-

CENTRO UNIVERSITÁRIO

INSTITUTO DE EDUCAÇÃO SUPERIOR DE BRASÍLIA
Pós-Graduação em Ciência de Dados
DISCIPLINA: Métodos estatísticos para apoio de decisão I

NOME DO ALUNO:

ERIC RIBEIRO FERNANDES

=

Brasília
Novembro de 2020

SUMÁRIO

1. Introdução

1.1 Contextualização macroeconômica dos bancos no ano de 2020

O ano de 2020 trouxe grandes desafios. A pandemia causada pela COVID-19, além de acarretar milhões de mortes, também trouxe consequências econômicas. Com o isolamento social algumas empresas tiveram redução em seus faturamentos e para diminuir seus custos demitiram funcionários, já outras não conseguiram sobreviver e acabaram encerrando suas atividades.

Com o desaquecimento econômico o Governo Federal do Brasil tomou algumas medidas a fim de suavizar essa redução de circulação de dinheiro, como o auxílio emergencial para pessoas físicas, saque do Fundo de Garantia do Tempo de Serviço (FGTS), suspensão de recolhimento de alguns impostos e diversas outras.

O crédito bancário também foi alvo de medidas do governo. Para facilitar sua tomada o Comitê de Política Monetária (COPOM) vem reduzindo a taxa de juros básica, chegando ao menor patamar histórico (2% a.a.), esse corte impactou o Indicador de Custo de Crédito (ICC) tornando a captação de crédito mais barata (Figura 1).

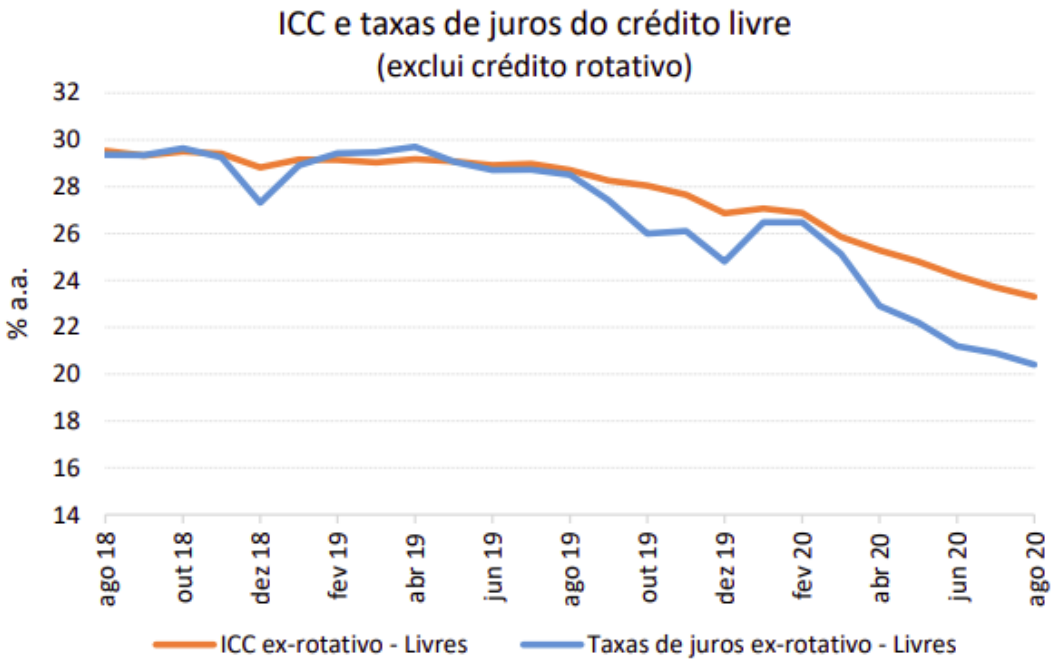


Figura 1 Fonte: Banco Central. Estatísticas monetárias e de crédito.

Os bancos, privados e públicos, também tomaram medidas acerca do crédito bancário. A principal delas foi a criação de linhas de crédito voltadas aos

microempreendedores. Porém, é interessante observar um certo dilema macroeconômico que os bancos vivem, enquanto uma crise econômica aumenta o risco dos tomadores de empréstimos não solverem seus débitos, o Banco Central (Bacen) força uma queda na taxa de juros de mercado, aumentando a demanda para o crédito bancário. A concessão de crédito vem aumentando (Figura 2) e os bancos procuram melhores formas de avaliar seus tomadores a fim de minimizar o risco de inadimplência (*default*). Modelos de aprendizagem de máquina se tornaram opções consideradas pelas instituições financeiras para avaliar ou quantificar o risco de crédito.

1.2 Objetivos do estudo

Diante disso, o presente estudo tem a finalidade de criar um modelo de aprendizado de máquina para prever se a concessão de crédito para determinada pessoa é boa ou ruim baseado nas técnicas de *Random Forets, Decision Tree, Logistic Regression e Gaussian Naïve Bayes* e avaliar qual modelo performou melhor.

