RELATÓRIO FINAL MINDPY

Nomes: Guilherme Rodrigues Cabreira, Guilherme Corvelo Bittencourt, Luiz Flávio Gonçalves de Paula, Rene Arthur Rocha de Souza e Wesley Fernandes Da Silveira

---

1. Escolha do Dataset e Definição do Problema

1.1. Escolha do Dataset:

Dado que a escolha do cenário era livre entre os grupos, decidimos desenvolver um projeto que gerasse valor para a comunidade, especialmente na área da saúde. Com esse objetivo em mente, iniciamos a busca por um dataset que nos permitisse abordar um tema relevante e cientificamente significativo.

Utilizamos a plataforma Kaggle, um site amplamente conhecido por oferecer uma vasta gama de datasets para análises e aprendizado de máquina.

Nossa busca foi orientada pela palavra-chave health, o que nos apresentou uma variedade de opções. Entre os muitos datasets disponíveis, identificamos o Alzheimer's Disease Dataset, que chamou nossa atenção por diversos motivos. Em primeiro lugar, o dataset atendia aos requisitos de tamanho, com informações sobre 2.149 pacientes, garantindo um volume suficiente para análises estatísticas e modelagens preditivas robustas. Além disso, a riqueza de variáveis presentes, como dados demográficos, históricos médicos, avaliações clínicas e diagnóstico de Alzheimer, nos forneceu um cenário ideal para explorar diferentes abordagens.

Outro fator determinante foi o impacto potencial de trabalhar com um tema tão relevante quanto o Alzheimer, uma condição que afeta milhões de pessoas em todo o mundo. Escolher esse dataset não apenas proporcionou um desafio técnico interessante, mas também alinhou nosso projeto a uma causa de grande importância social. Assim, a escolha foi feita com base na qualidade, na abrangência das informações e no impacto que os resultados poderiam trazer para a comunidade e a área da saúde.

1.2.Definição do Problema:

Com o objetivo de gerar valor para a comunidade e abordar um problema real na área da saúde, decidimos desenvolver um modelo de classificação para a doença de Alzheimer utilizando técnicas de Inteligência Artificial.

O diagnóstico precoce do Alzheimer é um desafio importante, pois pode impactar diretamente a qualidade de vida dos pacientes e permitir um melhor planejamento dos tratamentos. No entanto, os métodos tradicionais de diagnóstico muitas vezes são subjetivos ou dependem de exames invasivos, o que pode atrasar o início de intervenções necessárias. Por isso, acreditamos que a análise de dados pode trazer contribuições valiosas para essa área.

A ideia principal do projeto é criar um classificador que consiga prever a presença da doença com base em variáveis clínicas, demográficas e de histórico médico presentes no dataset escolhido. Além de obter bons resultados preditivos, também pretendemos explorar como esses fatores se relacionam com o diagnóstico, ajudando a trazer novos insights clínicos.

Com essa iniciativa, buscamos unir tecnologia e saúde, propondo uma solução prática e que possa ser útil tanto no ambiente clínico quanto como base para futuras pesquisas.

2. Análise exploratória dos Dados

Para iniciar o trabalho com o dataset escolhido, realizamos uma análise exploratória para compreender melhor as características dos dados, identificar possíveis valores faltantes ou inconsistências e extrair informações relevantes para orientar as próximas etapas do projeto. Esta etapa incluiu a verificação geral da composição do dataset, a análise de distribuições, identificação de valores discrepantes (outliers) e a criação de gráficos que ajudaram a guiar as decisões relacionadas à classificação.

