
Universidade Federal
de Ouro Preto

# Algoritmo Classificação

**Gabriel Costa, Lucas Rodrigues,**
**Alexsandro, Jean Pierre**

# 1 Introdução

O conjunto de dados **German Credit** é um conjunto de dados disponível publicamente baixado do UCI Machine Learning Repository.

Os empréstimos fazem parte integrante das operações bancárias. No entanto, nem todos os empréstimos são prontamente devolvidos e, portanto, é importante que um banco monitore de perto seus pedidos de empréstimo. Este projeto é uma análise dos dados de crédito alemães. Ele contém detalhes de 1.000 solicitantes de empréstimo com 20 atributos e a classificação se um solicitante é considerado um risco de crédito bom ou ruim.

Neste projeto, a relação entre o risco de crédito e vários atributos será explorada por meio de técnicas estatísticas básicas e apresentada por meio de visualizações.

# 2 Preparação e Limpeza dos Dados

Algum NaN ou None no DataFrame: False

O conjunto de dados contém 21 variáveis e 1000 observações. 8 variáveis são do tipo numérico e 13 do tipo objeto. Como as variáveis do tipo objeto não possuem nenhum valor nulo, podemos concluir que elas são do tipo categórico.

Em seguida, vamos rotular as variáveis para facilitar o uso. O documento que descreve o conjunto de dados pode ser consultado para isso:

Com base na descrição, nomeamos as colunas:

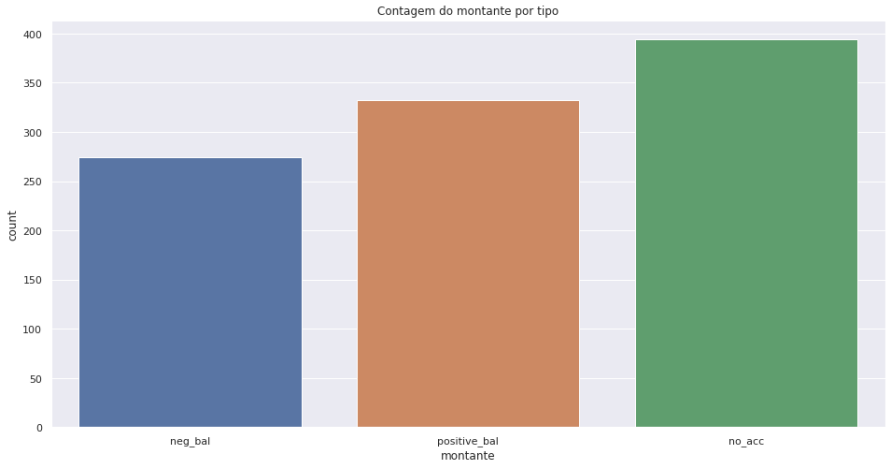
```
df.columns =
        ["montante", "duracao", "historico_credito",
         "proposito", "montante_credito", "poupanca",
         "tempo_empregado", "taxa_parcelamento","risco"
         "estado_civil_sexo", "tipo_participacao_credito",
         "tempo_moradia", "propriedade", "idade", "emprego",
         "gastos_adicionais", "habitacao", "quantidade_creditos",
         "dependentes", "telefone", "trabalhador_estrangeiro"]
```

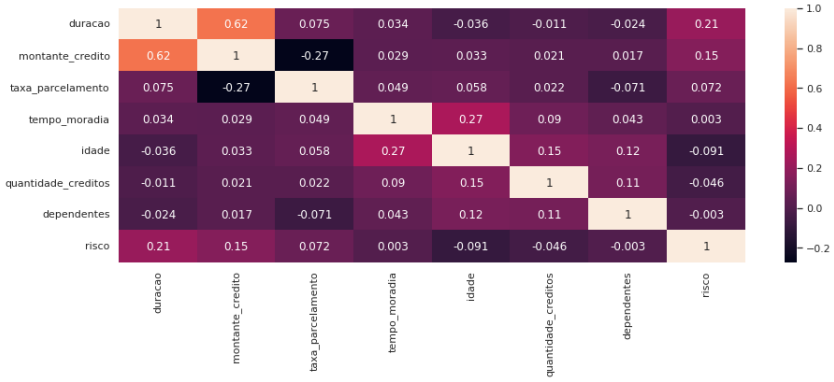
# 3 Análise e visualização dos dados:

