MODELO PARA A ENTREGA DAS ATIVIDADES

COMPONENTE CURRICULAR: Aprendizado de Máquina

NOME COMPLETO DO ALUNO: Willians Carvalho da Silva -10416087

João Pedro Santos Oliveira – 10423752

Costas-10289655

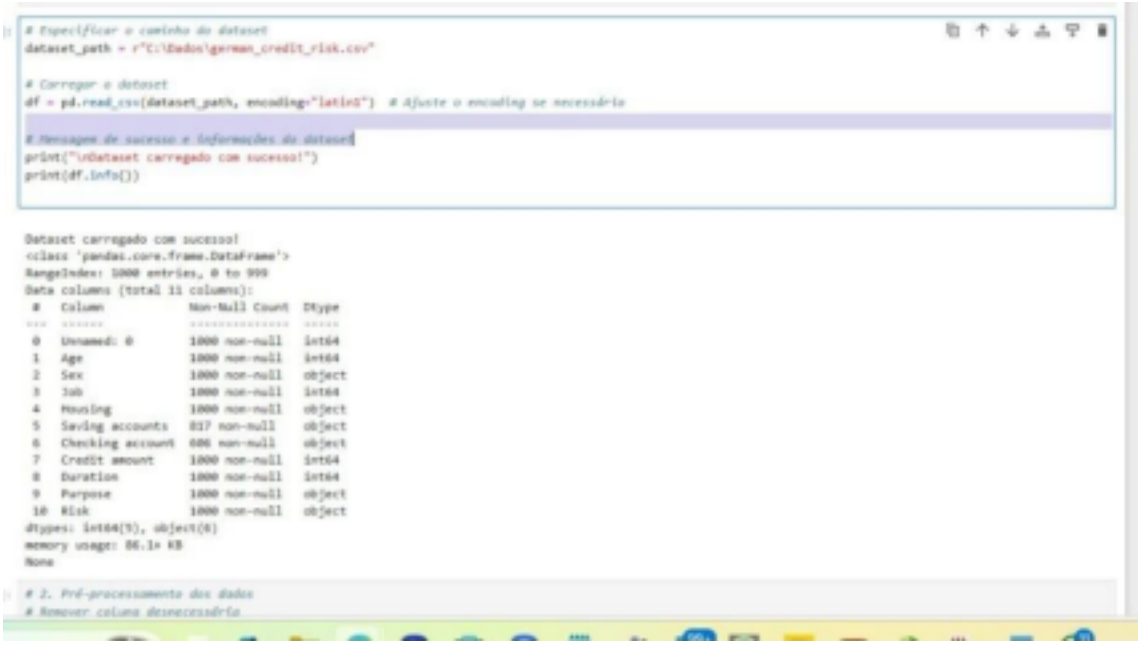
Cesar Valentim Silva – 10416087

Luciano Guimaraes

A análise de risco de crédito representa um dos principais desafios no setor financeiro, dado o impacto direto na sustentabilidade econômica das instituições e na experiência dos clientes. Neste contexto, a utilização de técnicas de aprendizado de máquina tem se mostrado uma ferramenta poderosa para identificar padrões e prever comportamentos, contribuindo para uma tomada de decisão mais assertiva. Este trabalho tem como objetivo aplicar algoritmos supervisionados, como Árvore de Decisão e Random Forest, para a análise preditiva e exploratória de uma base de dados desbalanceada de crédito, composta por variáveis demográficas, econômicas e financeiras. A partir da avaliação de métricas e visualizações gráficas, busca-se compreender as relações entre as variáveis, os fatores determinantes do risco e as limitações dos modelos utilizados. Este estudo não apenas explora as nuances de um problema complexo, mas também propõe estratégias para a melhoria da acurácia e da eficácia dos modelos, destacando a relevância da análise de dados como um pilar na gestão de risco e tomada de decisão.

```
# Importação de bibliotecas

import pandas as pd

import numpy as np
```

```
from sklearn.model_selection import train_test_split

from sklearn.preprocessing import LabelEncoder

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification_report, confusion_matrix, accuracy_score


# Especificar o caminho do dataset

dataset_path = r"C:\Dados\german_credit_risk.csv"


# Carregar o dataset

 df = pd.read_csv(dataset_path, encoding="latin1") # Ajuste o encoding se necessário


# Mensagem de sucesso e informações do dataset

print("\nDataset carregado com sucesso!")

print(df.info())
```