2.1. Estrutura e Composição do Dataset

Inicialmente, utilizamos a função .info() para verificar a quantidade de entradas, os tipos de variáveis e a existência de valores nulos no dataset. A partir desse levantamento, confirmamos que o dataset continha 2.149 entradas únicas e não apresentava valores nulos ou duplicados.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2149 entries, 0 to 2148
Data columns (total 35 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   PatientID                  2149 non-null   int64
 1   Age                        2149 non-null   int64
 2   Gender                     2149 non-null   int64
 3   Ethnicity                  2149 non-null   int64
 4   EducationLevel             2149 non-null   int64
 5   BMI                        2149 non-null   float64
 6   Smoking                    2149 non-null   int64
 7   AlcoholConsumption         2149 non-null   float64
 8   PhysicalActivity           2149 non-null   float64
 9   DietQuality                2149 non-null   float64
 10  SleepQuality               2149 non-null   float64
 11  FamilyHistoryAlzheimers    2149 non-null   int64
 12  CardiovascularDisease      2149 non-null   int64
 13  Diabetes                   2149 non-null   int64
 14  Depression                 2149 non-null   int64
 15  HeadInjury                 2149 non-null   int64
 16  Hypertension               2149 non-null   int64
 17  SystolicBP                 2149 non-null   int64
 18  DiastolicBP                2149 non-null   int64
 19  CholesterolTotal           2149 non-null   float64
...
 33  Diagnosis                  2149 non-null   int64
 34  DoctorInCharge             2149 non-null   object
dtypes: float64(12), int64(22), object(1)
memory usage: 587.7+ KB
```

Figura 1. Análise geral dos dados do dataset.

Em seguida, utilizamos a função .describe() para calcular métricas estatísticas como média, desvio padrão, valores mínimos e máximos, permitindo uma análise inicial dos atributos. Com base nos resultados, interpretamos as variáveis mais relevantes para o estudo, destacando tendências e padrões observados nos dados.

2.2. Interpretação das Principais Variáveis

- **Age (Idade):** A idade média dos pacientes é de 74 anos, com variação considerável (desvio padrão de 9 anos). A faixa etária (60 a 90 anos) reflete o foco do estudo em uma população mais suscetível ao Alzheimer.