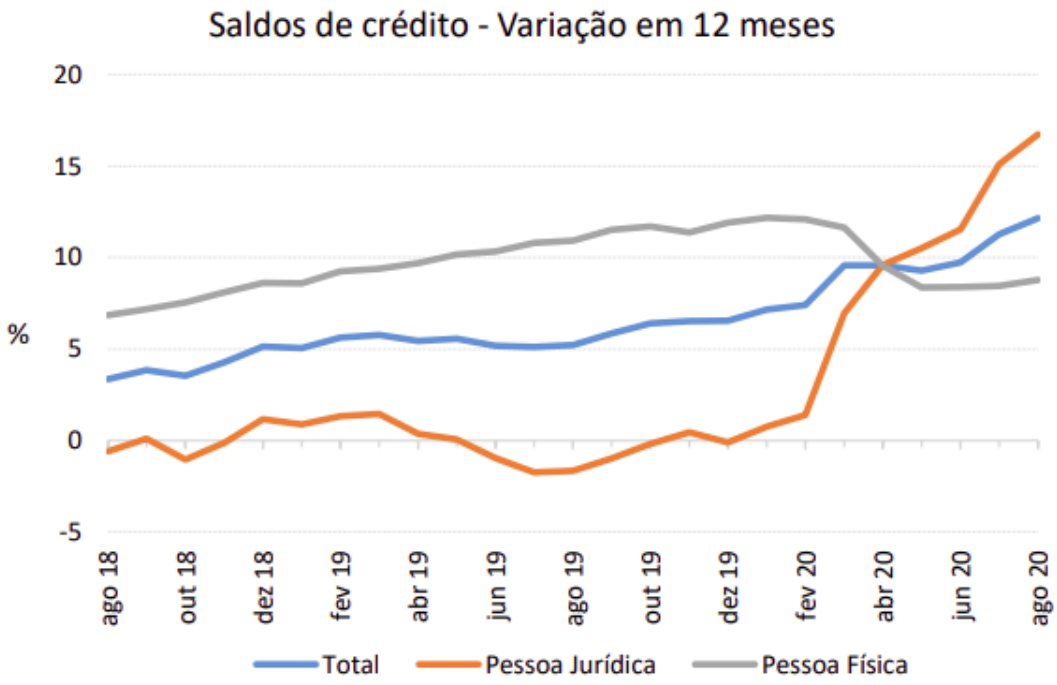


Figura 2 Fonte: Banco Central. Estatísticas monetárias e de crédito

2. Referencial teórico

3.1 Risco de crédito

Segundo Brito e Assaf Neto (2008) o crédito, para uma instituição financeira, refere-se à atividade de colocar um valor à disposição de um tomador de recursos. Pode ser sob a forma de financiamento ou empréstimo e contempla a promessa de pagamento em data futura.

Logo, o risco de crédito está relacionado à possibilidade de não pagamento, pelo tomador, na data pactuada com o devedor. Na estatística o risco pode ser definido como a variabilidade dos retornos esperados de um ativo, sendo muitas vezes associado ao desvio-padrão (σ) de uma média ($\bar{X}$) de retorno esperado (FIGUEIRA, 2001, p.14).

Vários fatores podem contribuir para que o credor não receba na data pactuada. Esses fatores são denominados fatores de risco. "Tais fatores nunca ocorrem sozinhos ou em proporções previamente definidas. Na realidade existe uma forte inter-relação e interdependência entre os diversos fatores de riscos" (BLATT, 1999, p.54).

3.2 Técnicas de aprendizagem de máquina (machine learning)

De acordo com Han e Kamber (2011) aprendizagem de máquina investiga como o computador pode aprender a resolver problemas se baseando em dados. Sua principal atuação se encontra em reconhecer padrões complexos que, normalmente, um ser humano teria dificuldade em descobrir. Existem várias categorias de aprendizado de máquina, sendo as principais listadas a seguir:

- Aprendizagem supervisionada: Essa categoria engloba os algoritmos que aprendem com bases de dados que possuem uma variável-alvo (*target*). Por exemplo, pretende-se criar um modelo que possa prever se irá chover em determinado dia, para isso é informado ao algoritmo uma base de dados com eventos históricos e especificado se nos dias passados choveu ou não.
- Aprendizagem não-supervisionada: Já esta categoria, ao contrário da antecessora, engloba algoritmos que não possuem uma variável *target*. Geralmente esses algoritmos são utilizados para descobrir padrões durante a análise exploratória de um problema, podendo criar grupos que possuem características em comum (*clusters*).

Algoritmos de aprendizagem de máquina possuem foco em sua acurácia, ou seja, na probabilidade que o algoritmo possa prever um resultado certo.

2.3 Risco de crédito e machine learning

Segundo o estudo realizado por Aniceto (2016) no período entre 1992 até 2014 foram encontrados 80 artigos relacionando os temas *risco de crédito* e *machine learning*, nos bancos de dados Science Direct e Web of Science. A pesquisadora demonstra um aumento de publicação de artigos a partir do ano de 2004 (Figura 3) e que a nacionalidade dos autores com maior frequência de publicação são, respectivamente, Taiwan, China, Estados Unidos e Grécia (Figura 4).

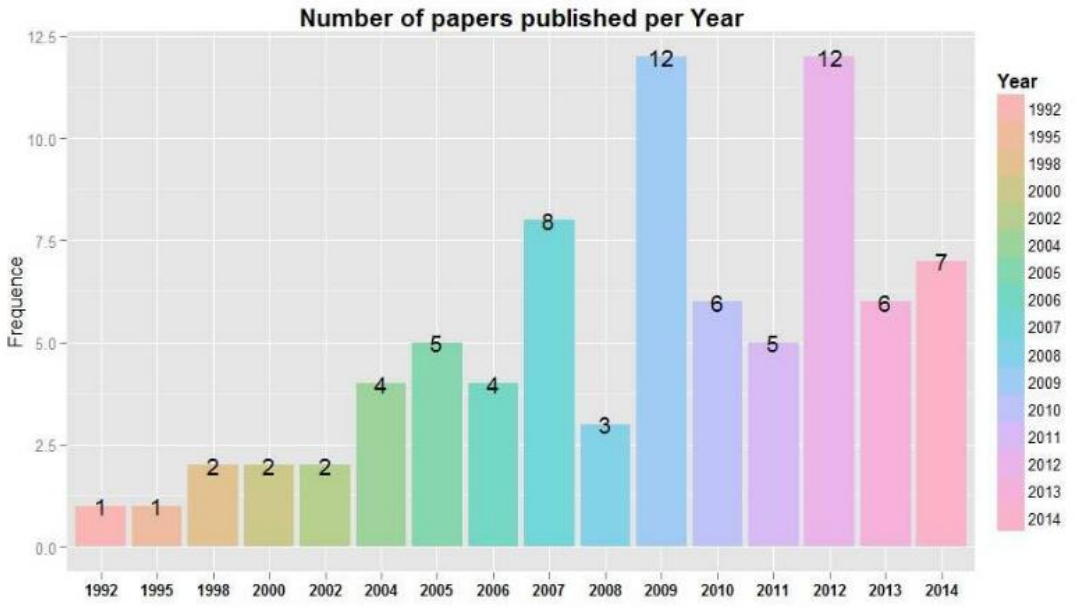


Figura 3 Aniceto. Estudo Comparativo entre Técnicas de Aprendizado de Máquina para Estimação de Risco de Crédito

Na revisão da literatura, realizada pela autora, as técnicas de *machine learning* mais abordadas, no contexto de risco de crédito, nos artigos pesquisados são *Artificial Neural Networks* (ANN), *Decision Trees* (DT) e *Support Vector Machines* (SVN). A análise também revelou que *Logistic Regression* e *Discriminant Analysis* são as técnicas secundárias mais encontradas.

