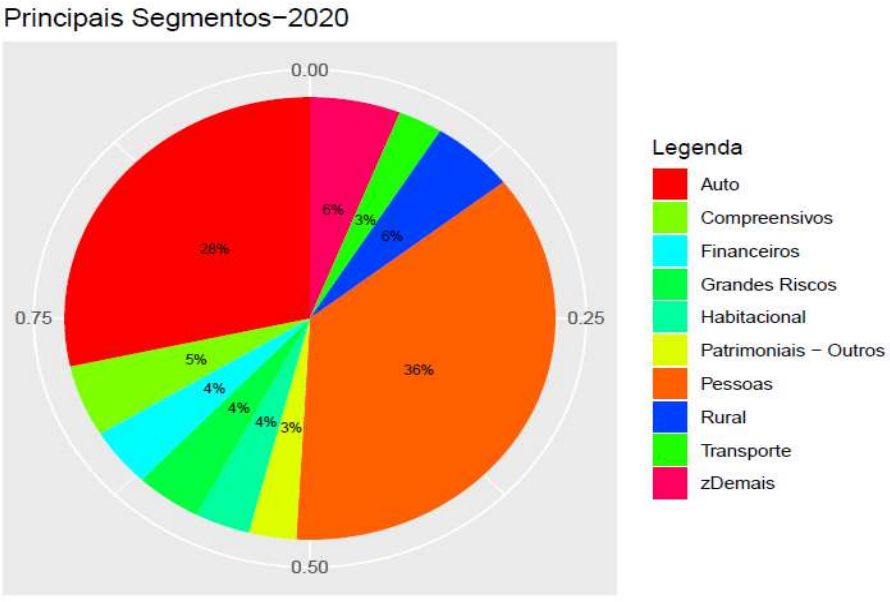
$$PE = F \times VM$$

O Prêmio de Risco é uma das variáveis base que compõem o prêmio total do seguro devendo incluir ainda despesas de comercialização e administrativas, impostos, lucro esperado e oscilação do risco (FILHO, 2011, p. 9,11).

1.2 Subscrição de Risco Seguro Automóvel

O segmento de seguro auto é o segundo maior ramo comercializado no Brasil conforme dados da SUSEP.

Figura 2 - Principais segmentos de Seguro no Brasil



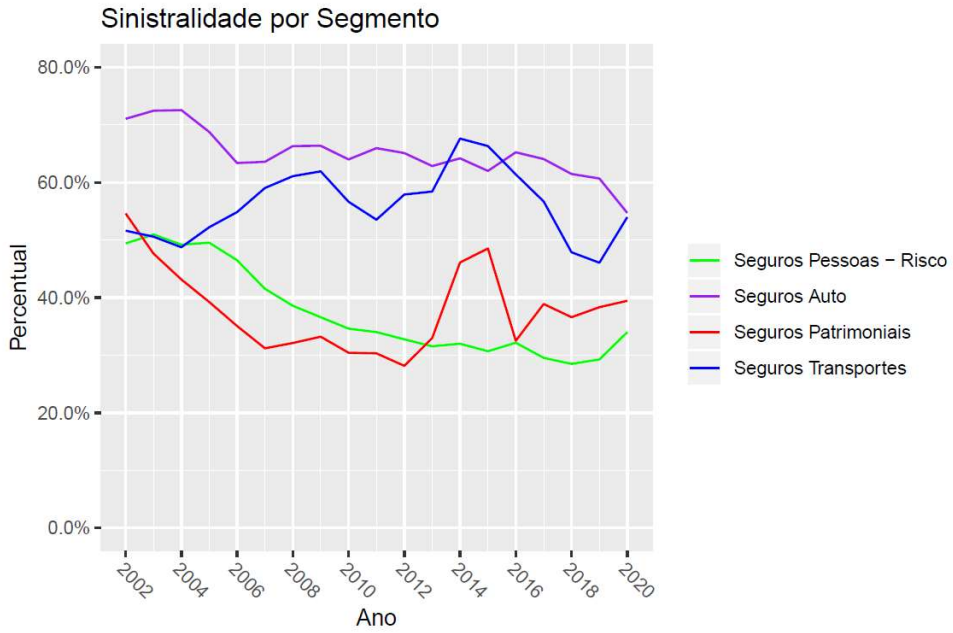
Fonte: SUSEP (2021).

Entre os seguros patrimoniais o seguro de automóveis é o mais avançado em relação ao processo de subscrição, dados disponíveis e métodos estatísticos e operação (PITA; DOMINGUEZ, 2011).

A subscrição do risco é um dos principais fatores de manutenção da saúde financeira das carteiras de Seguro, uma vez que é responsável pela análise do risco para aceitação e precificação de novas propostas (CNSEG, 2018).

Conforme dados da SUSEP em 2020 a sinistralidade do ramo Automóvel foi próxima a sessenta por cento.

Figura 3 - Taxa de Sinistralidade por Segmento



Fonte: SUSEP (2021).

Esses dados demonstram a importância da análise assertiva de dados da apólice do seguro para garantir o equilíbrio entre as emissões (prêmios) e regastes (sinistros).

1.2 Modelo de Classificação

O método de classificação faz uso da *aprendizagem supervisionada*, ou seja, existem dados já processados aos quais o modelo aprende para atribuir rótulo a uma amostra com base nos atributos (HARRISON, 2020).

De acordo com Provost e Foster (2016) o modelo irá determinar a que classe a nova proposta de seguro pertence representando a probabilidade da ocorrência de sinistros em relação aos demais indivíduos.

Além da análise de dados para a subscrição do risco e precificação a automatização do processo de captura e análise dos dados é essencial para a manutenção das carteiras de Seguro visto o número crescente de informações produzidas pelos mais diversos meios (AMARAL, 2016) se torna fundamental para garantia de ajustes cada vez mais eficazes.

Nas próximas etapas do projeto iremos testar os modelos para classificar o risco a possibilidade de sinistro em uma determinada proposta de seguros considerando variáveis como idade, sexo e região geográfica.

2 METODOLOGIA

2.1 Fases da Metodologia

Para iniciar o projeto foram definidos os seguintes passos para chegar ao resultado esperado:

1. Definição do conjunto de dados, análise exploratória e limpeza dos dados
2. Criação e transformação de variáveis, tratamento de dados faltantes (*missing*), dados discrepantes (outliers), dados desbalanceados, etc.
3. Separação dos conjuntos de Testes e Treinamento
4. Aplicação de algoritmos de Classificação ou Regressão: *Naive Bayes, K-Nearest Neighbor* (KNN) e Regressão Logística
5. Demonstração dos resultados e escolha do algoritmo.

2.2 Conjunto de Dados

Para realizar o estudo iremos utilizar os dados disponibilizados pela Superintendência de Seguros Privados através do sistema AUTOSEG (Dados estatísticos do Seguro de Automóveis) em forma de dados abertos.