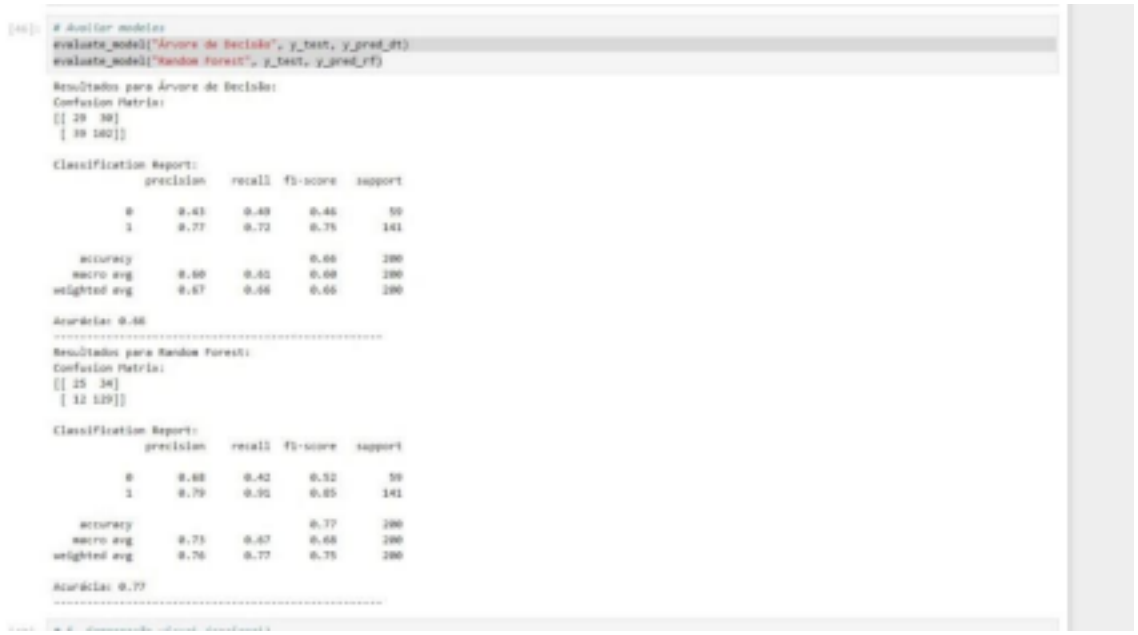

```
# 2. Pré-processamento dos dados
# Remover coluna desnecessária

df.drop(columns=['Unnamed: 0'], inplace=True)



# Tratar valores ausentes

df.fillna({'Saving accounts': 'unknown', 'Checking account': 'unknown'}, inplace=True)

# Converter variáveis categóricas para numéricas

label_encoders = {}

for column in ['Sex', 'Housing', 'Saving accounts', 'Checking account', 'Purpose', 'Risk']:

le = LabelEncoder()

 df[column] = le.fit_transform(df[column])

 label_encoders[column] = le



# Dividir dados em variáveis preditoras (X) e alvo (y)
```


```
X = df.drop(columns=['Risk']) # 'Risk' é a variável alvo

y = df['Risk']

# Divisão em treino e teste
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# 3. Treinamento de modelos

# Modelo 1: Árvore de Decisão

dt_model = DecisionTreeClassifier(random_state=42)

dt_model.fit(X_train, y_train)

y_pred_dt = dt_model.predict(X_test)

# Modelo 2: Random Forest

rf_model = RandomForestClassifier(random_state=42, n_estimators=100)

rf_model.fit(X_train, y_train)

y_pred_rf = rf_model.predict(X_test)

# 4. Avaliação dos modelos

# Função para exibir métricas

def evaluate_model(model_name, y_test, y_pred):

 print(f"Resultados para {model_name}:")

 print("Confusion Matrix:")

 print(confusion_matrix(y_test, y_pred))

 print("\nClassification Report:")

 print(classification_report(y_test, y_pred))

 print(f"Acurácia: { accuracy_score(y_test, y_pred):.2f}")
```


```
 print("-" * 50)

# Avaliar modelos

evaluate_model("Árvore de Decisão", y_test, y_pred_dt)

evaluate_model("Random Forest", y_test, y_pred_rf)

# 5. Comparação visual (opcional)
```

import matplotlib.pyplot as plt



Importâncias das features (Random Forest)

feature_importances = pd.DataFrame({

'Feature': X.columns,

'Importance': rf_model.feature_importances_

}).sort_values(by='Importance', ascending=False)

plt.barh(feature_importances['Feature'], feature_importances['Importance'])

plt.xlabel('Importância')

plt.ylabel('Feature')

plt.title('Importância das Features - Random Forest')

plt.show()

import seaborn as sns

from sklearn.metrics import confusion_matrix

Matriz de confusão para o modelo Random Forest

```
cm = confusion_matrix(y_test, y_pred_rf)


plt.figure(figsize=(6, 4))

sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')

plt.title('Matriz de Confusão - Random Forest')

plt.xlabel('Predito')

plt.ylabel('Real')

plt.show()

from sklearn.metrics import roc_curve, roc_auc_score


# Probabilidades preditas pelo modelo Random Forest

y_proba_rf = rf_model.predict_proba(X_test)[:, 1]

fpr, tpr, thresholds = roc_curve(y_test, y_proba_rf)


plt.figure(figsize=(6, 4))
```


```
plt.plot(fpr, tpr, label=f"AUC: {roc_auc_score(y_test, y_proba_rf):.2f}")

plt.plot([0, 1], [0, 1], 'k--', label='Random Guessing')

plt.xlabel('False Positive Rate')

plt.ylabel('True Positive Rate')

plt.title('Curva ROC - Random Forest')

plt.legend()

plt.show()


# Distribuição das classes
```

```
df['Risk'].value_counts().plot(kind='bar', color=['skyblue', 'orange'], figsize=(6, 4))
plt.title('Distribuição de Risco na Base de Dados')
plt.xlabel('Classe (Risk)')
plt.ylabel('Número de Instâncias')
plt.show()


# Boxplot de Credit amount por Risk
plt.figure(figsize=(6, 4))
sns.boxplot(data=df, x='Risk', y='Credit amount', palette='viridis')
plt.title('Distribuição do Valor de Crédito por Risco')
plt.xlabel('Classe (Risk)')
plt.ylabel('Credit amount')
plt.show()
```


```
plt.figure(figsize=(8, 5))
sns.boxplot(data=df, x='Risk', y='Credit amount', palette='viridis')
plt.title('Distribuição do Valor de Crédito por Risco')
plt.xlabel('Classe (Risk)')
plt.ylabel('Valor do Crédito (Credit amount)')
plt.show()
plt.figure(figsize=(8, 5))
sns.kdeplot(data=df[df['Risk'] == 1], x='Credit amount', label='Risco Baixo', shade=True, color='green')
sns.kdeplot(data=df[df['Risk'] == 0], x='Credit amount', label='Risco Alto', shade=True, color='red')
```

```
plt.title('Distribuição de Valores de Crédito por Classe de Risco')

plt.xlabel('Valor do Crédito (Credit amount)')

plt.ylabel('Densidade')

plt.legend()

plt.show()

risk_housing = df.groupby(['Housing', 'Risk']).size().unstack()

plt.figure(figsize=(8, 5))

sns.heatmap(risk_housing, annot=True, fmt='d', cmap='coolwarm')

plt.title('Concentração de Risco por Tipo de Moradia')

plt.xlabel('Risco')

plt.ylabel('Tipo de Moradia (Housing)')
```


```
plt.show()

risk_purpose = df.groupby(['Purpose', 'Risk']).size().unstack()

risk_purpose_norm = risk_purpose.div(risk_purpose.sum(axis=1), axis=0)


risk_purpose_norm.plot(kind='bar', stacked=True, figsize=(10, 6), colormap='viridis')

plt.title('Proporção de Risco por Finalidade do Crédito')

plt.xlabel('Finalidade (Purpose)')

plt.ylabel('Proporção')

plt.legend(['Risco Alto', 'Risco Baixo'], title='Risco')

plt.show()

# Criar o pairplot

plot = sns.pairplot(df, vars=['Age', 'Credit amount', 'Duration'], hue='Risk',
palette='viridis')
```

```
# Adicionar o título vertical na lateral direita
plot.fig.text(1.02, 0.5, 'Relação entre Variáveis Numéricas e o Risco',
rotation=270, va='center', fontsize=12)

# Exibir o gráfico
plt.show()
```

