



**Professor: Renan Gomes Vieira**

**Equipe**

- Igor Santana Sampaio
- Jorge Roniel de Paula Souza
- Luis Eduardo Martins Barbosa
- Victor Lopes Mendes

## **Trabalho Final Ciência dos Dados**

## **Aprendizagem de Máquina**

### **Sobre os dados:**

Este conjunto de dados contém vários indicadores de saúde e fatores de risco relacionados a doenças cardíacas. Parâmetros como idade, gênero, pressão arterial, níveis de colesterol, hábitos de fumo e padrões de exercício foram coletados para analisar o risco de doenças cardíacas e contribuir para pesquisas em saúde. O conjunto de dados pode ser utilizado por profissionais de saúde, pesquisadores e analistas de dados para examinar tendências relacionadas a doenças cardíacas, identificar fatores de risco e realizar diversas análises relacionadas à saúde.

### **Desafios encontrados:**

No desenvolvimento deste estudo, foram enfrentados desafios significativos que impactaram a condução da pesquisa. A primeira dificuldade consistiu na seleção criteriosa dos modelos de aprendizado de máquina, uma vez que a literatura oferece uma vasta gama de algoritmos com comportamentos distintos frente a dados clínicos. Determinar quais classificadores seriam mais adequados para capturar a complexidade das patologias cardíacas exigiu uma análise profunda das características de cada método.

Adicionalmente, a implementação e estruturação do código representaram um obstáculo técnico relevante. A construção de um fluxo de trabalho robusto, que integrasse de forma eficiente as etapas de pré-processamento, a execução das rodadas de validação cruzada (K-Fold) e a aplicação rigorosa dos testes estatísticos, demandou um esforço considerável para garantir a reprodutibilidade dos experimentos e a integridade dos resultados obtidos.

### **Aprendizado Adquirido:**

O desenvolvimento deste estudo trouxe aprendizados importantes a partir das dificuldades enfrentadas. A necessidade de selecionar modelos de aprendizado de máquina diante da diversidade de algoritmos disponíveis evidenciou a importância de uma análise crítica e fundamentada para compreender suas limitações e potencialidades frente a dados clínicos complexos.

Além disso, a implementação de um fluxo de trabalho robusto mostrou que a organização e a integração cuidadosa das etapas de pré-processamento, validação cruzada e testes estatísticos são essenciais para assegurar a reprodutibilidade e a confiabilidade dos resultados, reforçando que rigor metodológico e competência técnica caminham juntos na pesquisa científica.

## **NLP:**

### **Sobre os dados:**

Este conjunto de dados contém diversos registros de interações e preferências de usuários em relação a obras cinematográficas. Parâmetros como títulos de filmes, gêneros, avaliações numéricas e etiquetas de conteúdo (tags) foram coletados para analisar padrões de consumo de mídia e contribuir para pesquisas em sistemas de personalização. O conjunto de dados pode ser utilizado por cientistas de dados, pesquisadores de inteligência artificial e analistas de sistemas para examinar tendências no setor do entretenimento, identificar perfis de preferência e realizar diversas análises relacionadas a algoritmos de recomendação.

### **Desafios encontrados:**

O dataset original continhas diversos atributos que não contribuíam para a nossa análise de NLP proposta, como IDs, nomes de diretores e URLs de imagens etc. Um dos desafios iniciais foi realizar a seleção de atributos que realmente eram relevantes, focando no conteúdo mais textual, como reviews e pegando também a nota dos críticos para usarmos mais tarde como indicador do que tem review boa e ruim. Em seguida, garantimos o tratamento de dados faltantes (NaN) e strings vazias, que poderiam comprometer a execução dos nossos algoritmos.

Outro desafio encontrado foi o desbalanceamento da classe "Fresh". Ao rodar os primeiros experimentos, a matriz de confusão revelou que o modelo apresentava um viés, classificando errado muitas críticas negativas como positivas. Vencemos esse desafio aplicando a técnica de random undersampling, igualando as amostras das classes "Fresh" e "Rotten", o que resultou em um modelo muito mais equilibrado e confiável

Por fim, enfrentamos dificuldades no ajuste de hiperparâmetros. Entender como cada impacto de parâmetro no espaço de busca do Grid Search foi desafiador. Foi necessário entender que a inclusão de bigramas era essencial para capturar o contexto de expressões de negação (ex: "not good"), apesar de aumentar a complexidade do modelo.

### **Aprendizados adquiridos:**

Com o desenvolvimento deste projeto, aprendemos que uma acurácia alta pode enganar se o modelo não for capaz de distinguir bem as classes individualmente. A análise da matriz de confusão, com as métricas de Precision, Recall e F1-Score, se tornou nossa principal ferramenta para validar a qualidade real do classificador e garantir que ele não estivesse apenas "chutando" a classe mais frequente.

O uso de WordClouds foi um aprendizado valioso para a explicabilidade do modelo. Inicialmente, as nuvens continham termos genéricos que não nos davam valor visual. Ao adicionarmos elas nas stopwords, conseguimos validar que o modelo estava de fato focando em termos relevantes e adjetivos que expressam opiniões, tornando a visualização muito mais clara e interpretável.