2.2.1 Campos Base de Dados

Dados extraídos da base de dados abertos da Susep (https://dados.gov.br/dataset/dados-estatisticos-do-seguro-de-automoveis-autoseg/resource/a5539a4a-17e2-4e2e-96cb-70ce917ac7e2)

Tabela 1 - Tabela arq_casco_comp disponibilizada pela Susep

Campo	Descrição
IDADE	código e descrição de faixas etárias
COD_TARIF	Código Região Tarifária
REGIAO	Descrição da Região Tarifária
COD_MODELO	Código Modelo Veículo
ANO_MODELO	Ano Modelo Veiculo
SEXO	código e descrição de sexo (masculino, feminino, jurídico)
EXPOSICAO1	Quantidade de veículos expostos (ver conceito de exposição acima)
PREMIO1	Soma dos valores de prêmios, ponderados pela exposição de cada apólice
IS_MEDIA	Média das Importâncias Seguradas das apólices incluídas no grupamento definido pela chave escolhida, ponderada pela exposição de cada uma delas
FREQ_SIN1	Quantidade de sinistros da cobertura roubo/furto
INDENIZ1	Total de indenizações de sinistros da cobertura roubo/furto
FREQ_SIN2	Quantidade de sinistros da cobertura colisão parcial
INDENIZ2	Total de indenizações de sinistros da cobertura colisão parcial
FREQ_SIN3	Quantidade de sinistros da cobertura colisão perda total
INDENIZ3	Total de indenizações de sinistros da cobertura colisão perda total
FREQ_SIN4	Quantidade de sinistros da cobertura incêndio
INDENIZ4	Total de indenizações de sinistros da cobertura incêndio
FREQ_SIN9	Quantidade de sinistros de outras coberturas, como assistência 24 hs, etc
INDENIZ9	Total de indenizações de sinistros de outras coberturas, como assistência 24 hs, etc

Fonte: SUSEP (2021).

a) Média das Importâncias Seguradas

O Sistema Autoseg apresenta as informações de acordo com o grupamento escolhido pelo usuário, que representa o conjunto de apólices incluídas na respectiva chave composta (categoria tarifária, região de circulação, ano/modelo do veículo, perfil do principal condutor, etc). A Importância Segurada Média (IS Média) representa a média das IS's das apólices incluídas no grupamento, ponderada pela exposição de cada uma delas (ver conceito de exposição abaixo).

b) Expostos

O conceito de exposição leva em conta o tempo em que cada apólice esteve vigente, dentro da janela de observação, que é o período semestral abrangido em cada atualização do Autoseg. Desta forma, o número de expostos, apurado para um período anual, representa o melhor estimador disponível para a quantidade de veículos segurados.

c) Prêmio médio (R$)

Da mesma forma que a IS Média, o prêmio médio representa a média dos prêmios das apólices incluídas no grupamento, ponderada pela exposição de cada uma delas.

d) Frequência de Sinistros

A Frequência (FREQ) é a quantidade de sinistros de incêndio, roubo, colisão e outras causas, enquanto que o campo INDENIZ representa o total de indenizações para cada cobertura.

2.2.2 Tabelas Complementares (Domínios)

a) Faixa de Idade

Tabela 2 - Tabela de domínio para o campo 'IDADE' disponibilizada pela Susep

código	descrição
0	Não informada
1	Entre 18 e 25 anos
2	Entre 26 e 35 anos
3	Entre 36 e 45 anos
4	Entre 46 e 55 anos
5	Maior que 55 anos

Fonte: SUSEP (2021).

b) Sexo

Tabela 3 - Tabela de domínio para o campo ‘SEXO' disponibilizada pela Susep

código	descrição
M	Masculino
F	Feminino
J	Jurídica

Fonte: SUSEP (2021).

c) Código e Descrição da Região Tarifária

Tabela 4 - Tabela de domínio para o campo ‘COD_TARIF’ disponibilizada pela Susep

código	descrição
1	RS - Met. Porto Alegre e Caxias do Sul
2	RS - Demais regiões
3	SC - Met. Florianópolis e Sul
4	SC - Oeste
5	SC - Blumenau e demais regiões
6	PR - F.Iguaþu-Medianeira-Cascavel-Toledo
7	PR - Met. Curitiba
8	PR - Demais regiões
9	SP - Vale do Paraíba e Ribeira
10	SP - Litoral Norte e Baixada Santista
11	SP - Met. de São Paulo
12	SP - Grande Campinas
13	SP - Ribeirão Preto e Demais Mun. de Campinas
14	MG - Triângulo mineiro
15	MG - Sul
16	MG - Met.BH-Centro Oeste-Zona Mata-C. Vertentes
17	MG - Vale do Aço-Norte-Vale Jequitinhonha
18	RJ - Met. do Rio de Janeiro
19	RJ - Interior
20	ES - Espírito Santo
21	BA - Bahia
22	SE - Sergipe
23	PE - Pernambuco
24	PB - Paraíba
25	RN - Rio Grande do Norte
26	AL - Alagoas
27	CE - Ceará

28	PI - Piaui
29	MA - Maranhão
30	PA - Pará
31	AM - Amazonas
32	AP - Amapá
33	RO - Rondônia
34	RR - Roraima
35	AC - Acre
36	MT - Mato Grosso
37	MS - Mato Grosso do Sul
38	DF - Brasília
39	GO - Goiás
40	TO - Tocantins
41	GO - Sudeste de Goiás

Fonte: SUSEP (2021).

2.3 Métodos de Classificação

2.3.1 K-Nearest Neighbor (KNN)

O *k*-vizinho mais próximo (ou KNN) é baseado no conceito de distância ou proximidade entre os dados.

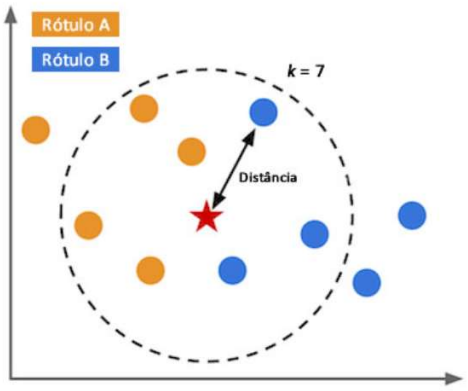
Nesse método cada objeto representa um ponto definido pelos atributos (ou espaço de entrada). Com isso, é possível calcular as distâncias entre cada ponto por meio de uma função. Assim, o algoritmo classifica um novo objeto com base nos experimentos do conjunto de treinamento mais próximos a ele, considerando k vizinhos (FACELLI et al., 2011).

O algoritmo KNN utiliza pesos uniformes, ou seja, cada ponto do conjunto local contribui uniformemente para a classificação de um ponto de consulta. Em algumas circunstâncias, pode ser vantajoso ponderar os pontos de forma que os pontos próximos contribuam mais para a regressão do que os pontos distantes (PEDREGOSA et al., 2011).

Embora seja um algoritmo considerado simples, uma vez que ele trabalha com a memorização dos dados de aprendizagem, ele é aplicável mesmo em problemas mais complexos. Além disso, o algoritmo é naturalmente incrementável, pois ao adquirir novos exemplos de treinamento, basta armazena-los na memória. (FACELLI et al., 2011).

Para usar esse algoritmo é necessário chegar à distância entre os valores mais próximos. Se os valores forem numéricos, é possível basear-se numa distância euclidiana, caso sejam dados categóricos então é possível usar métrica de sobreposição onde é classificado por semelhança (TAULLI, 2020).