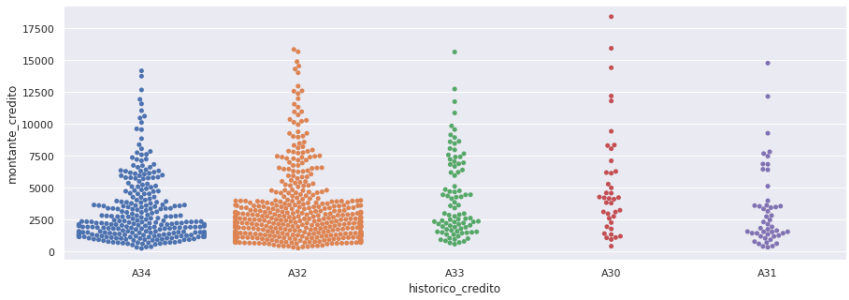
**Análise de correlação entre as variáveis do dataset:**

**Examinando a distribuição da coluna de risco:**

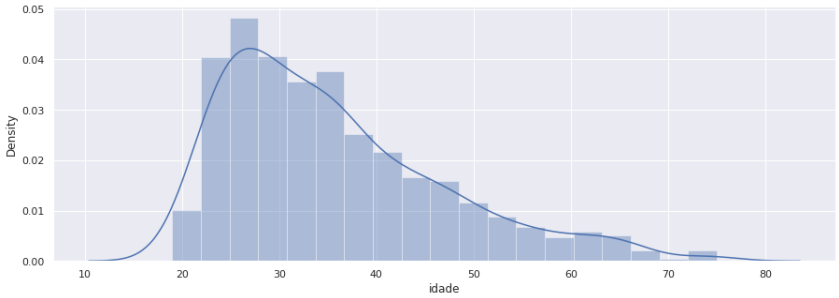
A coluna risco tem dois valores:
1: representando um bom empréstimo
2: representando um mau empréstimo (default).
A convenção usual é usar '1' para empréstimos ruins e '0' para empréstimos bons. Vamos substituir os valores para cumprir a convenção.



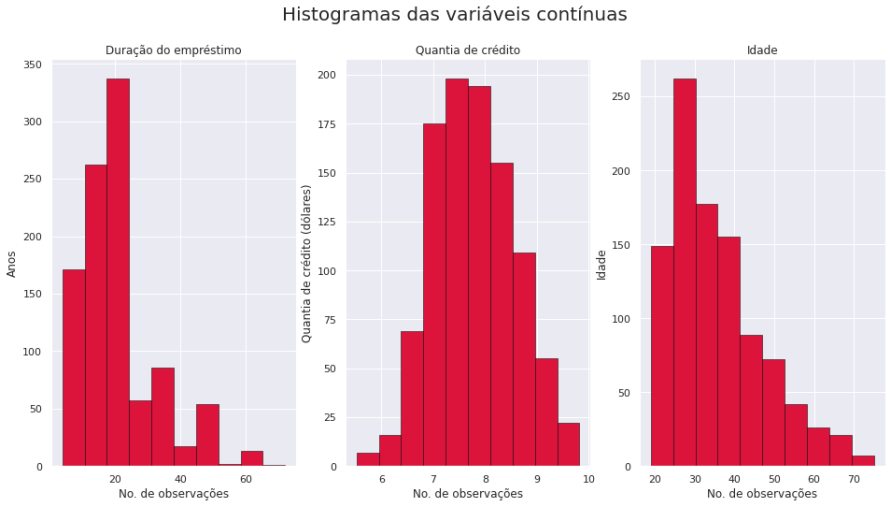

```
0    0
1    1
2    0
3    0
4    1
Name: risco, dtype: int64
```




**Exploração das variáveis contínuas:**

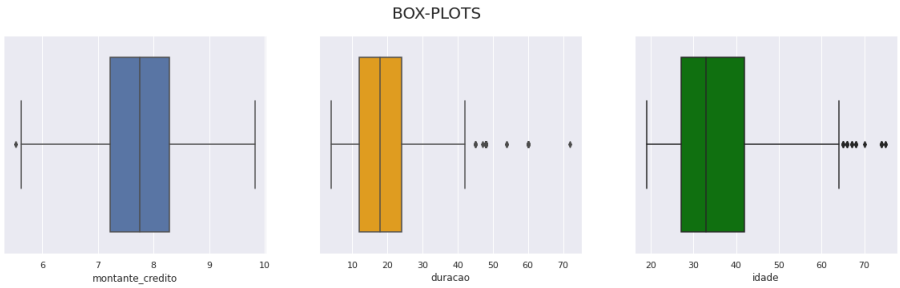
**Utilizaremos:** Estatísticas resumidas, Histogramas, Box-plots

**Observações:** Uma olhada na distribuição das variáveis contínuas mostra que as variáveis estão em intervalos diferentes. O histograma sugere que a maioria das observações cai no primeiro quantil da variável. Isso pode ser verificado pelo box-plot. Os box-plots mostram que a maioria dos valores dos créditos estão entre 1000 a 4500 dólares. O valor do crédito é positivamente enviesado. A maior parte da duração do empréstimo é de 15 a 30 meses. A maioria dos requerentes de empréstimo tem idade entre 28 e 43 anos.