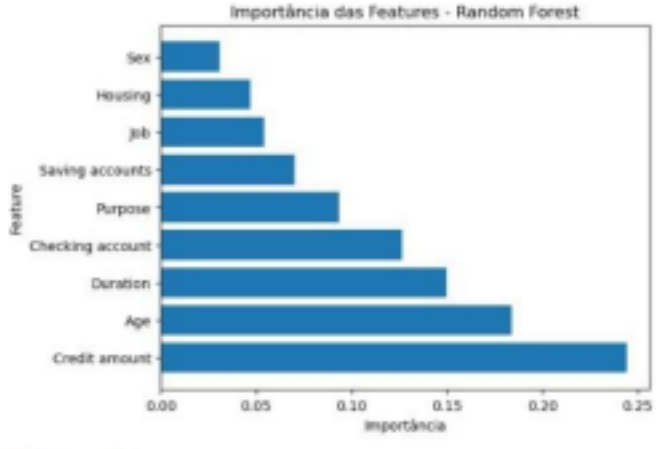
interpretação Geral

As três variáveis mais importantes (Credit amount, Age, e Duration) são diretamente relacionadas ao crédito solicitado e ao perfil financeiro dos clientes. Essas variáveis são centrais para a avaliação do risco.

Variáveis como Housing e Sex têm pouco impacto no modelo, sugerindo que podem ser menos úteis para a tomada de decisão.

A importância de Saving accounts e Checking account reforça a relevância de informações financeiras para prever a capacidade de pagamento.

```
plt.barh(feature_importances["Feature"], feature_importances["Importance"])
plt.xlabel("Importância")
plt.ylabel("Feature")
plt.title("Importância das Features - Random Forest")
plt.show()
```



O gráfico a seguir exibido é uma matriz de confusão, representada como um gráfico de calor. Ele mostra o desempenho do modelo de Random Forest na classificação dos dados de teste. Aqui está a análise detalhada:

Interpretação dos Valores

A matriz de confusão é composta por quatro quadrantes que representam os acertos e erros do modelo:

Verdadeiros Positivos (TP):

Valor: 129

O modelo corretamente classificou 129 instâncias da classe 1 (Risco Baixo) como Risco Baixo.

Falsos Negativos (FN):

Valor: 12

O modelo classificou erroneamente 12 instâncias da classe 1 (Risco Baixo) como Risco Alto.

Falsos Positivos (FP):

Valor: 34

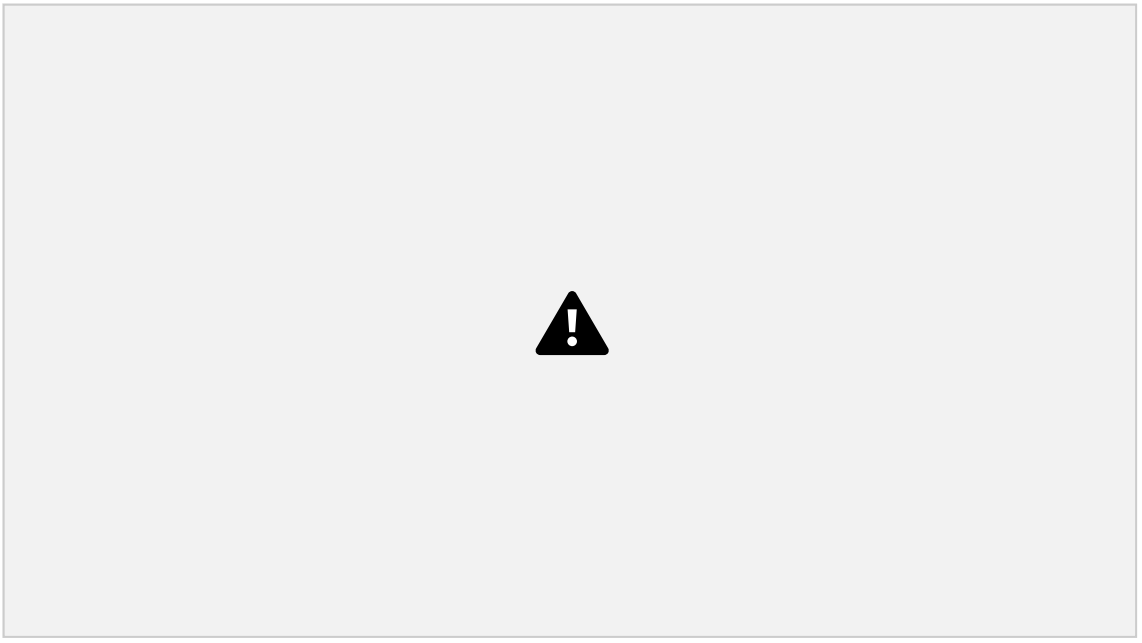
O modelo classificou erroneamente 34 instâncias da classe 0 (Risco Alto) como Risco Baixo.

Verdadeiros Negativos (TN):

Valor: 25

O modelo corretamente classificou 25 instâncias da classe 0 (Risco Alto) como Risco Alto.

O gráfico a seguir é um gráfico de Curva ROC (Receiver Operating Characteristic) para o modelo de Random Forest, e ele apresenta o desempenho do modelo em termos de separação entre as classes (Risco Baixo e Risco Alto). Além disso, inclui a métrica AUC (Área Sob a Curva), que quantifica o desempenho global do modelo.Análise da Curva ROC



Interpretação da Curva:

O eixo X representa a Taxa de Falsos Positivos (FPR), ou seja, a proporção de casos da classe negativa (Risco Alto) que foram incorretamente classificados como positivos (Risco Baixo).

O eixo Y representa a Taxa de Verdadeiros Positivos (TPR), ou seja, a proporção de casos da classe positiva (Risco Baixo) corretamente identificados.