Figura 4 - Exemplo de aferição das distâncias de uma amostra com dois rótulos de classe e com k = 7



Fonte: Pacheco (2017).

a) Definindo melhor valor de k

2.3.2 Naive Bayes

O algoritmo Naive Bayes se baseia em probabilidades, em um problema de duas classes definido por atributos booleanos de forma linear (FACELLI, 2011).

Conforme Kononenko (1991 *apud* FACELI et al, 2011, p76) o classificador Naive Bayes é robusto e apresenta boa performance mesmo com a presença de ruídos e atributos irrelevantes.

Esse algoritmo é baseado no Teorema de Bayes (proposto por Thomas Bayes) e tem o objetivo de encontrar a probabilidade posteriormente, ou seja, o classificador bayesiano apresenta uma maneira de calcular a probabilidade da ocorrência de um evento, baseando-se em probabilidades obtidas da análise dos eventos passados. Para Mitchell (1997 *apud* PIVETTA,2013). Estas informações são utilizadas para formar a base de conhecimento da aplicação ou o conjunto de dados responsáveis pela classificação correta das hipóteses.

O objetivo do classificador Naive Bayes é verificar se uma amostra analisada pertence ou não a uma determinada classe (PIVETTA, 2013). A obtenção desta resposta realiza-se através de uma análise estatística das informações coletadas sobre as instâncias apresenta o seu funcionamento através da Equação:

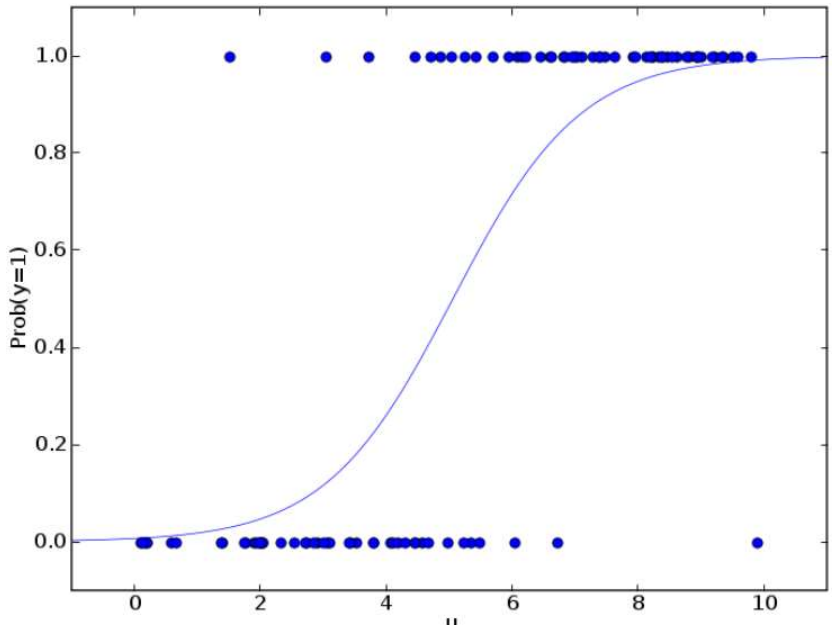
$$P(h|D) = \frac{P(D|h).P(h)}{P(D)}$$

2.3.3 Regressão Logística

A regressão logística é um recurso que permite estimar a probabilidade associada à ocorrência de determinado evento em face de um conjunto de variáveis explanatórias usando uma função logística (HARRISON, 2020), ou seja, busca estimar a probabilidade da variável dependente assumir um determinado valor em função dos conhecidos de outras variáveis

De acordo com Provost e Foster (2016), para a estimativa d probabilidade, a regressão logística usa o mesmo modelo linear, sendo que a saída do modelo é interpretada como log-chances de pertencer a classe, ou seja, qual é a probabilidade de pertencer a determinada classe.

Figura 5 - Curva da regressão logística



Fonte: Carvalho, Góes (2018).

2.4 Métricas de Validação e Avaliação dos Modelos

O objetivo do uso das métricas é medir a qualidade do modelo (SILVA; BORGES, 2019, p.282), serão utilizadas as seguintes métricas para o desenvolvimento desse trabalho:

2.4.1 Precisão Geral (*Accuracy*)

Refere-se ao número de acertos (positivos) dividido pelo número total de exemplos:

$$Accuracy = \frac{Numero\ de\ predições\ correta}{Numero\ total\ de\ predições\ feitas}$$

2.4.2 F1 Score

O F1 Score é a média entre precisão e recall

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

2.4.3 Precisão (*Precision*)

Trata-se do número de exemplos classificados como pertencentes a uma classe, as quais de fato são da classe (positivos verdadeiros) dividido pela soma entre esse número e o número de exemplos classificados na classe, porém que pertencem a outras classes (falso positivos).

$$Precision = \frac{PositivosVerdadeiros}{PositivosVerdadeiros + FalsosPositivos}$$

2.4.4 *Recall*

Refere-se ao número de exemplos classificados como pertencentes a uma classe, que realmente são daquela classe, dividido pela quantidade total de exemplos que pertencem a esta classe, mesmo sendo classificados em outra.

$$Recall = \frac{PositivosVerdadeiros}{PositivosVerdadeiros + FalsosNegativos}$$

2.4.5 Matriz de Confusão

A matriz de confusão consiste em uma matriz que contém os dados reais e os valores preditos pelo modelo (FRANCESCHI,2019)

Conforme Castro e Braga (2011 *apud* FRANCESCHI, 2019) ao longo da diagonal principal (na figura 10 em azul), estão representadas as predições corretas do modelo , ou, verdadeiros positivos (TP) e verdadeiros negativos (TN), os elementos fora dessa diagonal representam os erros, ou, falsos positivos (FP) e falsos negativos (FN).

Figura 6 - Resumo da Matriz de Confusão e métricas para avaliação de modelos.

Matriz de Confusão e métricas derivadas para avaliação de modelos de classificação.							
Matriz de confusão		Previsto pelo Modelo		Sensibilidade	Especificidade	Precisão	Acurácia
		Não churner	Churner				
Situação real	Não churner (N)	Verdadeiro Negativo (TN)	Falso Positivo (FP)	$= \frac{TP}{TP+FN}$	$= \frac{TN}{TN+FP}$	$= \frac{TP}{TP+FP}$	$= \frac{TP+TN}{TP+TN+FP+FN}$
	Churner (P)	Falso Negativo (FN)	Verdadeiro Positivo (TP)				

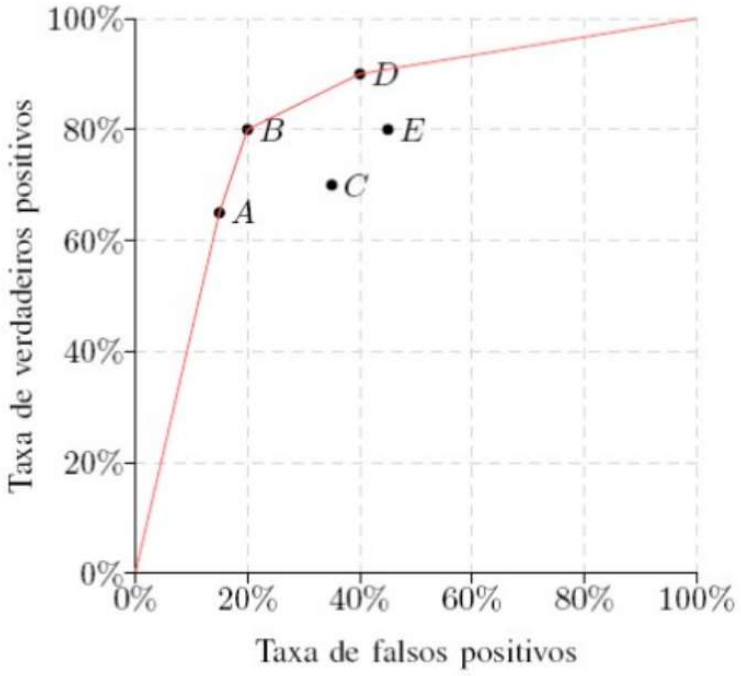
Fonte: Franceschi, 2019.