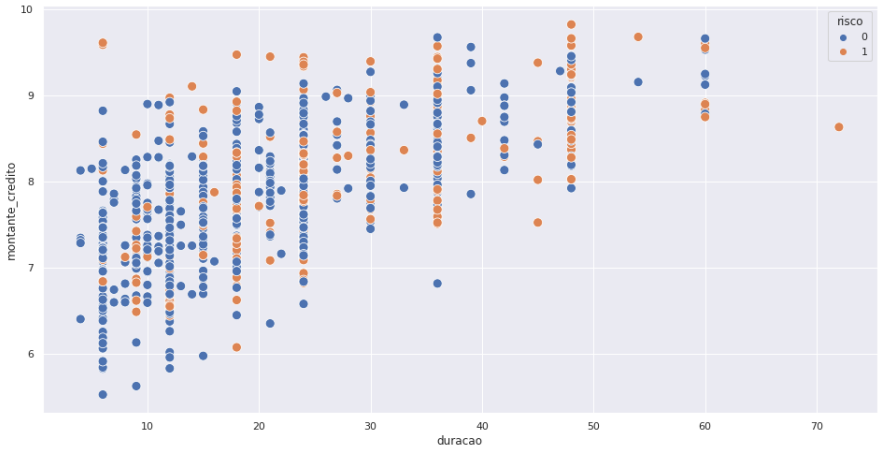



**Relação entre o valor do crédito e a duração do reembolso:**

**Utilizaremos:** Gráfico de dispersão




**Observações:** O gráfico de dispersão mostra que, em geral, empréstimos maiores têm maior duração de reembolso. Casos em que grandes empréstimos são concedidos com curto prazo de reembolso acabaram por ser maus empréstimos.




**Exploração de variáveis categóricas:**

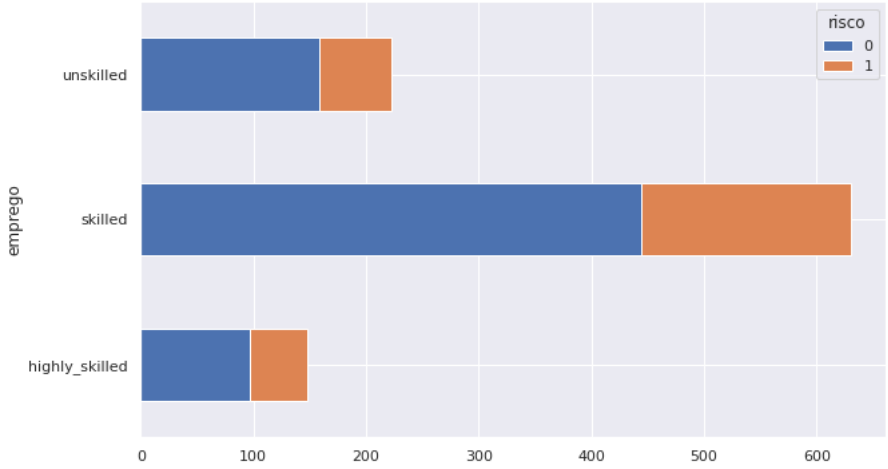
**Relação entre risco de crédito e habilidades do solicitante do empréstimo:**