A linha azul mostra como o modelo se comporta para diferentes limiares de classificação. Quanto mais próxima essa curva estiver do canto superior esquerdo, melhor o desempenho do modelo.

AUC (Área Sob a Curva):

O valor de AUC = 0.78 indica que o modelo tem um bom desempenho

geral. A AUC varia de 0 a 1:

0.5: Indica um modelo aleatório (sem poder preditivo).

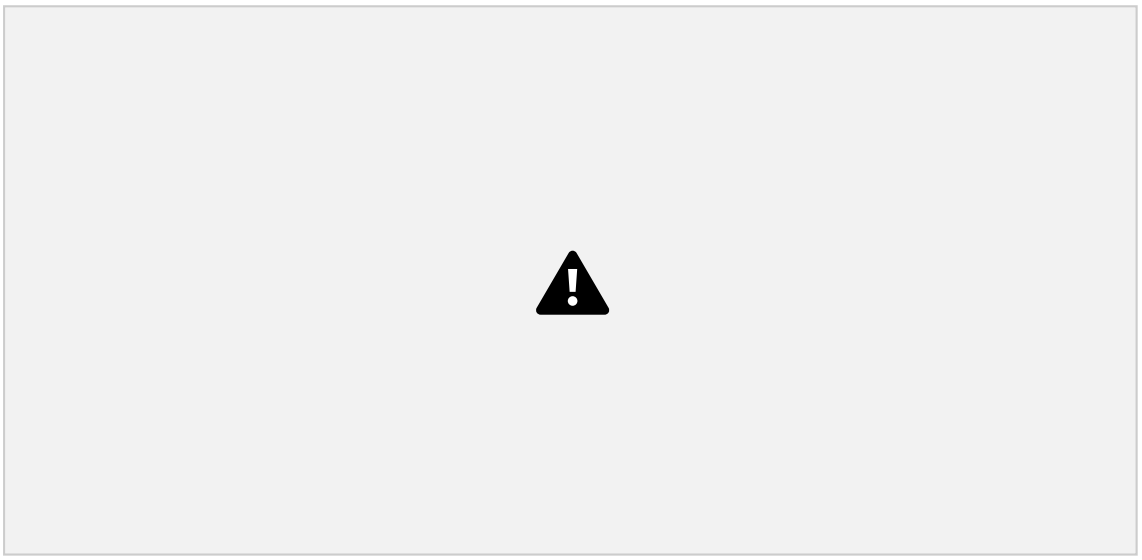
1.0: Indica um modelo perfeito.

0.78: Sugere que o modelo é capaz de diferenciar bem entre Risco Alto e Risco Baixo.

Linha de Random Guessing:

A linha tracejada (diagonal) representa um modelo que classifica aleatoriamente, ou seja, sem qualquer habilidade preditiva. O modelo de Random Forest claramente supera essa linha, indicando que ele tem valor preditivo.

O gráfico a seguir é um gráfico de barras mostra a distribuição das classes na base de dados para a variável-alvo Risk (Risco). Ele revela a quantidade de instâncias (amostras) para cada classe, que representam os grupos Risco Baixo (1) e Risco Alto (0).Análise da Distribuição



Desbalanceamento das Classes:

O gráfico deixa claro que há um desequilíbrio significativo entre as classes:

A classe 1 (Risco Baixo) possui mais de 700 instâncias.

A classe 0 (Risco Alto) possui menos de 300 instâncias.

Este desbalanceamento é comum em problemas de crédito, já que, geralmente, a maioria dos clientes apresenta um perfil de menor risco.

Impacto do Desequilíbrio nos Modelos:

O modelo de aprendizado de máquina tende a favorecer a classe majoritária (1), pois possui mais exemplos para aprender durante o treinamento.

Isso pode levar a:

Baixo recall para a classe minoritária (0), ou seja, muitos casos de Risco Alto não são identificados corretamente.

Alta precisão para a classe majoritária (1), mas com erros de classificação na classe minoritária.

Considerações Estatísticas:

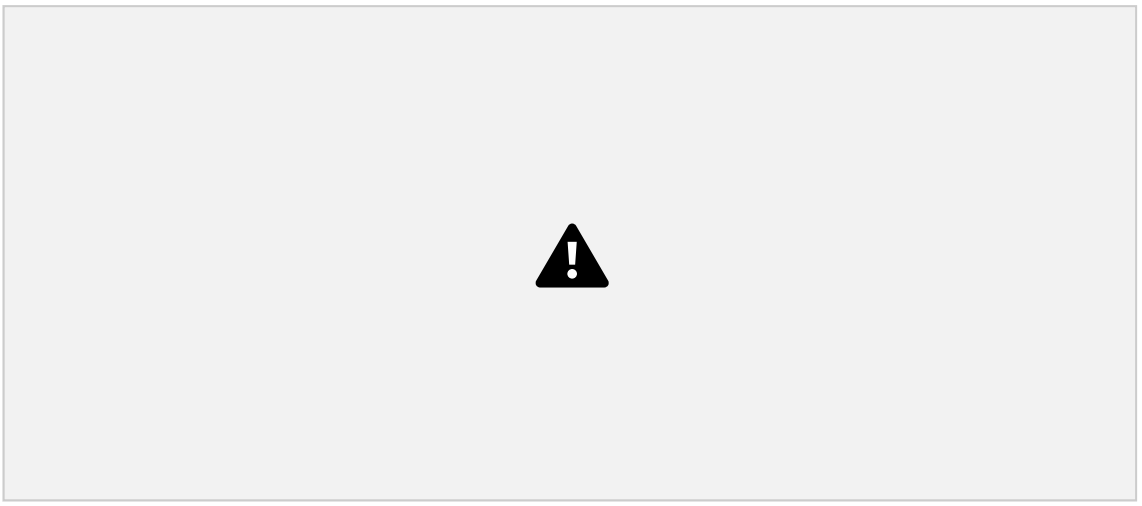
A classe 1 representa aproximadamente 70% dos dados, enquanto a classe 0 representa apenas 30%.

Um bom modelo para este cenário deve equilibrar a capacidade de prever ambas as classes de forma justa, mesmo com menos exemplos da classe 0.

