**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.

**Análise de Classificação de Score de Crédito**

**1. Introdução**

Neste trabalho, como parte do time de analistas da Quantum Finance, exploramos uma base de dados originalmente utilizada para classificação de score de crédito.

Inicialmente, foi realizada uma **Análise Exploratória de Dados (EDA)** para entender a distribuição das variáveis e identificar possíveis inconsistências.  
Como parte dessa etapa:

* **Colunas consideradas não relevantes** para a análise foram **removidas**:
  + ID, Customer\_ID, Month, Name, SSN, Monthly\_Inhand\_Salary, Type\_of\_Loan, Num\_of\_Delayed\_Payment e Amount\_invested\_monthly.
* **Tratamento dos dados**:
  + Remoção de outliers (como valores de idade superiores a 80 anos).
  + Padronização e preenchimento de valores ausentes.
  + Correções de inconsistências em variáveis categóricas e numéricas.

Além disso, foi gerada uma **matriz de correlação** entre as variáveis numéricas para identificar relações importantes:

* Verificou-se uma **correlação moderada positiva** entre Outstanding\_Debt (dívida pendente) e Delay\_from\_due\_date (atrasos de pagamento), sugerindo que atrasos estão relacionados ao aumento da dívida.
* Observou-se também uma **correlação negativa significativa** entre Credit\_Mix e Outstanding\_Debt, indicando que clientes com um mix de crédito mais diversificado tendem a apresentar dívidas menores.
* A variável Credit\_Score, principal alvo de interesse, mostrou **correlações fracas** com a maioria das demais variáveis, sugerindo que múltiplos fatores combinados são necessários para predizer o score de crédito.

A distribuição das principais variáveis foi visualizada nos gráficos a seguir:

**Gráfico**

|  |  |
| --- | --- |
|  | Distribuição das classes de **Credit Score** (0, 1 e 2). |

**A graph of a credit score

AI-generated content may be incorrect.**

Distribuição do **Credit Utilization Ratio**, indicando um comportamento próximo ao normal.

**A graph of credit utility

AI-generated content may be incorrect.**

Após a preparação da base, aplicamos técnicas de Machine Learning supervisionadas e não supervisionadas com o objetivo de construir modelos capazes de apoiar decisões financeiras estratégicas.

**Matriz de correlação**

**A screenshot of a graph

AI-generated content may be incorrect.**

**2. Avaliação dos Modelos Supervisionados e Não Supervisionados**

**2.1 Árvore de Decisão**

| **Classe** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| 0 | 0.67 | 0.66 | 0.66 | 1655 |
| 1 | 0.72 | 0.72 | 0.72 | 2689 |
| 2 | 0.74 | 0.74 | 0.74 | 4753 |

* **Acurácia geral:** 72%
* **Macro F1-Score:** 71%

**Interpretação:**  
O modelo de Árvore de Decisão demonstrou desempenho razoável, com resultados equilibrados entre as classes. No entanto, a performance poderia ser melhorada com ajuste de hiperparâmetros ou técnicas de ensemble.

**2.2 Random Forest**

| **Classe** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| 0 | 0.79 | 0.76 | 0.78 | 1655 |
| 1 | 0.79 | 0.82 | 0.81 | 2689 |
| 2 | 0.82 | 0.81 | 0.82 | 4753 |

* **Acurácia geral:** 81%
* **Macro F1-Score:** 80%

**Interpretação:**  
O Random Forest apresentou o melhor desempenho geral, com ótima capacidade de generalização entre as classes. Trata-se do modelo mais robusto para uma possível implementação prática.

**2.3 Regressão Logística**

| **Classe** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| 0 | 0.00 | 0.00 | 0.00 | 1655 |
| 1 | 0.00 | 0.00 | 0.00 | 2689 |
| 2 | 0.52 | 1.00 | 0.69 | 4753 |

* **Acurácia geral:** 52%
* **Macro F1-Score:** 23%

**Interpretação:**  
A Regressão Logística demonstrou baixo desempenho. O modelo classificou quase todos os exemplos como pertencentes à classe 2, o que explica o recall de 100% nesta classe, mas um F1-score muito baixo para as demais. Essa limitação é típica de problemas com dados desbalanceados ou não linearmente separáveis.

**2.4 K-Means (Não Supervisionado)**

| **Classe** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| 0 | 0.19 | 0.03 | 0.05 | 1655 |
| 1 | 0.00 | 0.00 | 0.00 | 2689 |
| 2 | 0.52 | 0.97 | 0.68 | 4753 |

* **Acurácia geral:** 51%
* **Macro F1-Score:** 24%

**Interpretação:**  
Como esperado em um algoritmo não supervisionado, o K-Means não conseguiu replicar perfeitamente as classes reais, mas conseguiu agrupar corretamente boa parte dos clientes da classe 2. O modelo revelou padrões ocultos, úteis para segmentações de mercado ou análises de comportamento.

**3. Conclusão**

* **Random Forest** se mostrou o modelo mais eficaz, atingindo 81% de acurácia e boa generalização entre as classes.
* **Árvore de Decisão** teve desempenho razoável, podendo ser usada com ajustes adicionais.
* **Regressão Logística** não foi adequada para este cenário, principalmente devido à sua limitação em bases complexas.
* **K-Means** revelou agrupamentos úteis, mas, como técnica não supervisionada, não substitui um modelo supervisionado para classificação formal.

**4. Aplicação Prática**

A partir dos resultados obtidos:

* Clientes classificados como **classe 2** podem ser considerados de **baixo risco** para concessão de crédito.
* Clientes classificados nas **classes 0 e 1** requerem análise mais cuidadosa ou condições de crédito diferenciadas.
* A segmentação via **K-Means** pode ser aproveitada para **campanhas específicas de marketing** ou **planos de crédito personalizados**