**Utilizaremos:** Gráfico de barras

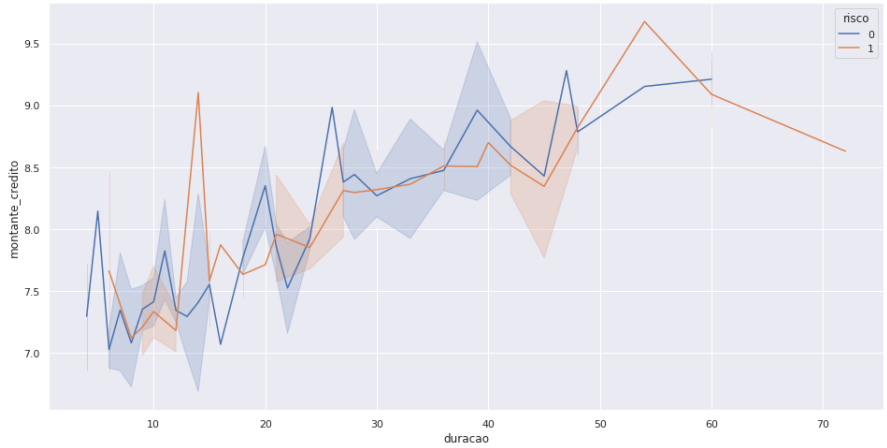
**Observações:** O gráfico mostra que os candidatos desempregados/não qualificados representam um alto risco.

**Relação entre o valor do crédito e a duração do empréstimo:**

**Utilizaremos:** Gráfico de linha




**Observação:** Existe uma relação linear entre o valor do crédito e a duração. Quanto maior o valor do crédito, maior é a duração do reembolso.

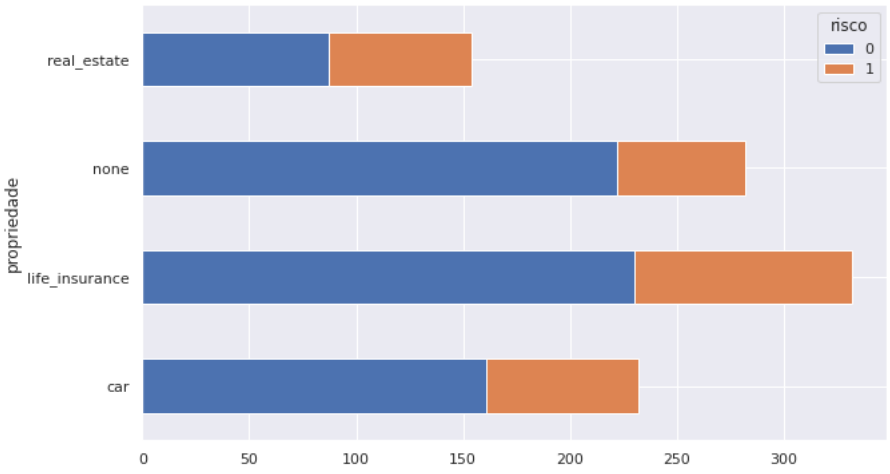



**Relação entre o ativo mais valioso do candidato e o valor do crédito, risco de crédito:**

**Utilizaremos:** Gráfico de barras empilhadas, Gráfico de dispersão

**A codificação categórica usada nos gráficos é:**
A121: imóveis
A122: se não A121: convênio de poupança/seguro de vida
A123: se não A121/A122: carro ou outro, não no atributo 6
A124: desconhecido/sem propriedade

**Observações:** Os gráficos mostram que as pessoas com ativos imobiliários são muito arriscadas.




# 4 Codificar variáveis categóricas

A maioria dos modelos de aprendizado de máquina não pode lidar com variáveis categóricas. Portanto, precisamos codificar as 13 variáveis categóricas que temos no conjunto de dados alemão.

Temos variáveis categóricas com 2 a 10 categorias. Vamos para a codificação de rótulo para variáveis com apenas duas categorias, enquanto para variáveis com mais de duas categorias, vamos para a codificação one-hot. Na codificação de rótulo, atribuímos cada categoria exclusiva em uma variável categórica com um número inteiro. Nenhuma nova coluna é criada. Na codificação one-hot, criamos uma nova coluna para cada categoria exclusiva em uma variável categórica. A única desvantagem da codificação one-hot é que o número de recursos (dimensões dos dados) pode explodir com variáveis categóricas com muitas categorias. Para lidar com isso, podemos executar a codificação one-hot seguida de PCA ou outros métodos de redução de dimensionalidade para reduzir o número de dimensões (enquanto ainda tentamos preservar as informações).

Para codificação de rótulo, usamos o LabelEncoder da biblioteca 'Scikit-Learn' e para codificação one-hot, a função 'get_dummies(df)' da biblioteca 'pandas'.

Agora que codificamos as variáveis, vamos continuar com a análise exploratória dos dados.

**Correlação entre as variáveis:** Vejamos as correlações entre os recursos e o destino usando o coeficiente de correlação de Pearson. Neste caso, uma

correlação positiva representa a correlação com a inadimplência de crédito, enquanto uma correlação negativa representa a correlação com o reembolso do crédito.