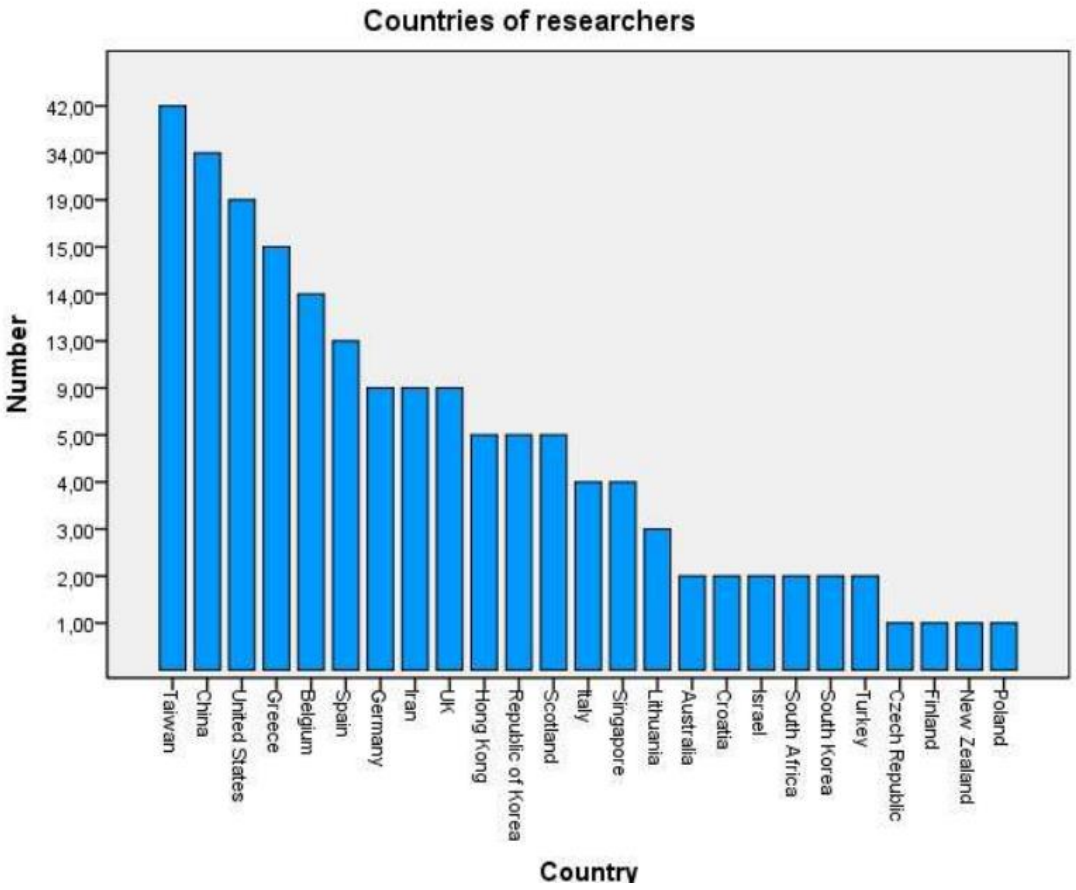


Figura 4 Aniceto. Estudo Comparativo entre Técnicas de Aprendizado de Máquina para Estimação de Risco de Crédito

3. Metodologia

4.1 Características do dataset do estudo

A base de dados escolhida para realizar o experimento foi encontrada no site https://www.kaggle.com, denominada *German Credit Risk - With Target*. Publicada no site em setembro de 2019 o *dataset* já fora objeto de diversos estudos relacionados à aprendizagem de máquina, sendo mais de 5.000 *downloads* já efetuados até a publicação do presente estudo.

A descrição das variáveis independentes (*features*) da amostra são detalhadas a seguir:

i) Age (quantitativa descritiva).
ii) Sex (qualitativa nominal: male, female).
iii) Job (qualitativa ordinal: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled).
iv) Housing (qualitativa nominal: own, rent, or free).
v) Saving accounts (qualitativa ordinal - little, moderate, quite rich, rich).
vi) Checking account (quantitativa contínua, em DM - Deutsch Mark).
vii) Credit amount (quantitativa contínua, em DM).
viii) Duration (quantitativa descritiva, in month).
ix) Purpose (qualitativa nominal: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others).

Já a variável dependente (*target*) é descrita a seguir:

i) Risk (qualitativa nominal binomial: good, bad).

O *dataset* possui 1.000 registros, sendo 700 classificados como *good* e 300 como *bad*. A composição dos dados faltantes é descrita na Tabela 1.

Feature	***Missings***	***Dtype***
Age	-	int64
Sex	-	object
Job	-	int64
Housing	-	object
Saving accounts	183	object
Checking account	394	object
Credit amount	-	int64
Duration	-	int64
Purpose	-	object
Risk	-	object

Tabela 1. Composição dos dados faltantes

4.2 Análise exploratória dos dados

A análise exploratória foi feita separando os dados pela variável *target*, como há uma diferença significativa entre a quantidade de dados os gráficos foram plotados em relação à porcentagem para fins de comparação. Cada gráfico apresenta os dados com a coloração verde para os itens classificados como *good* e em vermelho para os classificados em *bad*.