- **Gender (Gênero):** A distribuição é equilibrada entre masculino e feminino, indicando que o estudo abrange ambos os gêneros de forma representativa.
- **BMI (IMC):** A maioria dos pacientes está na categoria de sobrepeso, com alguns casos extremos que variam de baixo peso até obesidade grau II.

- **EducationLevel (Nível Educacional):** Em média, os pacientes completaram o ensino médio, um dado relevante para análises relacionadas à reserva cognitiva.
- **FamilyHistoryAlzheimers (Histórico Familiar):** Cerca de 25% dos pacientes possuem histórico familiar de Alzheimer, indicando um fator de risco importante.
- **Diagnosis (Diagnóstico):** Aproximadamente 35% dos pacientes possuem diagnóstico de Alzheimer, oferecendo uma base robusta para o desenvolvimento de classificadores.

2.3. Identificação de Outliers e Limpeza de Dados

A análise de outliers foi realizada por meio de boxplots, onde identificamos a presença de valores extremos em algumas variáveis. No entanto, após avaliação, constatamos que os outliers em colunas como *Ethnicity* e *EducationLevel* não representavam erros nos dados, sendo considerados "falsos outliers". Por isso, não houve necessidade de tratamento adicional.

Além disso, verificamos a ausência de valores duplicados ou nulos, garantindo a integridade dos dados para as próximas etapas.

2.4. Visualização e Insights

Geramos gráficos para analisar distribuições e possíveis correlações entre variáveis.

- **Porcentagem de diagnóstico de Alzheimer por pessoa com antecedente familiar**
  - Descrição: O gráfico compara a porcentagem de diagnóstico positivo e negativo de Alzheimer em pessoas com e sem histórico familiar da doença.
  - Insight:
    - A presença de um histórico familiar parece ter um impacto relevante no diagnóstico, com uma maior proporção de diagnósticos positivos em indivíduos com histórico familiar(essa informação tiramos com base em pesquisas, mas no nosso dataset, a captação de dados que foi usada consta o contrário).
    - Este dado sugere que fatores genéticos podem ser significativos no risco de Alzheimer, reforçando a importância do acompanhamento em pessoas com histórico familiar.

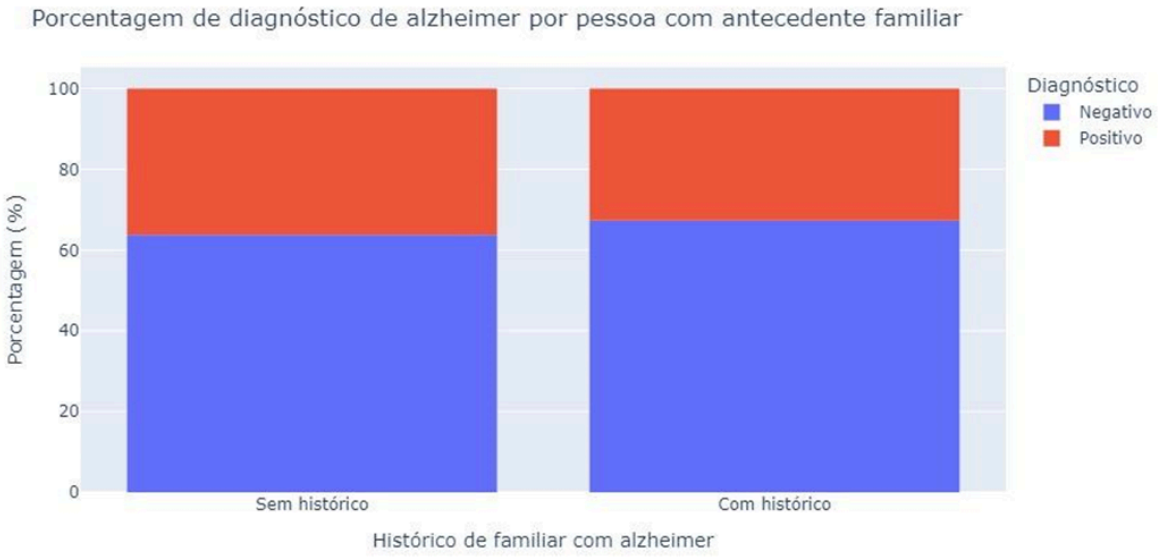

Figura 2. Porcentagem de diagnóstico de Alzheimer por pessoa com antecedente familiar.