O gráfico a seguir exibido é um boxplot, que mostra a distribuição do valor do crédito solicitado (Credit amount) para cada classe de risco (Risk). Ele é útil para identificar diferenças entre as classes e possíveis outliers nos dados.Interpretação

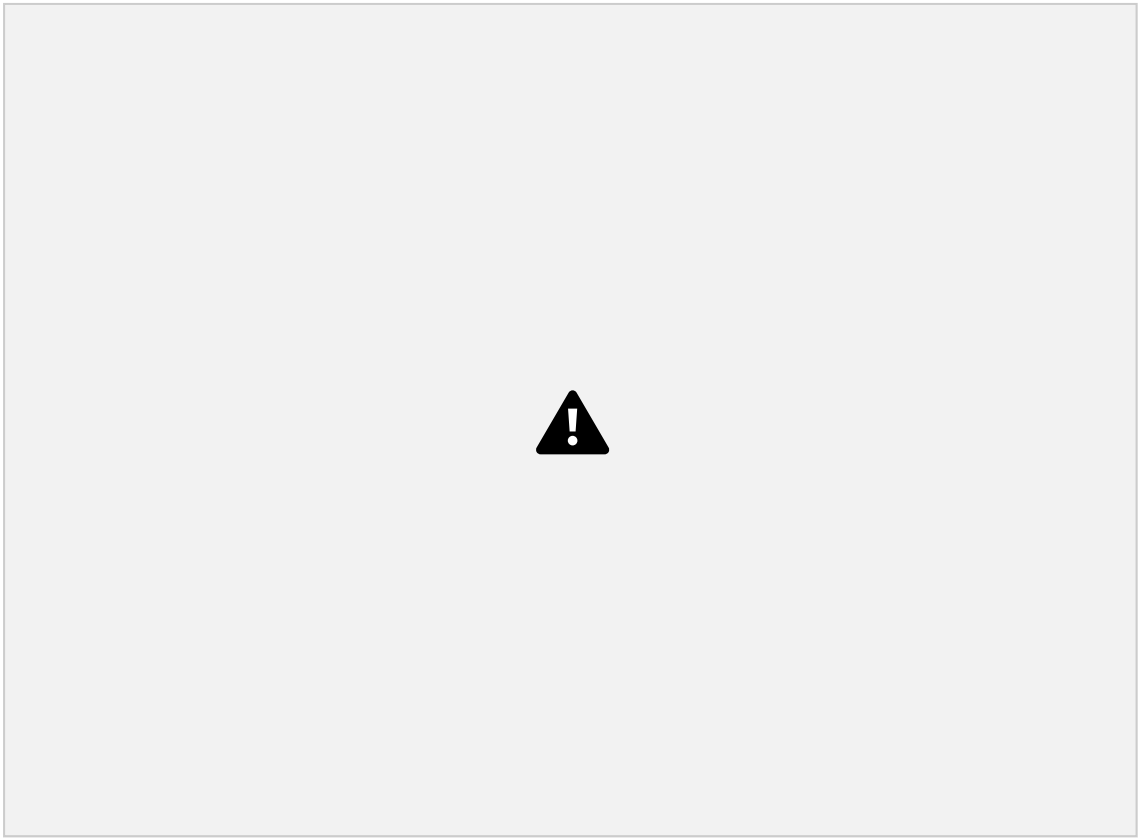
Este gráfico sugere que o valor do crédito solicitado é um indicador relevante para a classificação de risco. Modelos de aprendizado de máquina, como o

Random Forest, podem estar capturando esse padrão ao usar o Credit amount como uma das variáveis mais importantes (como observado no gráfico de importância das features).



O gráfico a seguir é um pairplot, que mostra a relação entre as variáveis numéricas Age (Idade), Credit amount (Valor do Crédito) e Duration (Duração do Crédito) com a variável-alvo Risk (Risco). Ele combina histogramas de

distribuição e gráficos de dispersão para destacar padrões e correlações.



Análise do Pairplot

Distribuição das Variáveis Numéricas:

Age (Idade):

Clientes de Risco Baixo (1) tendem a se concentrar em idades um pouco mais avançadas.

Clientes de Risco Alto (0) possuem uma distribuição ligeiramente mais uniforme, mas também apresentam maior concentração em idades mais jovens.

Credit amount (Valor do Crédito):

Clientes de Risco Alto (0) frequentemente solicitam valores maiores de crédito, enquanto clientes de Risco Baixo (1) solicitam valores menores.

Há sobreposição significativa, mas com uma leve tendência de risco alto associado a valores elevados.

Duration (Duração):

Clientes de Risco Alto (0) geralmente apresentam períodos de crédito mais longos em comparação a clientes de Risco Baixo (1).

Créditos de curta duração (por exemplo, abaixo de 12 meses) são predominantemente associados a Risco Baixo.

Relação entre Variáveis:

Age vs. Credit amount:

Clientes mais jovens com altos valores de crédito parecem estar mais associados ao Risco Alto (0).

Credit amount vs. Duration:

Créditos de maior duração e maior valor tendem a estar associados ao Risco Alto (0).

Age vs. Duration:

Há uma leve concentração de clientes mais jovens com créditos de longa duração na classe de Risco Alto (0).

Padrões Identificados:

Risco Alto (0): Geralmente associado a:

Idades mais jovens.

Valores de crédito maiores.

Durações de crédito mais longas.

Risco Baixo (1): Geralmente associado a:

Idades mais avançadas.

Valores de crédito menores.

Durações de crédito mais curtas.