**Observações:**

**Correlação positiva:** Pessoas com contas correntes com saldo negativo (account_bal_A11) provavelmente deixarão de pagar o empréstimo.

Empréstimos de maior duração (duration) tendem a ficar inadimplentes.

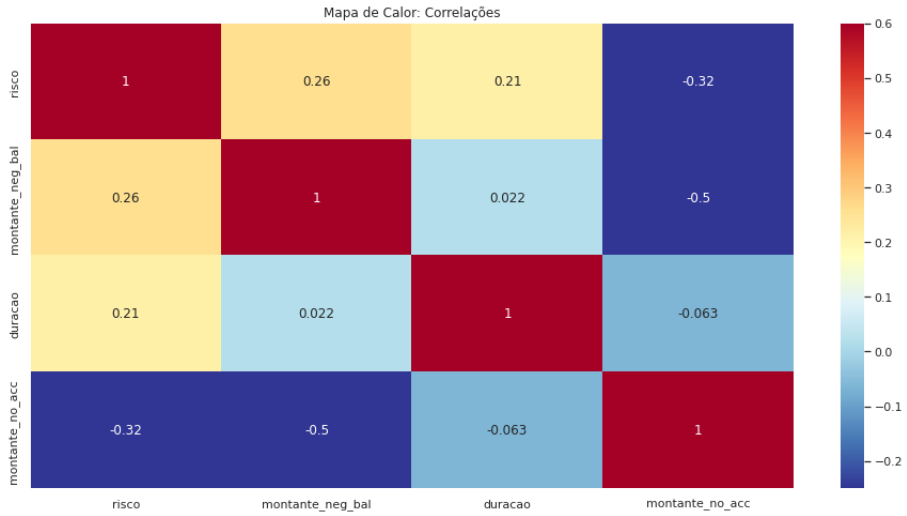
**Correlação negativa:** Pessoas sem conta corrente (account_bal_A14) provavelmente pagarão o empréstimo.

**Correlações positivas:**

```
estado_civil_sexo_A92   0.075493
habitacao_A153          0.081556
montante_positive_bal   0.089895
habitacao_A151          0.092785
gastos_adicionais_A141  0.096510
proposito_A40           0.096900
tempo_empregado_A72     0.106397
montante_credito        0.109570
propriedade_real_estate 0.125750
historico_credito_A31   0.134448
historico_credito_A30   0.144767
poupanca_A61            0.161007
duracao                 0.214927
montante_neg_bal        0.258333
risco                   1.000000
Name: risco, dtype: float64
```

**Correlações negativas:**

```
montante_no_acc         -0.322436
historico_credito_A34   -0.181713
habitacao_A152          -0.134589
poupanca_A65            -0.129238
propriedade_none        -0.119300
gastos_adicionais_A143  -0.113285
proposito_A43           -0.106922
proposito_A41           -0.099791
idade                   -0.091127
poupanca_A64            -0.085749
trabalhador_estrangeiro -0.082079
estado_civil_sexo_A93   -0.080677
tempo_empregado_A74     -0.075980
poupanca_A63            -0.070954
tempo_empregado_A75     -0.059733
```




**Vejamos o mapa de calor de correlações significativas:**

# 5 Engenharia de recursos:

A engenharia de recursos refere-se à criação de recursos mais úteis a partir dos dados. Isso representa um dos padrões do aprendizado de máquina: a engenharia de recursos tem um retorno sobre o investimento maior do que a construção de modelos e o ajuste de hiperparâmetros.

A engenharia de recursos refere-se a um processo geral e pode envolver tanto a construção de recursos: adicionar novos recursos a partir dos dados existentes, quanto a seleção de recursos: escolher apenas os recursos mais importantes ou outros métodos de redução de dimensionalidade. Existem muitas técnicas que podemos usar para criar recursos e selecionar recursos.

Para este problema, tentaremos construir características polinomiais.

**Recursos polinomiais:** Aqui, encontramos interações entre os recursos significativos. A correlação entre os recursos de interação são verificados. Se os recursos de interação tiverem maior correlação com o destino em comparação com os recursos originais, eles serão incluídos no modelo de aprendizado de máquina, pois podem ajudar o modelo a aprender melhor.