2.4.6 Curva ROC e AUC

Análise ROC (*Receiver Operating Characteristic*), trata-se de um método de avaliação e seleção de métodos de classificação, é uma técnica para visualizar, avaliar, organizar e selecionar classificadores baseados em suas performances.

Segundo Prati et al (2008) o gráfico ROC é baseando na probabilidade de detecção, ou taxa de verdadeiros positivos ($tpr = (P\ (Y|X))$, e na probabilidade de falsos alarmes, ou taxa de falsos positivos ($fpr - P(\ Y|\overline{X}\ )$). Para construir o gráfico ROC plota-se fpr no eixo x (ordenadas) e tpr no eixo y (abscissas).

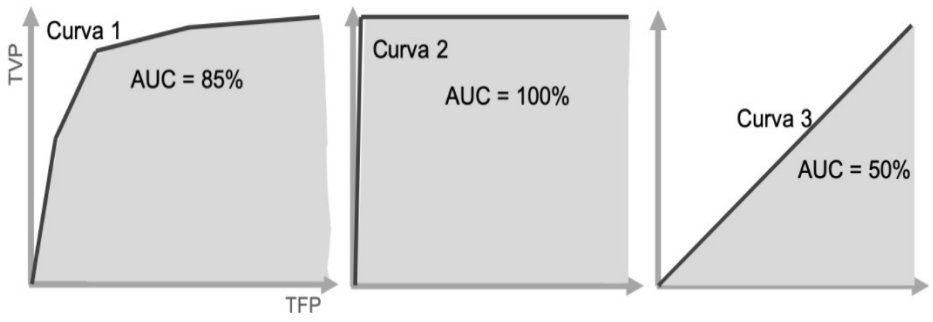
Figura 7 - Modelos de Classificação no espaço ROC



Fonte: Prati et al (2008).

Para avaliação de desempenho de diferentes classificadores (em um problema de classificação binária) utilização AUC (*Area Under the Curve*) que mede a área abaixo da curva ROC, para cada um dos classificadores, o classificador com maior valor de AUC é o que apresenta o melhor desempenho (médio) para o conjunto de exemplo (WANDERLEY, 2010).

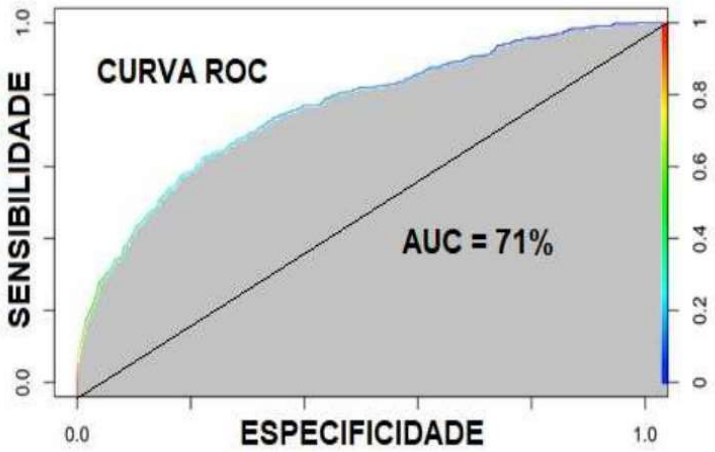
Figura 8 - Exemplo de valores de AUC



Fonte: Prates (2020).

No exemplo abaixo para uma AUC calculada em 71% significa que há esse percentual de chances de que o modelo seja capaz de distinguir entre classe positiva ou classe negativa.

Figura 9 - Curva ROC e AUC



Fonte: Franceschi, 2019.

3 MODELAGEM

3.4 Aplicação algoritmos e Tratamento de classes desbalanceadas

O objetivo da aplicação do modelo é chegar em um conjunto de modelo treinado para que ao ser informado os dados: Região, Categoria Tarifária, Ano Modelo, Idade (segurado), IS Média, Sexo, Exposição, seja possível calcular a probabilidade a um nível confiável de ocorrência de sinistro para a apólice e compor a precificação e subscrição do seguro.

Algumas premissas foram necessárias para a aplicação dos modelos citados acima, para isso foi necessário realizar o tratamento e seleção dos dados, chegando no seguinte modelo:

Tabela 5 - Tratamento de Dados

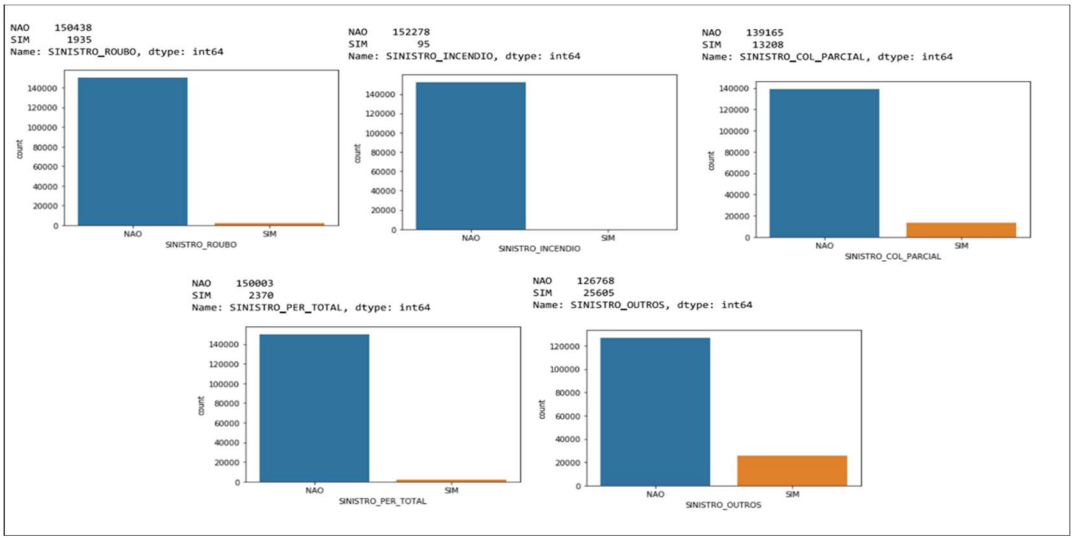
	Campo	Descrição
1	'COD_CAT_TARIF',	Código da Categoria Tarifária
2	'COD_MODELO_GRUPO',	Código Modelo do Veículo
3	'ANO_MODELO',	Ano Modelo Veículo
4	'IDADE',	Idade
5	'EXPOSICAO1',	Quantidade de Expostos
6	'PREMIO1',	Prêmio
7	'IS_MEDIA',	Importância Segurada Média
8	'FREQ_ROUBO',	Quantidade de Sinistro Roubo
9	'INDENIZ_ROUBO',	Indenização Sinistro Roubo

