**Quantum Finance - Projeto Integrado**

**Apresentação de aplicação em Inteligência Artificial/Machine Learning (Turma 10DTS)**

Grupo:

* RM: **357043 Laura Munhoz Friozi**
* RM: **358133 Cristian Eugenio Maximiliano Insfrán**
* RM: **358275 Júlio Cesar Kenji Monzem**
* RM: **358421 Júlio Cesar Falco**

**Apresentação do desafio**

- O que deve ser apresentado?

Quantum Finance – A partir da base de classificação de credit score, a qual vocês têm acesso aos dados, alguns códigos e resultados.

**Trabalho:**

Neste trabalho, como parte do time de analistas da Quantum Finance, vocês deverão explorar uma base de dados originalmente utilizada para classificação de score de crédito, disponível no Kaggle (https://www.kaggle.com/datasets/parisrohan/credit-score-classification), utilizando técnicas de Análise Exploratória de Dados (EDA) e algoritmos de Machine Learning supervisionados ou não supervisionados. ​

O objetivo é aplicar e interpretar os resultados obtidos, assim como criar um sistema que gere valor a partir da análise da base de dados. ​

Exemplo 1: Modelo de Classificação Supervisionada​

Desenvolver um modelo de classificação supervisionada para prever a classificação de crédito dos indivíduos presentes na base. ​

Passos esperados:​

* Realizar uma análise exploratória dos dados (EDA) para entender as características principais da base e as relações entre variáveis; ​
* Implementar o modelo de classificação (ex.: Decision Tree, Random Forest, Logistic Regression, ou outro);​
* Avaliar os resultados utilizando métricas como acurácia, F1-score, matriz de confusão, entre outras relevantes; ​
* Apresentar os resultados e explicar como o modelo pode ser utilizado para decisões financeiras. ​

Parte 2: Algoritmo Não Supervisionado OU Sistema Especialista​

​Escolher entre: Implementar um algoritmo não supervisionado (ex.: K-Means ou PCA) de forma a agrupar os dados e analisar padrões presentes na base. ​

Criar um sistema especialista baseado em regras que interprete e classifique os dados. ​

​Passos esperados:​

* Realizar uma EDA inicial para definir as estratégias da aplicação do algoritmo não supervisionado ou para definir as regras do sistema especialista; ​
* Aplicar a técnica escolhida (algoritmo ou sistema especialista) na base; ​
* Documentar a experiência, destacando os resultados obtidos, os desafios enfrentados, e a utilidade prática da abordagem escolhida.
* Trabalho precisa ser entregue impreterivelmente dia 27/04/2025, até as 23:59.
* Na entrega no portal subir código e outros materiais em um arquivo zip.
* Como falado em sala: tentem se dividir nas equipes do startup one e começar por um bom processo de EDA, também se possível para aqueles que forem modelar o problema como classificação, favor prover resultados de testes com diferentes classificadores

**Classificação de Score de Crédito**

**1. Introdução**

Neste trabalho, como parte do time de analistas da Quantum Finance, foi realizada uma análise de dados e o desenvolvimento de modelos de **Machine Learning supervisionados** para **prever a classificação de crédito** dos indivíduos presentes na base.

Antes da aplicação dos modelos, foi realizada uma etapa de **Análise Exploratória de Dados (EDA)**, onde exploramos a distribuição das variáveis, investigamos valores ausentes, outliers e relações entre as variáveis principais.

Em seguida, foi realizado o **tratamento dos dados**, incluindo:

* Preenchimento de valores ausentes
* Limpeza de inconsistências
* Padronização de variáveis

**Para as variáveis categóricas**, optamos por utilizar a técnica de **One-Hot Encoding** (Dummies), garantindo que a representação das categorias fosse adequada aos algoritmos de classificação supervisionada. Essa abordagem evita a criação de ordens artificiais entre categorias e melhora a performance dos modelos.

O objetivo principal foi aplicar técnicas de classificação, interpretar os resultados obtidos e criar um sistema que gere valor a partir da análise da base, auxiliando decisões financeiras.