4.2.1. Gráficos de variáveis qualitativas

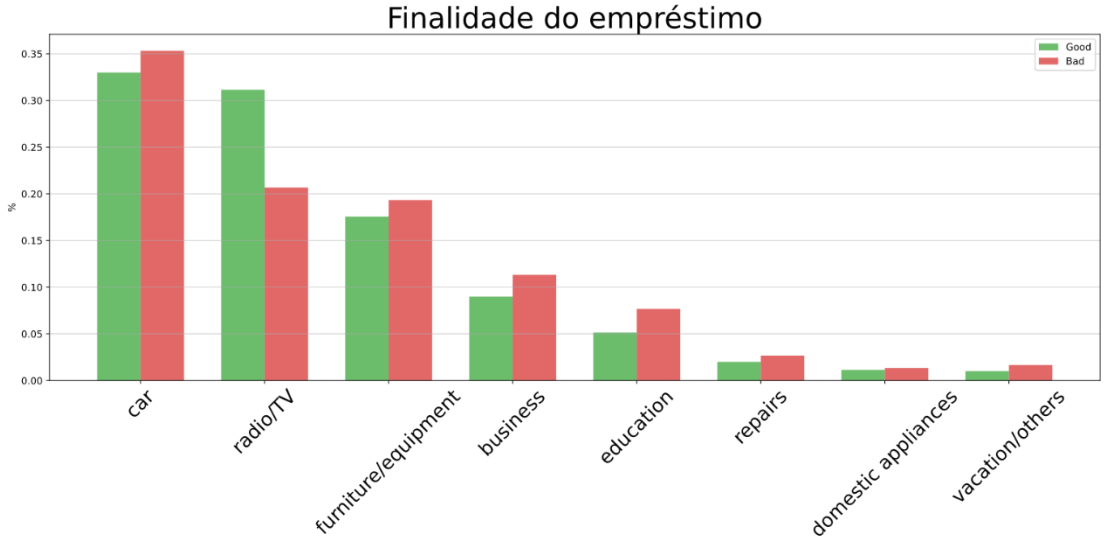


Figura 5. Fonte: Elaboração própria. Variável: *Purpose*

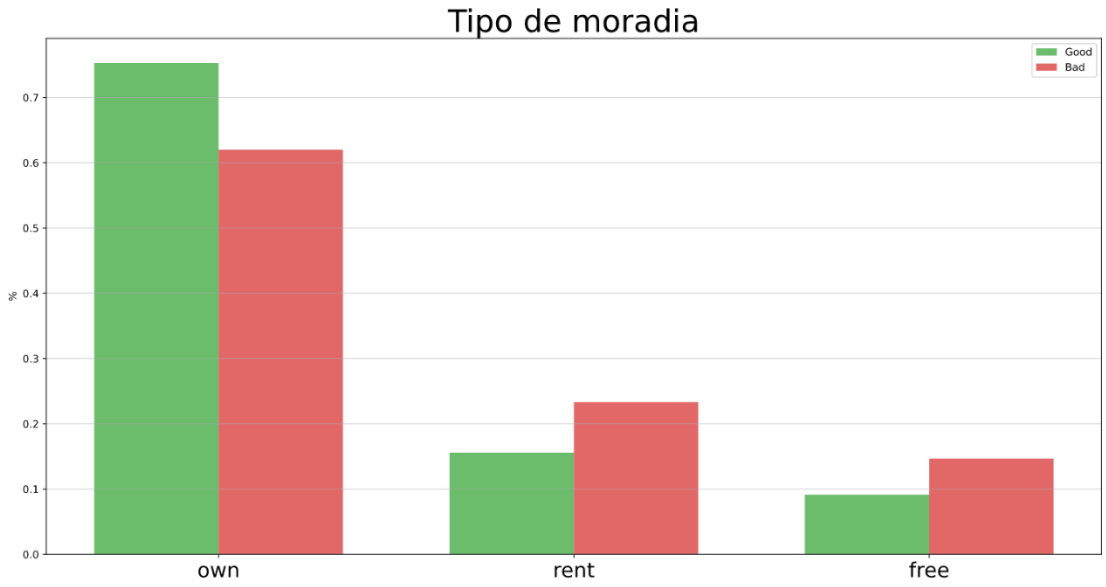


Figura 6. Fonte: Elaboração própria. Variável: *Housing*.