10	'FREQ_COL_PARCIAL',	Quantidade de Sinistro Colisão Parcial
11	'INDENIZ_COL_PARCIAL',	Indenização Sinistro Colisão Parcial
12	'FREQ_PERDA_TOTAL',	Quantidade de Sinistro Perda Total
13	'INDENIZ_PERDA_TOTAL',	Indenização Sinistro Perda Total
14	'FREQ_INCENDIO',	Quantidade de Sinistro Incêndio
15	'INDENIZ_INCENDIO',	Indenização Sinistro Incêndio
16	'FREQ_OUTROS',	Quantidade de Sinistro Outras Coberturas
17	'INDENIZ_FREQ_OUTROS',	Indenização Sinistro Outras Coberturas
18	'COD_NUM_SEXO',	Código Sexo
19	'REGIAO',	Código Região
20	'SINISTRO_ROUBO',	Flag Sinistro Roubo
21	'SINISTRO_COL_PARCIAL',	Flag Sinistro Colisão Parcial
22	'SINISTRO_PER_TOTAL',	Flag Sinistro Perda total
23	'SINISTRO_INCENDIO',	Flag Sinistro Incêndio
24	'SINISTRO_OUTROS',	Flag Sinistro Outros
25	'HOUVE_SINISTRO',	Descrição Flag Sinistro Geral
26	'HOUVE_SINISTRO_BIN']	Flag Sinistro Geral

Fonte: Próprio Autor (2022).

Para otimizar a validação e demonstração dos resultados, vamos focar nas variáveis de quantidade de frequência de sinistros, para isso foi houve a construção de novas colunas data frame com o valor booleano de frequência por tipo de sinistro: Roubo e Furto (FREQ_SIN1), Colisão Parcial (FREQ_SIN2), Perda Total (FREQ_SIN3), Incêndio (FREQ_SIN4), Sinistro Outras Coberturas (FREQ_SIN9) e se houve sinistro de qualquer natureza.

Figura 10 - Análise de Registros de Sinistro por tipo de Natureza

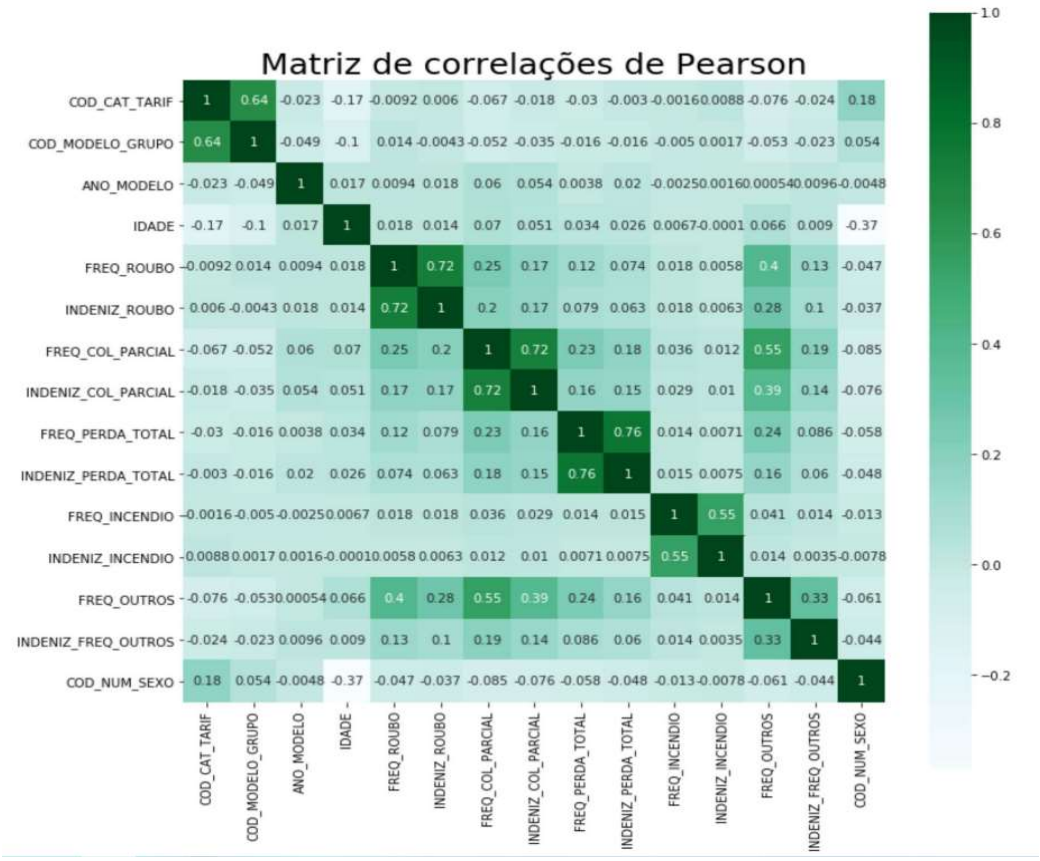


Fonte: Próprio Autor (2022).

Foi realizada a avaliação da correlação dos dados através da matriz de correlações de Pearson, onde o coeficiente assume valores de -1 e 1, onde o valor 1 significa uma correlação positiva perfeita e -1 uma correlação negativa perfeita entre as variáveis, o valor 0 significa que não há uma correlação.

No gráfico é possível observar as variáveis com maior correlação na base:

Figura 11 - Matriz de correlações de Pearson

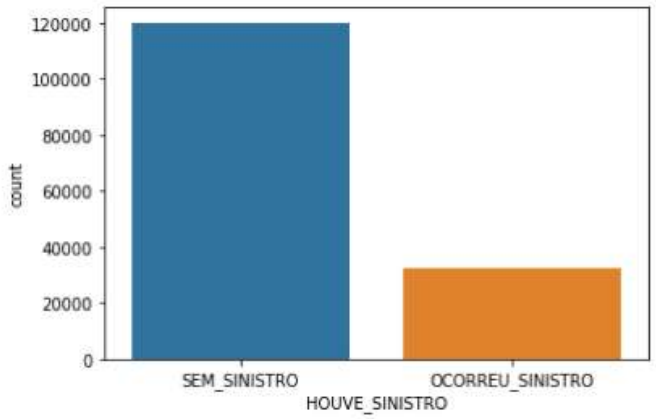


Fonte: Próprio Autor (2022).

Criação de uma coluna auxiliar considerando se houve qualquer sinistro:

Figura 12 - Ocorrência de Sinistros todas as Naturezas

```
SEM_SINISTRO        120019
OCORREU_SINISTRO     32354
Name: HOUVE_SINISTRO, dtype: int64
```



Fonte: Próprio Autor (2022).