- **Correlações entre diagnóstico de Alzheimer e outras variáveis**
  - Descrição: Um heatmap de correlação mostrando a relação entre o diagnóstico de Alzheimer e diferentes variáveis clínicas e comportamentais.
  - Insights:
    - MemoryComplaints apresenta uma correlação positiva (0.31), indicando que queixas de memória são preditores relevantes para um diagnóstico de Alzheimer.
    - FunctionalAssessment (-0.36) e ADL (-0.33) possuem correlações negativas significativas, sugerindo que a deterioração funcional e as dificuldades nas atividades de vida diária estão fortemente associadas à progressão da doença.
    - MMSE (-0.24) também mostra uma relação negativa, indicando que escores mais baixos no Mini-Mental State Examination estão associados ao diagnóstico.
    - BehavioralProblems (0.22), embora menor, ainda sugere uma relação notável entre problemas comportamentais e o diagnóstico.
    - Esses achados ajudam a destacar as áreas prioritárias para avaliação clínica e monitoramento de pacientes com risco.

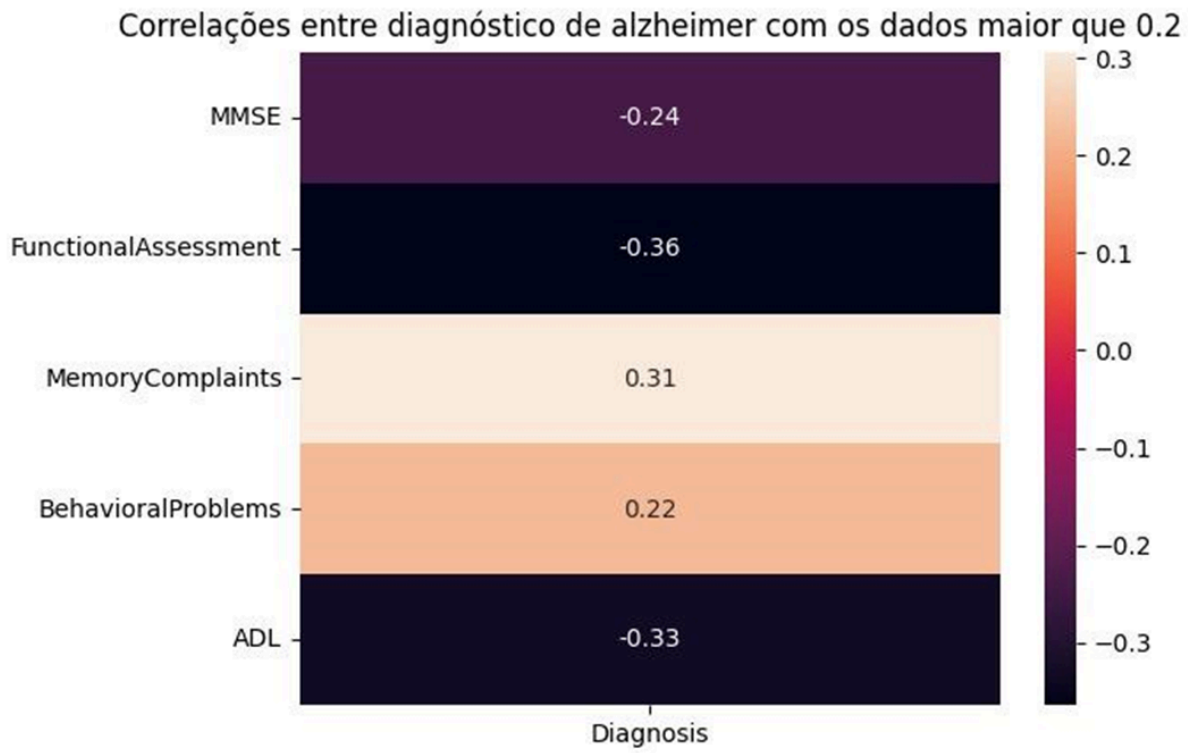

Figura 3. Correlação entre diagnóstico de alzheimer

- **Porcentagem de diagnóstico de Alzheimer por nível de educação**
  - Descrição: O gráfico apresenta a distribuição de diagnósticos positivos e negativos em diferentes níveis de educação.
  - Insights:
    - Há uma tendência de menor porcentagem de diagnósticos positivos em indivíduos com níveis mais altos de educação (bacharelado e pós-graduação).
    - Este padrão suporta a hipótese da "reserva cognitiva", onde maior escolaridade pode oferecer alguma proteção contra o desenvolvimento de sintomas clínicos de Alzheimer.
    - Indivíduos sem escolaridade apresentam um risco visivelmente maior, sugerindo a importância de promover educação como um fator de proteção cognitiva ao longo da vida.

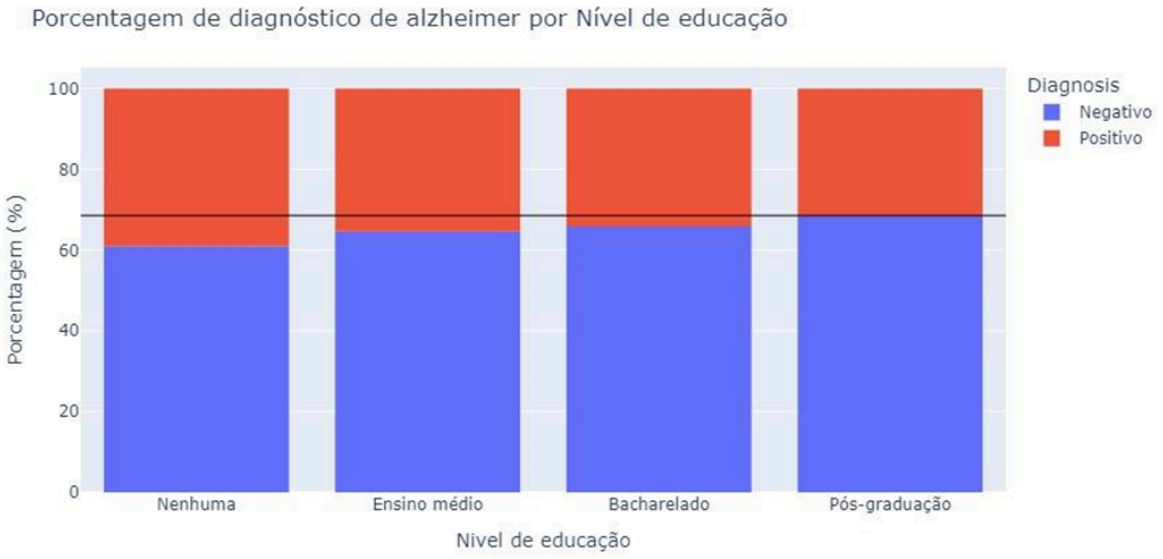

Figura 4. Nível de educação.

A análise exploratória revelou que o dataset é bem estruturado, com variáveis diversificadas e relevantes para o estudo. As informações extraídas forneceram um panorama claro para a construção do modelo de classificação e a definição das estratégias de pré-processamento e treinamento. A etapa de análise também reforçou a confiabilidade do dataset, que apresenta dados consistentes e sem valores faltantes ou duplicados.