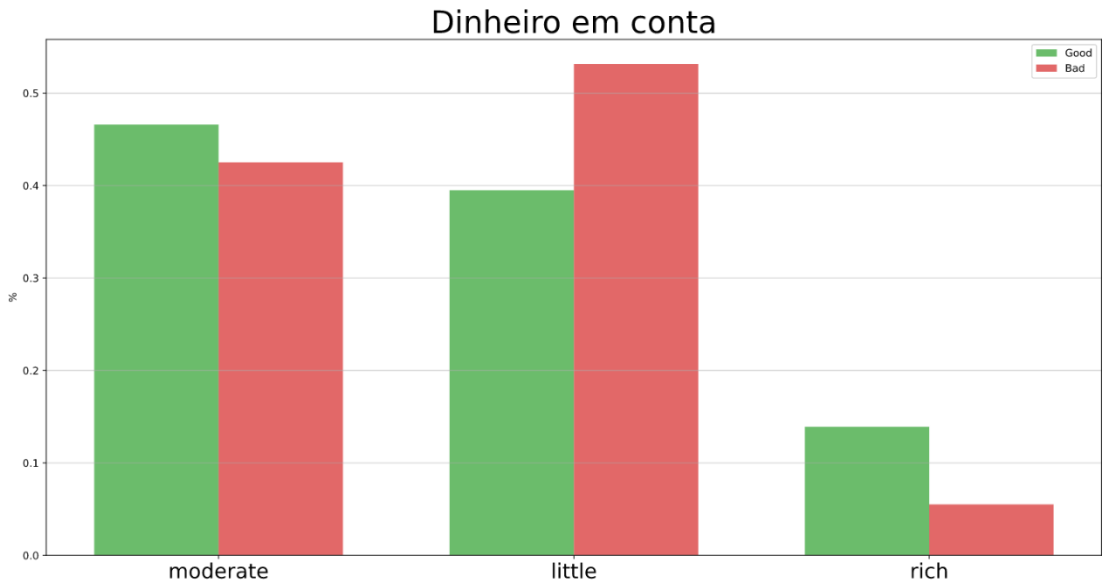


Figura 7. Fonte: Elaboração própria. Variável: *Saving accounts*

4.2.2. Gráficos de variáveis quantitativas

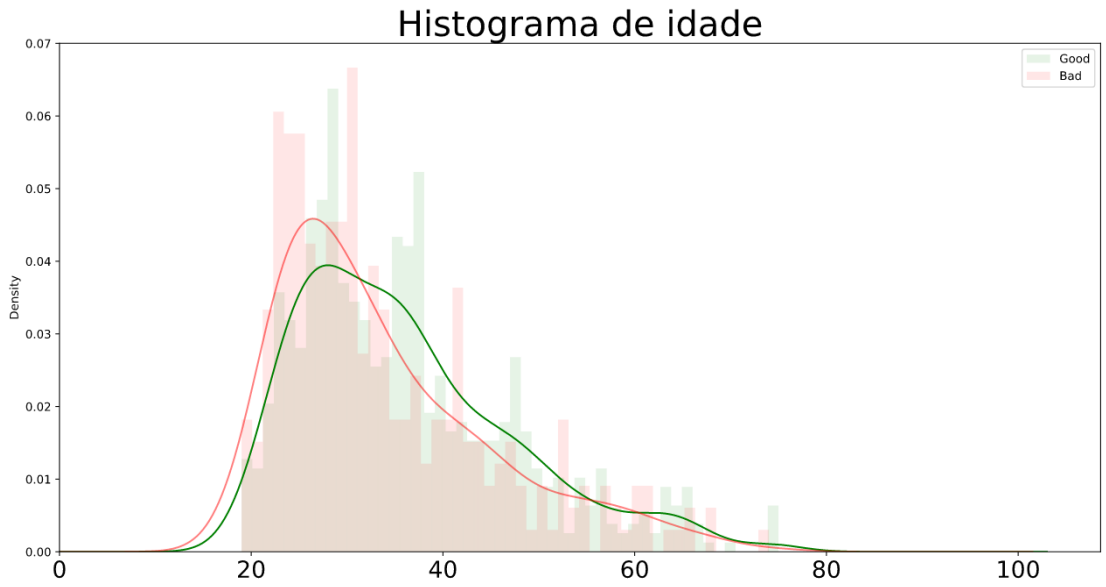


Figura 8. Fonte: Elaboração própria. Variável: *Age*

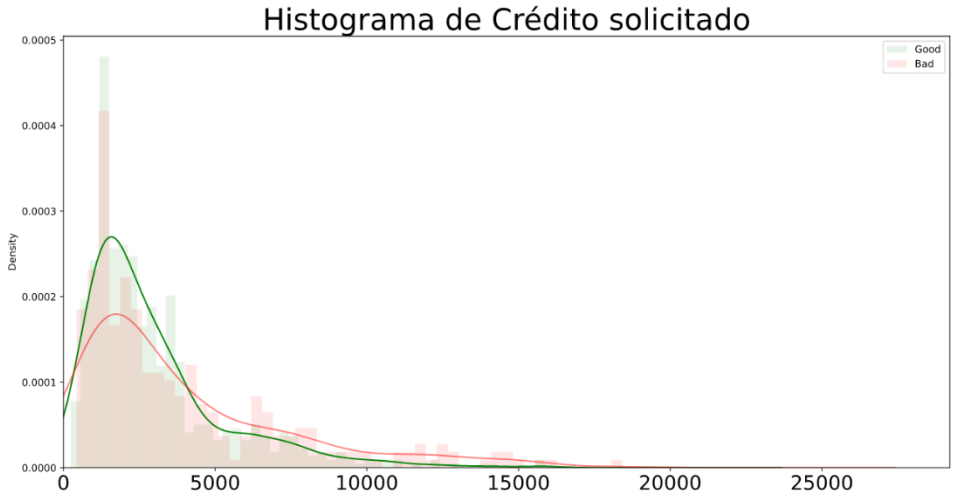


Figura 9. Fonte: Elaboração própria. Variável:*Credit amount.*

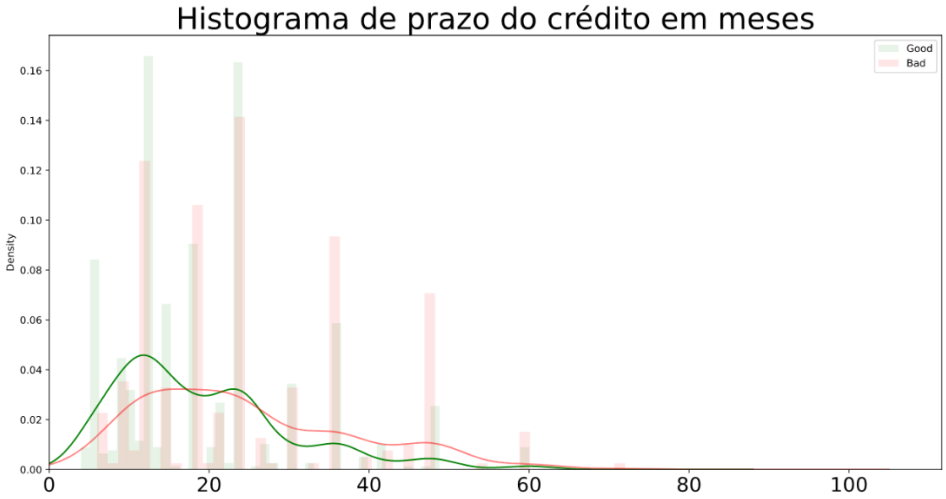


Figura 10. Fonte: Elaboração própria. Variável: Duration

4.2.3 Verificação de correlações entre features e target

Como a variável target possui característica de classe binária sua correlação com variáveis de classe foi verificada pelo teste de Chi-Quadrado. Para isso, foram criadas tabelas de dupla entrada para verificar a distribuição de seus valores. Já para as variáveis numéricas a opção utilizada foi correlação de linear de *Spearman*. Para isso a variável target precisou ser convertida em número, assumindo valor 0 para os valores classificados como *bad* e 1 para os classificados como *good*, conforme Tabela 2. Para testar suas distribuições fora realizado teste de normalidade de *Shapiro* e de diferença de distribuição de *Kolmogorov-Smirnov*.

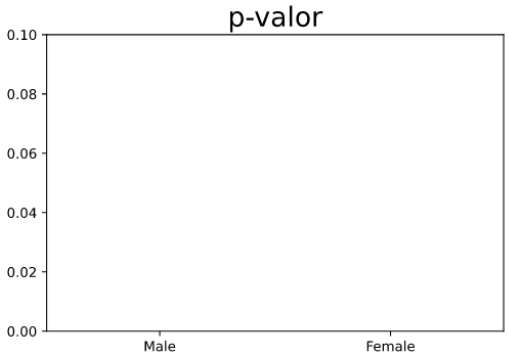
Classificação	
Nominal	Binária
BAD	0
GOOD	1