Definição de um dataframe de treinamento e teste com valores chave

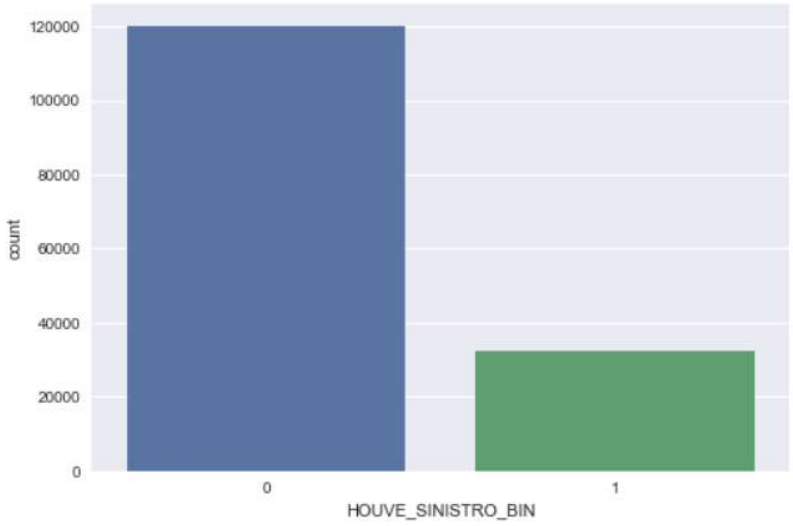
Figura 13 - DataFrame Utilizado para Aplicação dos Métodos de Classificação

	COD_CAT_TARIF	REGIAO	ANO_MODELO	IDADE	PREMIO1	IS_MEDIA	COD_NUM_SEXO	EXPOSICAO1	HOUVE_SINISTRO_BIN
571036	1	11	2009	2	663.547698974609	26765.4571573431	2.0	0.827397227287292	0
1583743	3	09	2016	5	20432.8662719727	104063.765086472	1.0	9.90684905275702	0
43249	2	18	2014	3	6126.14697265625	165967.006156271	1.0	0.504109561443329	0
2538332	3	13	2010	1	797.22783946991	6281.97479473013	3.0	1.51506843231618	0
2356493	4	07	2012	3	894.645751953125	104106.002822646	3.0	0.287671238183975	0
2074603	3	13	2016	3	4914.80593109131	49117.3703370386	1.0	2.90684919804335	1
3091360	1	8	2016	0	0	0	2.0	0	1

Fonte: Próprio Autor (2022).

Verificação do balanceamento da classe que será considerado pelo modelo para predizer de houve ou não sinistro:

Figura 14 - Balanceamento das Classes




```
: 0    120019
  1     32354
  Name: HOUVE_SINISTRO_BIN, dtype: int64
```

Fonte: Próprio Autor (2022)

Pela imagem é possível verificar que os registro onde não houve ocorrência de sinistro (SEM_SINISTRO) correspondem a 78,8% do total da base enquanto registros sem sinistro são 21.2% demonstrando que a classe está desbalanceada.

Para fins de testes foi aplicado os modelos sobre esse viés, primeiro definindo 80% da base para treinamento e 20% para testes

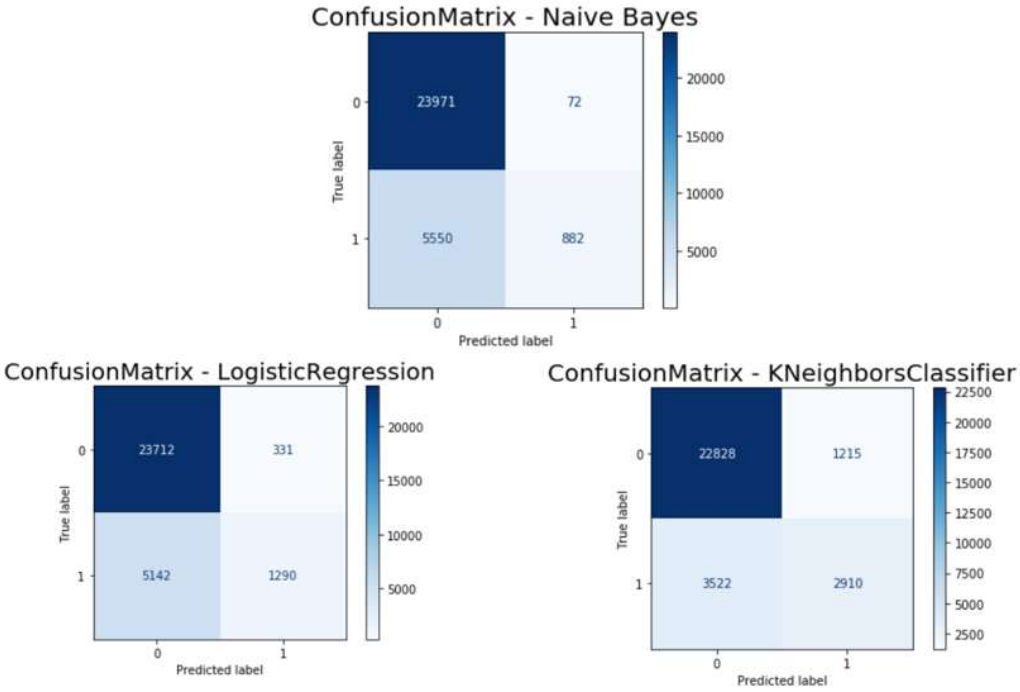
Após a separação dos conjuntos de treinamento e teste foi aplicado os modelos de Regressão Logística, K-vizinhos e Naive Bayes, chegando aos resultados conforme tabela abaixo:

Tabela 6 - Relatório dos Classificadores Regressão Logística, KNN e Naive Bayes

	Accuracy	**F1 score**	**Precision**	**Recall**
Regressão Logística	0.83	0.65	0.83	0.62
KNN	0.84	0.72	0.77	0.69
Naive Bayes	0.82	0.64	0.76	0.62

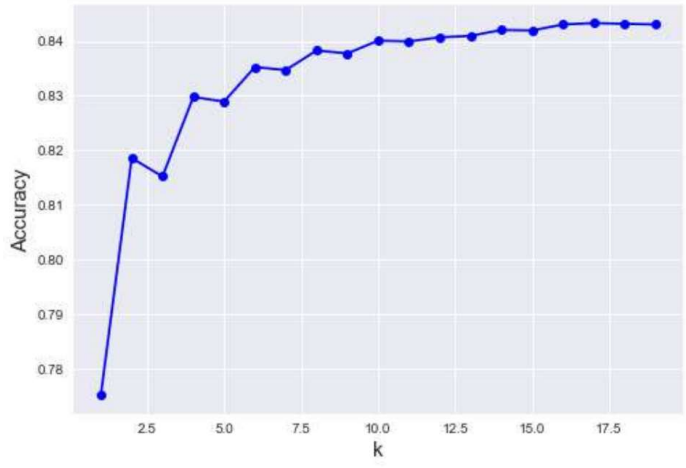
Fonte: Próprio Autor (2022).

Figura 15 – Matriz de Confusão Modelos



Fonte: Próprio Autor (2022).