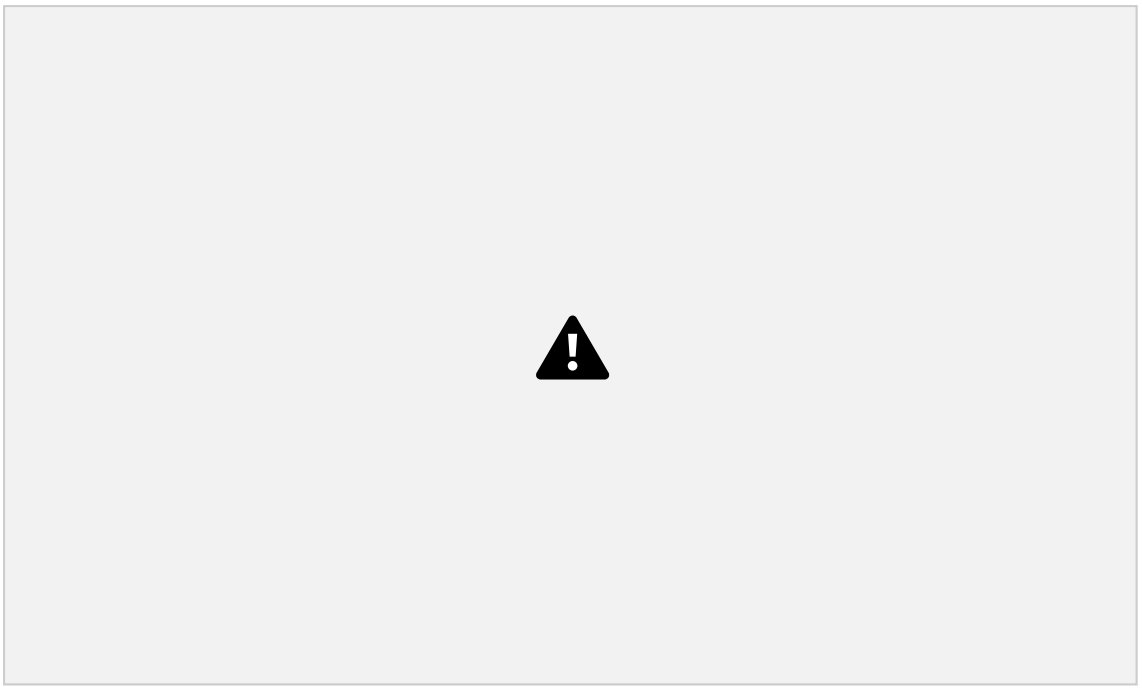
Insights

Este pairplot ajuda a destacar:

A importância das variáveis Credit amount e Duration como indicadores para a classificação de risco, algo que já havia sido sugerido nos gráficos anteriores.

Uma leve correlação entre idade, valor do crédito e duração com o risco, indicando que essas variáveis podem ser usadas para construir um modelo preditivo eficaz.

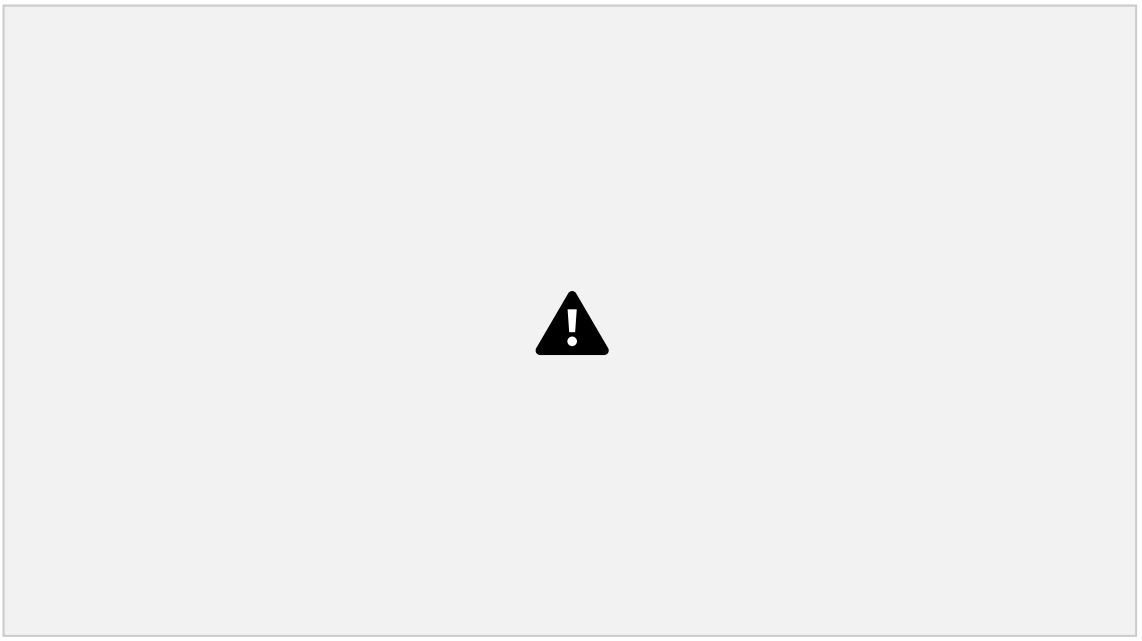
O gráfico mostrado é um KDE Plot (Kernel Density Estimation), que apresenta a distribuição dos valores de crédito (Credit amount) para as duas classes de risco:



Risco Baixo (1) e Risco Alto (0). Ele permite uma visualização suave das

densidades das variáveis, mostrando onde as observações estão mais concentradas em cada classe.Créditos maiores (acima de 10.000) estão mais frequentemente associados ao Risco Alto, sugerindo que esses valores poderiam ser monitorados com maior atenção em sistemas de crédito.

O gráfico mostrado é um heatmap (gráfico de calor) que exibe a concentração de risco por tipo de moradia (Housing). Ele é útil para visualizar a relação entre o tipo de moradia e as classes de risco, destacando onde as instâncias estão mais concentradas.



Análise do Gráfico

Eixos do Gráfico:

Eixo Y (Tipo de Moradia):

0: Moradia gratuita.

1: Casa própria.

2: Alugada.

Eixo X (Risco):

0: Risco Alto.

1: Risco Baixo.

Distribuição por Tipo de Moradia:

Moradia Gratuita (0):

Concentração de 64 clientes de Risco Baixo (1) e 44 clientes de Risco Alto (0). Relativamente equilibrado, mas com leve predominância de Risco Baixo.

Casa Própria (1):

527 clientes com Risco Baixo (1), dominando a categoria.

Apenas 186 clientes com Risco Alto (0), o que sugere que possuir casa própria está fortemente associado a Risco Baixo.

Alugada (2):

109 clientes de Risco Baixo (1) e 70 clientes de Risco Alto (0).

Proporção mais equilibrada, mas ainda com maior concentração na classe de Risco Baixo.

Padrões Identificados:

Casa Própria está claramente associada a um perfil de menor risco (1).

Clientes com moradia gratuita ou alugada apresentam maior proporção relativa de Risco Alto, indicando maior instabilidade financeira.