Tabela 2. Classificação da variável Target

4.2.3.1 Correlação de Chi-Quadrado

Sexo

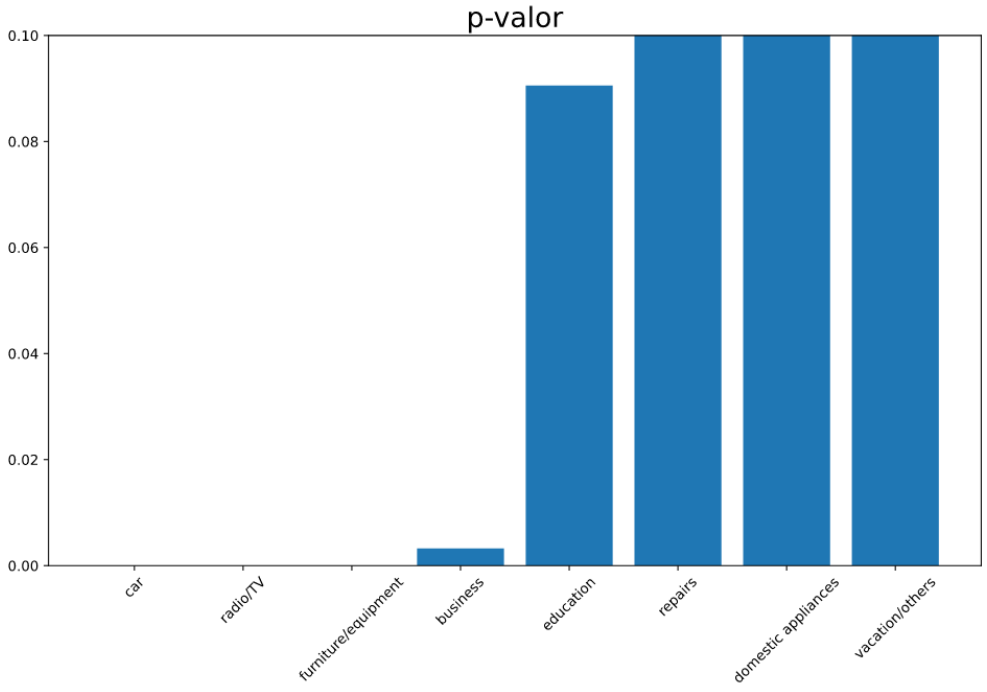
	male	female	Total
good	499	201	700
bad	191	109	300
Total	690	310	1000



A variável possui todos os p-valores abaixo no nível de significância 0.05. Logo, todos os valores da classe possuem alguma relação com a variável *target*.

Finalidade do empréstimo

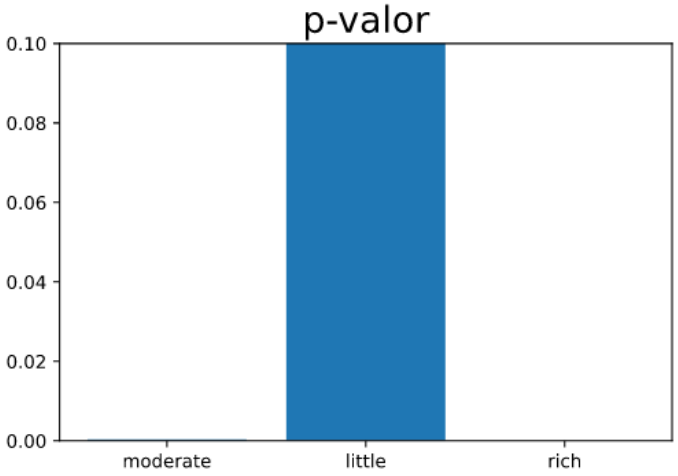
	Car	radio/Tv	furniture/equipament	business	education	repairs	domestic appliances	vacation/others	Total
good	231	218	123	63	36	14	8	7	700
bad	106	62	58	34	23	8	4	5	300
Total	337	280	181	97	59	22	12	12	1000



Podemos perceber que 4 valores da classe ultrapassaram o nível de significância de 0.05, ou seja, os valores não possuem relação com a variável *target.*

Valor em conta

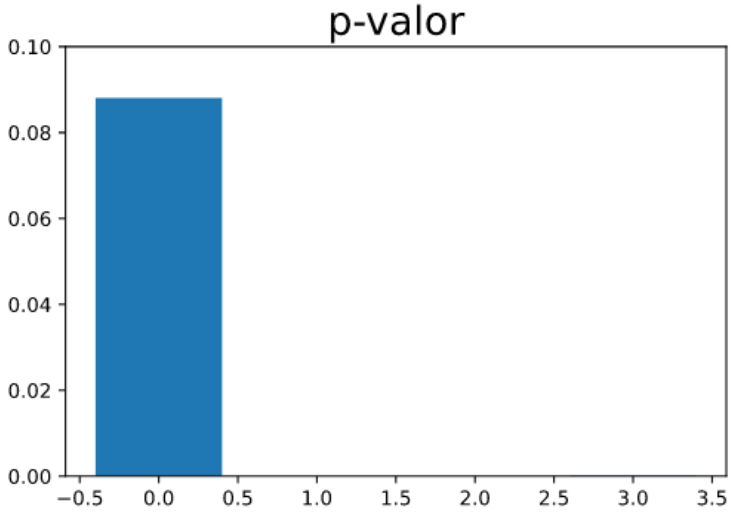
	little	moderate	rich	Total
good	139	164	49	352
bad	135	105	14	254
Total	274	269	63	606



Podemos perceber que 1 valor da classe ultrapassou o nível de significância de 0.05, ou seja, não possui relação com a variável *target.*

Tipo de emprego

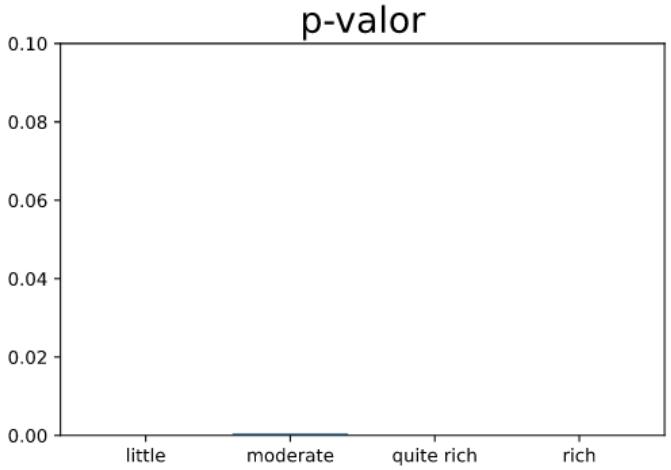
	0	1	2	3	Total
good	15	144	444	97	700
bad	7	56	186	51	300
Total	22	200	630	148	1000



Podemos perceber que 1 valor da classe ultrapassou o nível de significância de 0.05, ou seja, não possui relação com a variável *target.*

Dinheiro guardado

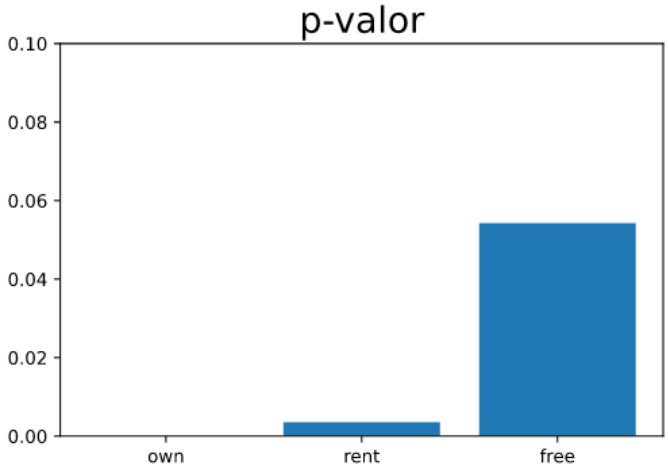
	little	moderate	quite rich	rich	Total
good	386	69	52	42	549
bad	217	34	11	6	268
Total	603	103	63	48	817



Podemos observar que essa variável possui 4 classes e todas elas influenciam no valor da variável *target*, o que pode indicar um poder preditivo para o modelo.