**Distribuição do Credit Utilization Ratio**

**A graph of credit utility

AI-generated content may be incorrect.**

Observa-se que a maioria dos clientes apresenta um Credit Utilization Ratio entre 25% e 40%, indicando uma utilização moderada do crédito disponível.

**Distribuição do Credit Score**

A distribuição dos clientes por score é relativamente desbalanceada:

* **Standard**: 53.1% dos clientes
* **Poor**: 29.0% dos clientes
* **Good**: 17.8% dos clientes

Este desbalanceamento foi levado em consideração na escolha das métricas de avaliação dos modelos.

A graph of a credit score

AI-generated content may be incorrect.

**2. Modelos Aplicados e Avaliação de Resultados**

**2.1 Árvore de Decisão**

| **Classe** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| False | 0.73 | 0.73 | 0.73 | 4344 |
| True | 0.76 | 0.75 | 0.75 | 4753 |

* **Acurácia geral:** 74%
* **Macro F1-Score:** 74%

**Interpretação:**  
O modelo de Árvore de Decisão apresentou desempenho equilibrado entre as duas classes. A acurácia de 74% indica que o modelo consegue classificar corretamente a maioria dos indivíduos. Tanto a precisão quanto o recall são consistentes para as classes "True" e "False", mostrando que o modelo é confiável para previsões de crédito básicas.

**2.2 Random Forest**

| **Classe** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| False | 0.77 | 0.84 | 0.80 | 4344 |
| True | 0.84 | 0.77 | 0.80 | 4753 |

* **Acurácia geral:** 80%
* **Macro F1-Score:** 80%

**Interpretação:**  
O modelo de Random Forest superou a Árvore de Decisão, alcançando 80% de acurácia. O balanceamento entre as métricas de precisão e recall mostra que o Random Forest consegue identificar tanto bons quanto maus pagadores com alta eficiência.  
Este modelo é o mais indicado para ser utilizado em produção, pois além de maior acurácia, oferece robustez contra overfitting e boa capacidade de generalização.

**2.3 Regressão Logística**

| **Classe** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| False | 0.00 | 0.00 | 0.00 | 4344 |
| True | 0.52 | 1.00 | 0.69 | 4753 |

* **Acurácia geral:** 52%
* **Macro F1-Score:** 34%

**Interpretação:**  
A Regressão Logística apresentou desempenho limitado. O modelo classificou praticamente todos os registros como "True", alcançando recall 100% para essa classe, mas falhando completamente para a classe "False".  
Isso demonstra que a Regressão Logística simples não é adequada para este conjunto de dados, possivelmente devido à não linearidade ou desbalanceamento das classes.

**2.4 K-Means (Modelo Não Supervisionado)**

| **Classe** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| False | 0.00 | 0.00 | 0.00 | 4344 |
| True | 0.52 | 1.00 | 0.69 | 4753 |

* **Acurácia geral:** 52%
* **Macro F1-Score:** 34%

**Interpretação:**  
O modelo K-Means, aplicado de forma não supervisionada, obteve desempenho semelhante ao da Regressão Logística no contexto atual, classificando praticamente todos os exemplos como "True" (bons pagadores).  
Embora tenha alcançado recall 100% para a classe positiva, não conseguiu identificar a classe negativa ("False").  
Isso mostra que o K-Means, por ser um algoritmo de agrupamento baseado apenas em características similares, não é a melhor abordagem para substituição direta de classificadores supervisionados neste problema.  
Contudo, ele ainda pode ser útil para descobrir padrões de comportamento ou segmentações naturais no perfil de clientes.

**3. Conclusão**

* **Random Forest** foi o modelo que apresentou melhor desempenho e deve ser priorizado para aplicações práticas de classificação de crédito.
* **Árvore de Decisão** oferece uma alternativa interpretável, ainda com boa acurácia.
* **Regressão Logística** e **K-Means** não atingiram desempenho satisfatório para fins de classificação supervisionada neste problema.