Figura 16 - Definição do valor de k para através do *cross-validate*

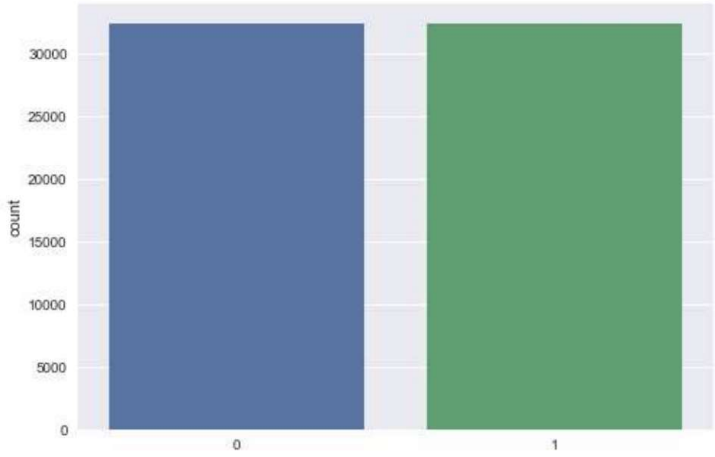



```
Melhor k: 17
```

Fonte: Próprio Autor (2022).

Após rodar os algoritmos com as classes desbalanceadas realizamos o balanceamento das classes.

Figura 17 - Balanceamento das classes utilizando NearMiss



Fonte: Próprio Autor (2022).

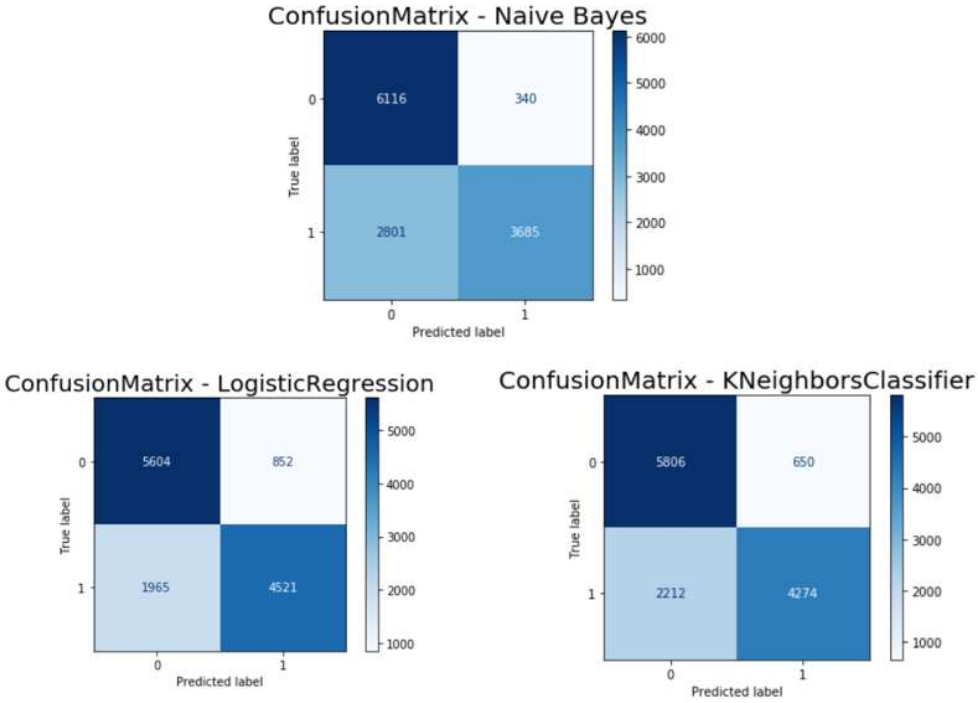
Após o balanceamento os algoritmos foram aplicados novamente gerando o resultado conforme tabela abaixo.

Tabela 7 - Comparação entre os modelos após balanceamento das classes

	Accuracy	**F1 score**	**Precision**	**Recall**
Regressão Logística	0.77	0.77	0.78	0.77
KNN	0.77	0.77	0.79	0.77
Naive Bayes	0.75	0.74	0.79	0.75

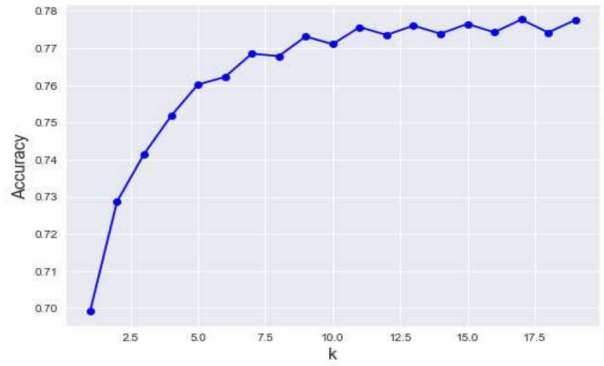
Fonte: Próprio Autor (2022).

Figura 18 – Matriz de Confusão modelos após balanceamento das classes



Fonte: Próprio Autor (2022).

Figura 19 - Novo Resultado k para KNN




```
Melhor k: 17
```

Fonte: Próprio Autor (2022).

3.7 Validação dos resultados

Para validação dos resultados entre com as classes desbalanceadas e balanceadas é possível verificar uma melhora considerável no recall que demonstra que as classes tem quantidades semelhante indicando que o modelo consegue classificar com mais precisão:

Figura 20 - Comparação de Resultados após balanceamento das classes

Resultado antes balanceamento das classes

	Accuracy	F1 score	Precision	Recall
Regressão Logística	0.83	0.65	0.83	0.62
KNN	0.84	0.72	0.77	0.69
Naive Bayes	0.82	0.64	0.76	0.62

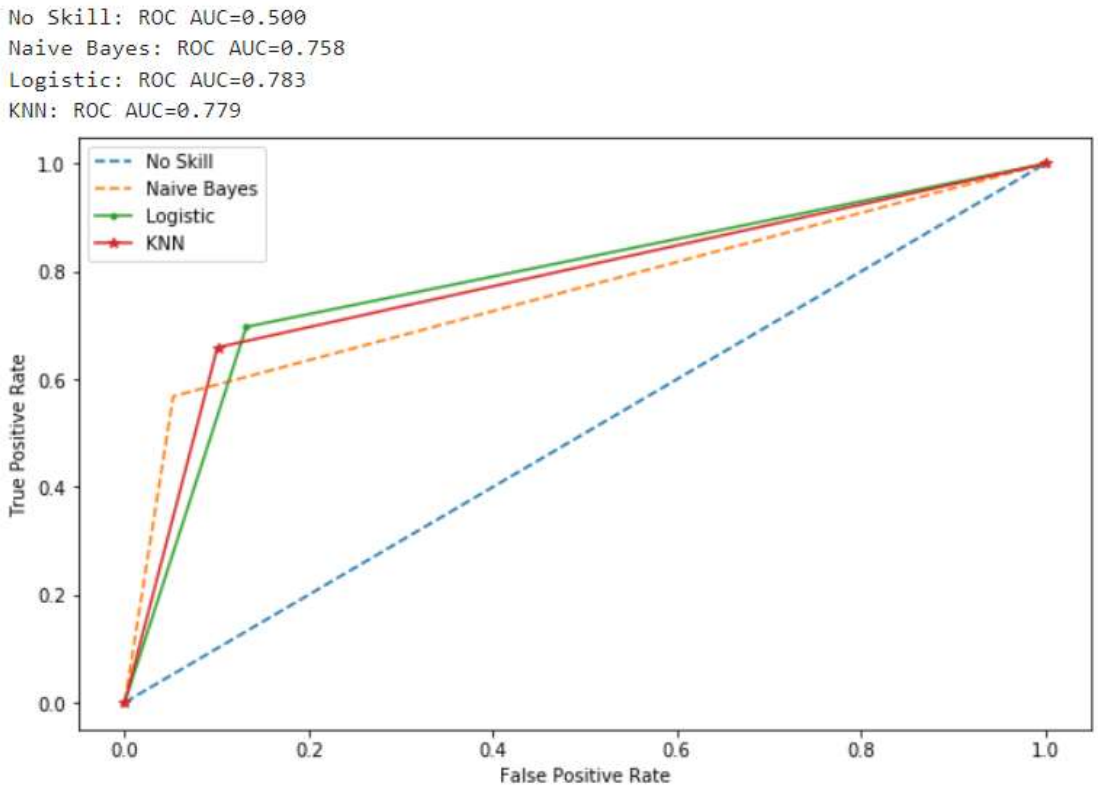
Resultado após balanceamento das classes