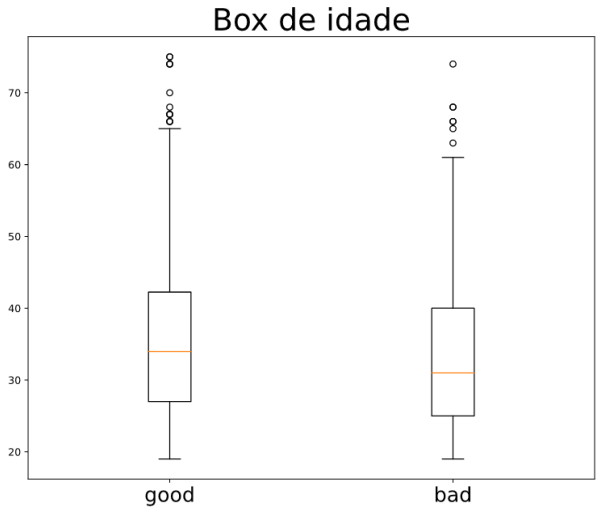
Tipo de moradia

	own	rent	free	Total
good	527	109	64	700
bad	186	70	44	300
Total	713	179	108	1000



A variável também possui bom potencial preditivo para o modelo, apesar de uma classe ultrapassar o nível de significância de 0.05 a diferença é mínima e a rejeição de H0 ainda pode fazer sentido.

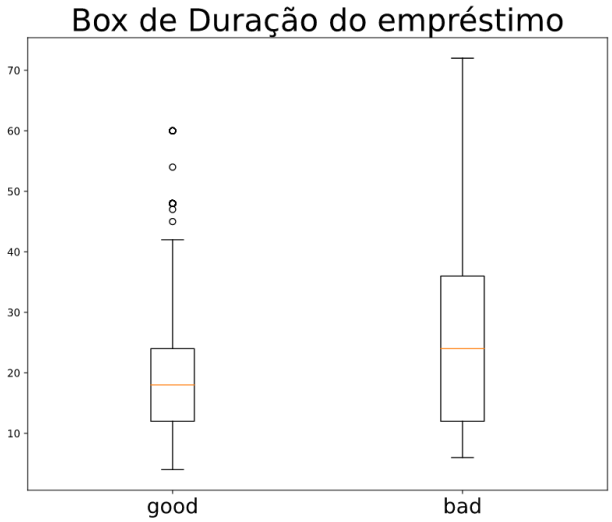
4.2.3.2 Correlações de Pearson *e* Spearman *e testes de* Shapiro *e* Kolmogorov-Smirnov



O teste de *Kolmogorov-Smirnov* retornou um p-valor menor que 0.05, indicando que as distribuições são diferentes e que as pessoas com idade inferior possuem menor chance de ter a solicitação de crédito aprovada em relação às pessoas mais velhas.

O Teste de *Shapiro* informa que nenhuma das duas distribuições são classificadas como *normais*. Por esse motivo a correlação escolhida é a de *Spearman*.

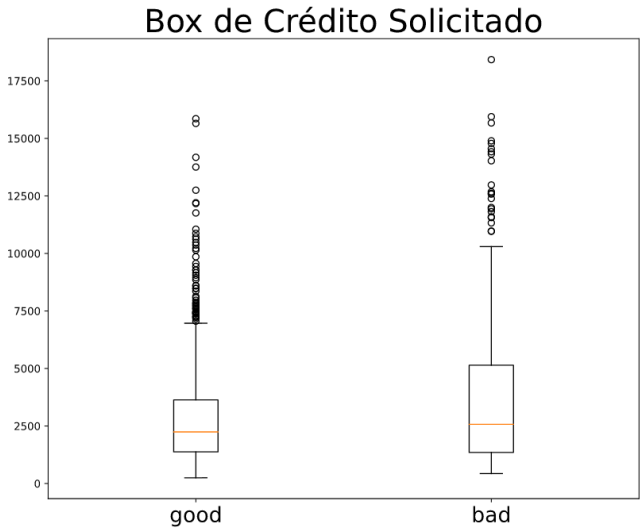
A correlação de *Spearman* indica que a correlação com a variável *target* é de aproximadamente 0.11, considerada *muito fraca*.



O teste de *Kolmogorov-Smirnov* retornou um p-valor menor que 0.05, indicando que as distribuições são diferentes e que as pessoas que solicitam um prazo maior de pagamento possuem maior chance de ter o crédito negado.

O Teste de *Shapiro* informa que nenhuma das duas distribuições são classificadas como *normais*. Por esse motivo a correlação escolhida é a de *Spearman*.

A correlação de *Spearman* indica que a correlação com a variável *target* é de aproximadamente -0.20, considerada *fraca*.



O teste de *Kolmogorov-Smirnov* retornou um p-valor menor que 0.05, indicando que as distribuições são diferentes e que as pessoas que solicitam um montante maior de crédito possuem maior chance de ser negado.

O Teste de *Shapiro* informa que nenhuma das duas distribuições são classificadas como *normais*. Por esse motivo a correlação escolhida é a de *Spearman*.

A correlação de *Spearman* indica que a correlação com a variável *target* é de aproximadamente -0.08, considerada *muito fraca* ou *inexistente*.

4.3. Pré-processamento dos dados

4.3.1. Seleção das features

Com base na análise das correlações as *features* escolhidas são descritas na Tabela 3.

Feature	***Missings***	***Dtype***
Sex	-	object
Housing	-	object
Saving accounts	183	object
Checking account	394	object
Credit amount	-	int64
Duration	-	int64
Purpose	-	object

Tabela 3. Features escolhidas para os modelos

4.3.2. Imputação de valores em dados faltantes

Como observado na Tabela 3 há duas variáveis que possuem valores faltantes, sendo ambas categóricas. O critério para imputação de valores faltantes não possui uma definição clara e a melhor forma de implementação varia de acordo com os critérios adotados pelos realizadores do estudo.