# 6 Modelos:

**Critério de avaliação:**

**Vamos dar uma olhada nas diferentes opções disponíveis de classificação:**

| Critério de avaliação | Descrição |
|---|---|
| Acurácia | (Verdadeiro positivo + Verdadeiro negativo) / Total observações |
| Precisão | Verdadeiro positivo / Total da predição positiva |
| Recall | Verdadeiro positivo / Total dos atuais positivos |
| F1 | 2 * Precisão * Recall / (Precisão + Recall) |
| AUC ROC | Área abaixo da curva ROC (TPR Vs. FPR para todos os limites de classificação) |

**Critério de avaliação:**

**Descrição:**
Acurácia (Verdadeiro positivo + Verdadeiro negativo) / Total observações
Precisão Verdadeiro positivo / Total da predição positiva
Recall Verdadeiro positivo / Total dos atuais positivos
F1 2 * Precisão * Recall / (Precisão + Recall)
AUC ROC Área abaixo da curva ROC (TPR Vs. FPR para todos os limites de classificação)

**Acurácia:** O conjunto de dados alemão é um conjunto de dados desequilibrado. A acurácia daria uma pontuação alta ao prever a classe majoritária, mas não conseguiria prever a classe minoritária, que são os inadimplentes. Portanto, essa não é uma métrica adequada para esse conjunto de dados.

**Precisão:** A precisão é uma boa métrica quando os custos de falsos positivos são altos. Exemplo, detecção de spam de e-mail.

**Recall:** Esta métrica é adequada quando os custos de falsos negativos são altos. Exemplo, prevendo um inadimplente como não inadimplente. Isso custa enorme perda para o banco. Portanto, esta é uma métrica adequada para o nosso caso.

**F1:** Medida de precisão e recall.

**AUC ROC:** É o gráfico de TPR vs FPR. Todos os outros critérios discutidos aqui assumem 0,5 como o limite de decisão para a classificação. No entanto, pode não ser sempre verdade. A AUC nos ajuda a avaliar o desempenho do modelo para todos os limites de classificação. Quanto maior o valor da métrica AUC, melhor o modelo.

Taxa de verdadeiro positivo (TPR) = TP / Total real positivo
Taxa de falsos positivos (FPR) = FP / Total real negativo
Usaremos a acurácia e o F1 como critério de métrica.

**Linha de base:**

```
0    0.7
1    0.3
```

Isso significa que a acurácia da linha de base é de 70%, ou seja, mesmo classificando todas as amostras como inadimplentes, teremos 70% de acurácia.

# 7 Classificadores :

**7.1 Random Forest :** Uma floresta aleatória é um meta estimador que ajusta vários classificadores de árvore de decisão em várias subamostras do conjunto de dados e usa a média para melhorar a precisão preditiva e controlar o ajuste excessivo. O tamanho da subamostra é controlado com o max samples parâmetro if bootstrap = True(padrão), caso contrário, todo o conjunto de dados é usado para construir cada árvore.

**Max Depth** A profundidade máxima da árvore. Se Nenhum, os nós são expandidos até que todas as folhas sejam puras ou até que todas as folhas contenham menos de min samples split amostras.

**Min Samples Leaf** O número mínimo de amostras necessárias para estar em um nó folha. Um ponto de divisão em qualquer profundidade só será considerado se deixar pelo menos min samples leafamostras de treinamento em cada um dos ramos esquerdo e direito. Isso pode ter o efeito de suavizar o modelo, especialmente na regressão.

**7.2 Decision Tree :** É um método de aprendizado supervisionado não paramétrico usado para classificação e regressão . O objetivo é criar um modelo que preveja o valor de uma variável de destino aprendendo regras de decisão simples inferidas a partir dos recursos de dados. Uma árvore pode ser vista como uma aproximação constante por partes.

**7.3 Logistic Regression :** A regressão logística é um processo de modelagem da probabilidade de um resultado discreto dada uma variável de entrada. Os modelos de regressão logística mais comunsum resultado binário; algo que pode assumir dois valores, como verdadeiro/falso, sim/não e assim por diante. A regressão logística multinomial pode modelar cenários onde há mais de dois resultados discretos possíveis. A regressão logística é um método de análise útil para problemas de classificação, onde você está tentando determinar se uma nova amostra se encaixa melhor em uma categoria. Como aspectos de segurança cibernética são problemas de classificação, como detecção de ataques, a regressão logística é uma técnica analítica útil.

# 8 Referências :


- Livro → Practical Time Series Analysis: Prediction with Statistics and Machine Learning - Aileen Nielsen
- Livro → Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython - *Wes McKinney*
- Kaggle - https://www.kaggle.com/
- Medium - https://medium.com/
- GeeksForGeeks → https://www.geeksforgeeks.org/
- Towards Data Science → https://towardsdatascience.com/