Insights

Possuir casa própria é um forte indicador de estabilidade financeira, reduzindo a probabilidade de ser classificado como Risco Alto.

Clientes com moradia gratuita ou alugada devem ser analisados com mais cuidado, especialmente se outras variáveis também indicarem instabilidade (como crédito alto ou duração longa).

Recomendações

Políticas de Crédito:

Clientes com casa própria podem ter condições de crédito mais favoráveis, enquanto aqueles que alugam ou têm moradia gratuita devem passar por uma análise mais detalhada.

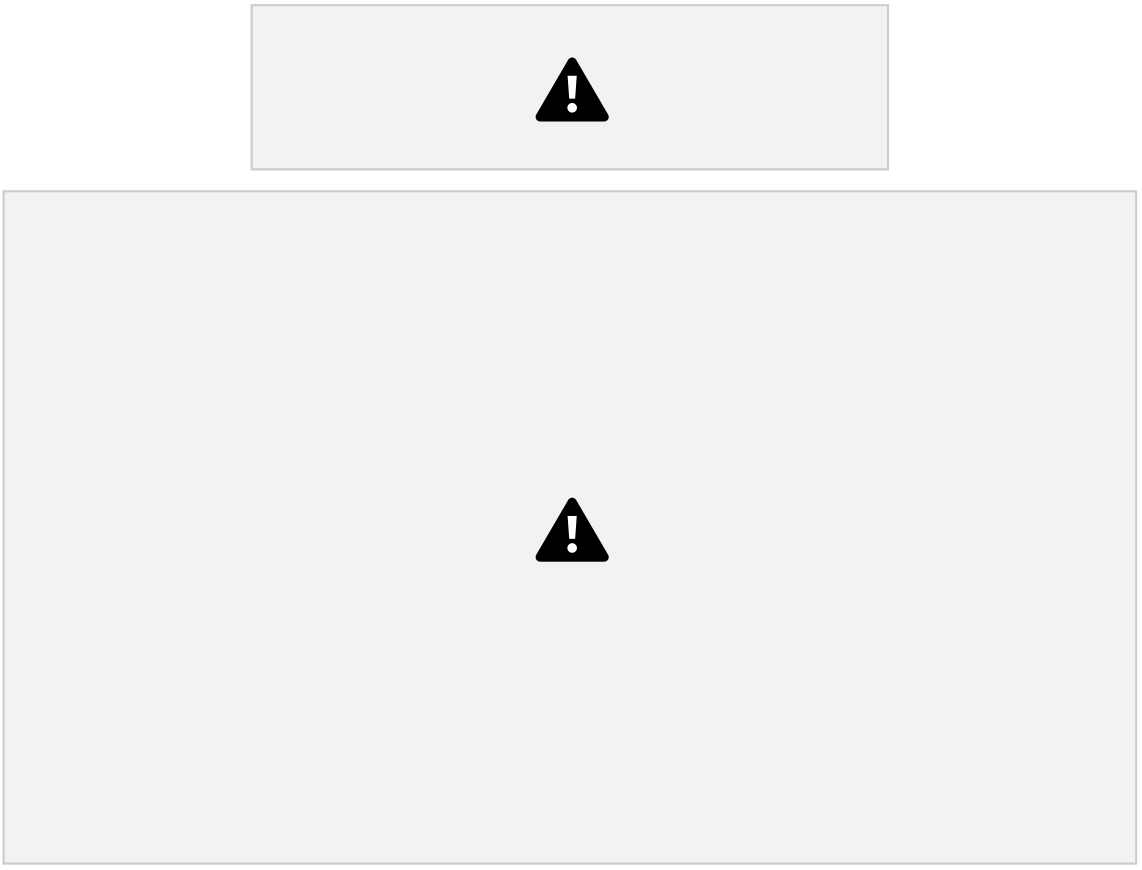
Integração com Outras Variáveis:

Usar o tipo de moradia em combinação com outras variáveis, como idade, valor do crédito e duração, para criar um perfil de risco mais completo.

Refinamento do Modelo:

Incluir o tipo de moradia como uma feature relevante no modelo preditivo, considerando seu impacto significativo na classificação de risco.

Este gráfico é um gráfico de barras empilhadas, que apresenta a proporção de risco (Risco Alto e Risco Baixo) em relação às finalidades do crédito (Purpose). Ele é útil para identificar padrões de risco associados a diferentes motivos pelos quais os clientes solicitam créditos.



Análise do Gráfico

Eixo X (Finalidade do Crédito):

Os valores de Purpose são categóricos e indicam diferentes finalidades,

como: 0: Compra de carro.

1: Reforma ou construção.

2: Educação.

3: Bens de consumo.

4: Viagens.

5: Saúde.

6: Investimento.

7: Outros.

Eixo Y (Proporção):

Representa a proporção relativa de cada classe de risco (Risco Alto e Risco Baixo) para cada finalidade.

Distribuição por Finalidade:

Finalidades com maior proporção de Risco Baixo (1):

A maioria das finalidades, como 0 (Compra de carro) e 5 (Saúde), apresentam uma predominância de clientes com Risco Baixo.

Finalidades como 4 (Viagens) e 7 (Outros) também possuem uma proporção elevada de clientes com menor risco.

Finalidades com maior proporção de Risco Alto (0):

Finalidades como 2 (Educação) e 3 (Bens de consumo) apresentam proporções relativamente maiores de clientes com Risco Alto.

Embora ainda predominem clientes de Risco Baixo, essas finalidades devem ser analisadas com maior atenção.

Padrões Identificados:

Finalidades que representam investimentos em ativos de longo prazo (e.g., 0 - Compra de carro) têm maior predominância de Risco Baixo.

Finalidades voltadas para consumo imediato ou menos tangíveis (e.g., 2 - Educação, 3 - Bens de consumo) possuem maior proporção de Risco Alto.

Insights

O tipo de finalidade do crédito está claramente relacionado ao perfil de risco. Finalidades como saúde e compra de bens duráveis indicam maior segurança financeira, enquanto bens de consumo e educação podem indicar instabilidade ou necessidade emergencial.

Este padrão pode ser explorado para criar políticas específicas de concessão de crédito, com critérios ajustados para cada tipo de finalidade.