Para este estudo o critério de imputação de dados é descrito na tabela 4.

Critério	Após a imputação a distribuição dos dados deverá respeitar a proporção original de classes separadas pela variável *target*.

Tabela 4. Critérios para imputação de dados

4.3.2.1. Imputação em Saving accounts

A seguir são demonstradas as proporções das classes da variável antes e depois da imputação separadas por variável target.

GOOD		
Classe	**Originais**	**Pós-Imputação**
little	70,31%	70,29%
moderate	12,57%	12,57%
quite rich	9,47%	9,43%
rich	7,65%	7,71%

BAD		
Classe	**Originais**	**Pós-Imputação**
little	80,97%	80,00%
moderate	12,69%	12,67%
quite rich	4,10%	4,67%
rich	2,24%	2,67%

4.3.2.2. Imputação em Checking account

A seguir são demonstradas as proporções das classes da variável antes e depois da imputação separadas por variável target.

GOOD		
Classe	**Originais**	**Pós-Imputação**
little	46,59%	46,71%
moderate	39,49%	39,43%
rich	13,92%	13,86%

BAD		
Classe	**Originais**	**Pós-Imputação**
little	53,15%	51,33%
moderate	41,34%	43,00%
rich	5,51%	5,67%

4.4 Modelos de aprendizado de máquina

4.4.1 Modelos utilizados no estudo

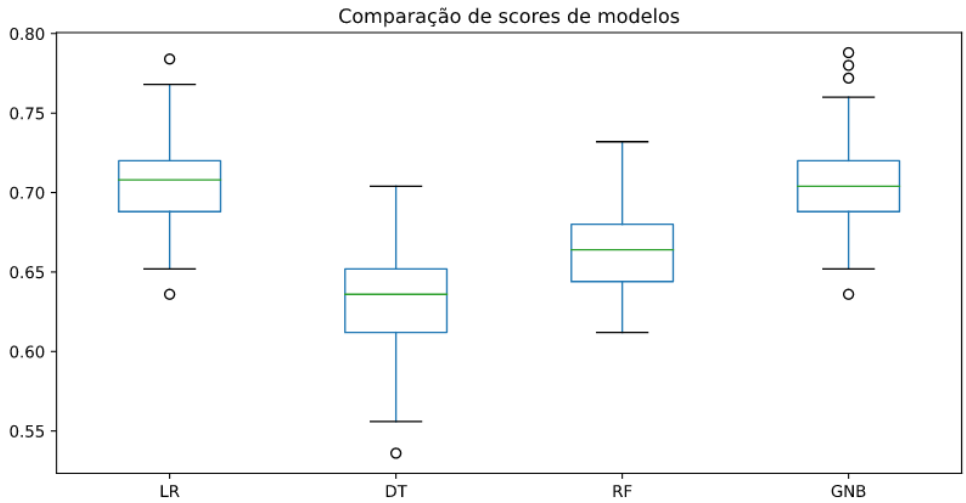
Os modelos de aprendizado de máquina utilizados nesse estudo são descritos na Tabela 5. Todos os modelos são próprios para problemas de classificação com variável supervisionada.

LR	Logistic Regression
DT	Decision Tree
RF	Random Forest
GNB	Gaussian Naive Bayes

Tabela 5. Modelos de aprendizagem de máquina utilizados do estudo.

4.4.2 Resultados e métricas de avaliação

Após realizar diversas iterações os scores dos modelos podem ser vistos a seguir.



Podemos observar que os modelos de RL e GNB foram os que obtiveram os melhores scores, com medianas próximas a 70% de acerto.

Além do *score* outra medida importante são as matrizes de confusão. A seguir podemos observar qual tipo de erro mais comum dos modelos testados.

Os erros do Tipo 1 acontecem quando o modelo classifica como positivo um valor que deveria ser falso, esse tipo de erro também é conhecido como *falso positivo*. Já os erros do Tipo 2 são conhecidos como *falso negativo*, correspondem aos valores que modelo não classificou como válido quando na verdade seriam. A relevância do tipo de erro por vezes não é a mesma, pois, o tipo de erro possui gravidade diferente dependendo da análise. No presente estudo os erros que devem ser evitados são o do Tipo 1, pois isso significa que o crédito está sendo aprovado para um possível mal pagador (*falso positivo*), esse tipo de erro possui um impacto maior no negócio do que se o crédito fosse reprovado para um bom pagador (erro Tipo2).

Analisando o gráfico de erros podemos observar que apesar da LR possuir um *score* melhor que os outros modelos ele também possui o maior índice de erro do Tipo 1, o que não é interessante para a questão do negócio. Em paralelo a isso o modelo GNB possui *score* semelhante à LR e ainda índice menor ocorrência de falsos positivos.

Portanto, conclui-se que o modelo mais eficaz para a questão de avaliação de risco de crédito se trata do modelo *Gaussian Naive Bayes*.

Conclusão

O presente estudo teve a finalidade de avaliar qual modelo de aprendizagem de máquina consegue melhor avaliar o risco de crédito no *dataset german_credit_risk_target*. Foram avaliadas as técnicas de *Logistic Regression, Decision tree, Random Forest* e *Gaussian Naive Bayes*.

Após realização de análise exploratória e pré-processamento dos dados os modelos foram avaliados. Com base nos resultados dos *scores* obtidos e avaliação dos tipos de erros conclui-se que o algoritmo mais adequado para resolução do problema de negócio é o *Gaussian Naive Bayes.*

4. Referências

ANICETO, Maísa Cardoso. Estudo comparativo entre técnicas de aprendizado de máquina para estimação de risco de crédito. 2016.

BRITO, Giovani Antonio Silva; ASSAF NETO, Alexandre. Modelo de classificação de risco de crédito de empresas. Rev. contab. financ., São Paulo, v. 19, n. 46, p. 18-29, Apr. 2008.

FIGUEIRA, Paulo Humberto. Gestão do risco de crédito: análise dos impactos da resolução 2682, do conselho monetário nacional, na transparência do risco da carteira de empréstimo dos bancos comerciais brasileiros. Dissertação (Mestrado em Gestão Empresarial) - FGV - Fundação Getúlio Vargas, Rio de Janeiro, 2001.

BLATT, Adriano. Avaliação de risco e decisão de crédito. São Paulo: Nobel, 1999.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. Data mining: concepts and techniques. Elsevier, 2011.