	Accuracy	F1 score	Precision	Recall
Regressão Logística	0.77	0.77	0.78	0.77
KNN	0.77	0.77	0.79	0.77
Naive Bayes	0.75	0.74	0.79	0.75

Fonte: Próprio Autor (2022).

Para a avaliação dos classificadores foi realizado a comparação seguindo o método da curva ROC ao considerar esse parâmetro o método de Regressão Logística tem uma pequena vantagem em relação aos demais.

Figura 21 - Curva ROC (*Receiver Operating Characteristic*) avaliação da qualidade da classificação



Fonte: Próprio Autor (2022).

4 CONCLUSÃO

4.1 Comentários

Para realizar a subscrição em carteira de seguros de automóvel além dos fatores de créditos comumente utilizados no mercado financeiro, além da análise do perfil do cliente (sexo, idade, ramo de atividade) é necessário realizar um estudo da natureza do ramo buscando

informações como região geográfica de circulação do veículo, histórico de sinistralidade, realização de vistoria veicular.

Nesse sentido a aplicação de modelos de classificação é uma ferramenta fundamental para realizar a análise do risco de forma eficaz.

Esse trabalho demonstrou que a utilização de classificadores são fundamentais para compor a análise do risco e a precificação do seguro uma vez que temos uma vasta base de dados com resultados de anos anteriores o que permite uma assertividade cada vez maior.

4.1 Trabalhos Futuros

Como trabalhos futuros sugere-se:

a) Aplicar os métodos de classificação para os dados de Indenização;
b) Realizar a composição do preço considerando o resultado desses classificadores;
c) Aplicar a metodologia com dados específicos para o veículo;
d) Aplicar a metodologia para ofertar produtos e treinamentos personalizados para o cliente.

REFERÊNCIAS

AMARAL, F.**Introdução à Ciência de dados**: Mineração de dados e Big Data. Rio de Janeiro: Alta Books, 2016.

CALDEIRA, L. **O Contrato de Seguro Privado e a Proteção do Consumidor**. Rio de Janeiro:Funenseg, 1997.

CARVALHO, A.X.Y; GOES, G.S. Introdução ao Software R e à Análise Econométrica. Disponível em < https://repositorio.enap.gov.br/bitstream/1/3452/3/Aula%202%20-%20Geraldo%20Goes%20e%20Alexandre%20Ywata%20-%20Introdu%C3%A7%C3%A3o%20%C3%A0%20Regress%C3%A3o%20Log%C3%ADstica.pdf> Acesso em: 19 fev. 2022.

CNSEG. Programa Educação em Seguros: Gerenciamento de Risco e o Seguro. p. 39.2017. Disponível em < https://cnseg.org.br/publicacoes/livretos-de-educacao-em-seguros.html> Acesso em: 23 abr. 2021.

FACELI, K.; LORENA, A.; GAMA, J.; DE CARVALHO, A. **Inteligência Artificial**: Uma Abordagem de Aprendizado de Máquina. Rio de Janeiro: LTC, 2011.

FILHO, O.L. **Seguros**: Fundamentos, Formação de Preço, Provisões e Funções Biométricas. São Paulo: Atlas, 2011.

FRANCESCHI, P.R. **Modelagens Preditivas de *Churn***: O caso do Banco do Brasil. Disponível em< http://www.repositorio.jesuita.org.br/bitstream/handle/UNISINOS/9087/Pietro%20Reinheimer%20de%20Franceschi_.pdf?sequence=1#page=75> Acesso em: 19 fev. 2022

HARRISON,M. **Machine Learning**: Guia de Referência Rápida. São Paulo:OREILLY,2020.

PACHECO, A. **K vizinhos mais próximos – KNN**. Disponível em <https://computacaointeligente.com.br/algoritmos/k-vizinhos-mais-proximos/#fukunaga> Acesso em: 19 fev. 2022.

PEDREGOSA, F. et al. **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011.

PITA, R.;DOMINGUEZ,A. **Seguro de Automóvel**. Rio de Janeiro:Funenseg.p.3.2011.

PIVETTA, S. **Classificação de Documentos do Exército Brasileiro Utilizando o Classificador Naive Bayes e Técnicas de Seleção de Sentenças.** Alegrete: Universidade Federal do Pampa, p. 24,25.2013

PRATES, W.R. **Curva ROC e AUC em Machine Learning**. Disponível em < https://cienciaenegocios.com/curva-roc-e-auc-em-machine-learning/> Acesso em: 19 fev. 2022

PRATI, R.C. et al. Curvas ROC para avaliação de classificadores. Disponível em < https://www.researchgate.net/profile/Ronaldo-Prati/publication/3455223_Evaluating_Classifiers_Using_ROC_Curves/links/06935dc2039d9e7f57bb23a6/Evaluating-Classifiers-Using-ROC-Curves.pdf> Acesso em: 19 fev. 2022

PROVOST, F.; FAWCETT, T. **Data Science para Negócios**: O que você precisa saber sobre mineração de dados e pensamento analítico de dados. Rio de Janeiro:Alta Books, p. 20, 2016.

SILVA, C. E; BORGES, L. G. **Uma Abordagem Inteligente para Classificação de Perfis de Aprendizagem de Discentes com Aplicação de Aprendizado de Máquina**. Disponível em <https://revistas.unifacs.br/index.php/rsc/article/view/6082> Acesso em: 19 fev. 2022

SUSEP. 9º Relatório de Análise e Acompanhamento dos Mercados Supervisionados.2021. Disponível em < http://www.susep.gov.br/menuestatistica/SES/relat-acomp-mercado-2021.pdf> Acesso em: 17 jul. 2021

TAULLI, T. **Introdução à Inteligência Artificial**: Uma abordagem não técnica. São Paulo: Novatec, p. 84. Disponível em < https://www.google.com.br/books/edition/Introdu%C3%A7%C3%A3o_%C3%A0_Intelig%C3%AAncia_Artificial/ON3FDwAAQBAJ?hl=pt-BR&gbpv=1&dq=Intelig%C3%AAncia+artificial:+uma+abordagem+de+aprendizado+de+m%C3%A1quina&printsec=frontcover> Acesso em: 12 out. 2021

WANDERLEY,M.F.B;BRAGA,A.P. Seleção de Características Baseadas em Análise de área abaixo da Curva ROC de Classificadores KDE-Bayesianos. Disponível em< http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.705.807&rep=rep1&type=pdf#page=57> Acesso em: 19 fev. 2022