O objetivo desta atividade foi realizar uma análise exploratória e preditiva dos dados relacionados ao risco de crédito, utilizando técnicas gráficas e algoritmos de aprendizado de máquina. Com base nos gráficos gerados e nas métricas extraídas, foi possível compreender os padrões dos dados, as variáveis mais influentes e o desempenho dos modelos utilizados para prever a classificação de risco.

Contexto da Base de Dados

A base de dados possui informações sobre clientes e seus pedidos de crédito, classificados em duas categorias:

1.Risco Alto (0): Clientes com maior probabilidade de inadimplência.

2.Risco Baixo (1): Clientes com menor probabilidade de inadimplência.

As variáveis incluem características demográficas (idade, sexo), econômicas (valor do crédito, duração do crédito, tipo de moradia) e financeiras (contas correntes e de poupança). A base apresentou um desbalanceamento significativo, com a classe de Risco Baixo sendo majoritária (cerca de 70% dos dados).

Análise Gráfica e Padrões Identificados

1. Distribuição de Risco

O gráfico de barras revelou um claro desbalanceamento entre as classes, com

a maioria dos clientes classificados como Risco Baixo (1). Esse desbalanceamento impacta diretamente os modelos de aprendizado de máquina, que tendem a favorecer a classe majoritária, resultando em baixa precisão para a classe de Risco Alto (0).

Ação Recomendada: Tratar o desbalanceamento por meio de técnicas como oversampling da classe minoritária (e.g., SMOTE) ou ajuste de pesos no modelo para melhorar a identificação de clientes de alto risco.

2. Importância das Features (Random Forest)

O gráfico de importância das features destacou as variáveis mais influentes para a classificação de risco:

Credit amount(Valor do Crédito): A variável mais importante, indicando que clientes que solicitam valores maiores possuem maior probabilidade de serem classificados como Risco Alto.

Age (Idade): A segunda variável mais importante, com clientes mais jovens apresentando maior risco.

Duration (Duração do Crédito): Créditos de maior duração também tendem a estar associados ao Risco Alto.

Variáveis como Saving accounts e Checking account tiveram relevância moderada, indicando que a condição financeira dos clientes influencia na classificação.

Ação Recomendada: Priorizar essas variáveis na modelagem e criar novas features derivadas, como a relação entre o valor do crédito e a duração do pagamento.

3. Distribuição de Valores de Crédito (Boxplot)

O boxplot mostrou que:

Clientes de Risco Alto (0)frequentemente solicitam valores de crédito mais altos, com uma maior presença de outliers (créditos acima de 15.000).

- Clientes de Risco Baixo (1) geralmente solicitam valores mais baixos, com menor variabilidade.

Ação Recomendada: Implementar políticas de crédito mais restritivas para valores muito altos, especialmente para clientes com características associadas ao Risco Alto.

4. Relações entre Variáveis Numéricas (Pairplot)

O pairplot revelou padrões interessantes:

Idade e Risco: Clientes mais jovens estão mais associados ao Risco Alto.

Duração e Valor do Crédito: Créditos de maior valor e duração tendem a estar concentrados na classe de Risco Alto.

A sobreposição entre as classes sugere que, embora essas variáveis sejam úteis, não são suficientes por si só para separar completamente as classes, exigindo combinações mais complexas no modelo preditivo.

Ação Recomendada: Criar interações entre variáveis, como a proporção entre o valor do crédito e a duração do contrato, para aumentar a capacidade preditiva do modelo.

5. Curva ROC e AUC

O gráfico da Curva ROC demonstrou que o modelo de Random Forest tem um AUC de 0.78, indicando bom desempenho geral na separação entre as classes. Contudo, há espaço para melhorias, especialmente na identificação de clientes de Risco Alto (classe 0).

Ação Recomendada: Ajustar o limiar de decisão do modelo para priorizar o recall da classe 0, considerando que erros na identificação de clientes de alto risco podem ter consequências financeiras significativas

Modelagem e Avaliação

Os modelos treinados (Árvore de Decisão e Random Forest) apresentaram os seguintes resultados:

Árvore de Decisão:

Acurácia: 66%

Limitada na identificação de clientes de alto risco, com recall de apenas 49% para a classe 0.

Random Forest:

Acurácia: 77%

Melhor desempenho geral, com recall de 91% para a classe 1 (Risco Baixo), mas recall de apenas 42% para a classe 0 (Risco Alto).

Esses resultados refletem a influência do desbalanceamento da base e a dificuldade de prever corretamente a classe minoritária.

Conclusões e Recomendações

A análise evidenciou padrões claros na base de dados e sugeriu melhorias para a modelagem de risco de crédito:

1.Foco nas Variáveis Mais Relevantes:

Priorizar o uso de variáveis como Credit amount, Age e Duration, que apresentaram forte influência na classificação de risco.

2. Tratamento do Desbalanceamento:

- Aplicar técnicas de oversampling, undersampling ou ajuste de pesos no modelo para melhorar a performance na classe de Risco Alto.

3. Engenharia de Features:

Criar variáveis derivadas, como a proporção entre valor do crédito e duração, ou a interação entre idade e valor do crédito.

4. Avaliação de Erros Críticos:

Ajustar o limiar de decisão para reduzir os falsos negativos (clientes de alto

risco classificados como baixo risco), mitigando possíveis impactos financeiros.

5. Exploração de Modelos Avançados:

Experimentar algoritmos como Gradient Boosting (XGBoost) ou Redes Neurais para capturar padrões mais complexos.

Reflexão Final

A análise exploratória e preditiva mostrou que o problema de risco de crédito é desafiador devido à sobreposição entre as classes e ao desbalanceamento da base de dados. No entanto, o uso de gráficos, como pairplots, boxplots e curvas ROC, foi fundamental para identificar padrões e direcionar as decisões de modelagem. Com ajustes adequados, é possível melhorar significativamente a identificação de clientes de alto risco, trazendo maior segurança financeira à operação de crédito.

Referências Bibliográficas

ntrodução ao Aprendizado de Máquina:

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.

Algoritmos e Modelo

Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

James, G., Witten, D., Hastie, T., Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*. Springer, 2013.

Banco de dados extraído:

https://github.com/alicenkbaytop/German-Credit-Risk Classification/blob/master/german_credit_risk.csv