3. Pré-Processamento:

No início do pré-processamento, verificamos duplicatas e inconsistências no dataset. Após uma análise cuidadosa, não foram identificadas duplicatas ou inconsistências nos valores ou tipos de dados das colunas.

Em seguida, realizamos a normalização dos dados para uniformizar as escalas das variáveis, essencial para algoritmos de aprendizado de máquina. Não foi necessário realizar transformações adicionais nos dados, pois as variáveis estavam bem definidas e compatíveis com o modelo pretendido.

Considerando o problema que desejamos solucionar, focamos na classificação da presença de Alzheimer com base em fatores de risco diretos. Esses fatores possuem impacto estabelecido na doença, seja pela contribuição ao declínio cognitivo, à progressão da condição ou pela associação em estudos epidemiológicos.

Além disso, identificamos fatores indiretos que podem ter uma relação com o Alzheimer, mas cuja influência depende de interações com outros fatores. Avaliamos suas correlações com o diagnóstico da doença para determinar sua relevância na análise.

Fatores Indiretos Avaliados: PatientID, Age, Gender, Ethnicity, EducationLevel, Smoking, PhysicalActivity, DietQuality, AlcoholConsumption, CholesterolTotal, CholesterolLDL, CholesterolTriglycerides, HeadInjury, PersonalityChanges e BehavioralProblems.

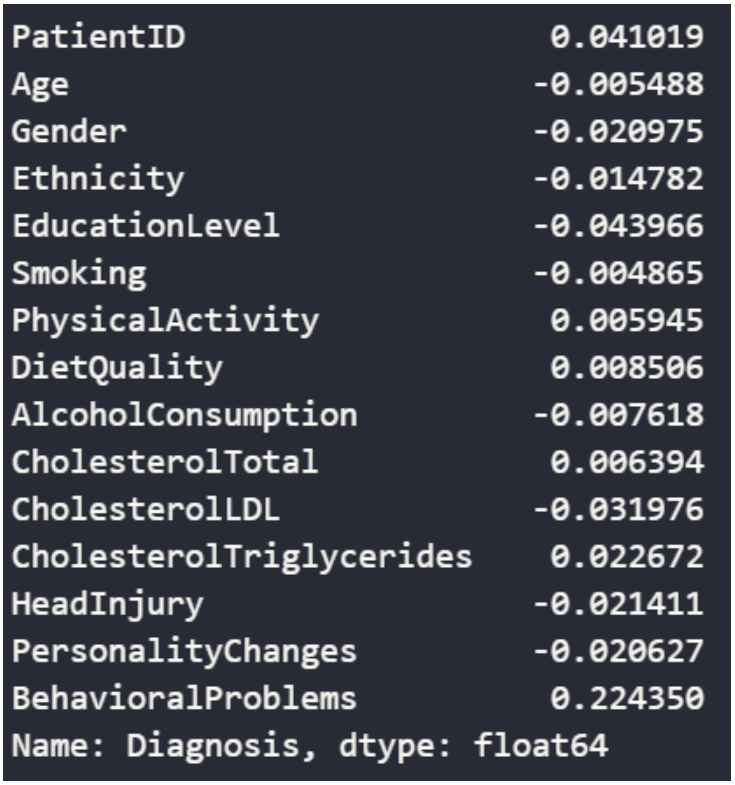

Figura 4. Correlação do diagnóstico de Alzheimer com fatores indiretos.

As análises mostraram que a maioria dessas colunas apresentava correlações muito baixas com o diagnóstico de Alzheimer. Embora a coluna BehavioralProblems tenha demonstrado uma correlação ligeiramente maior, sua utilidade foi limitada pela falta de detalhes sobre os tipos de problemas registrados.

Portanto, optamos por remover essas colunas, pois não contribuem significativamente para o processo de classificação, permitindo que o modelo se concentre nas variáveis de maior impacto.

4. **Modelagem de IA:**

Para a etapa de modelagem, realizamos a divisão dos dados em três subconjuntos: treinamento, validação e teste, visando garantir um processo eficiente de treinamento, ajuste e avaliação do modelo de aprendizado de máquina. A estratégia adotada seguiu a proporção padrão de 70% para treinamento, 15% para validação e 15% para teste.

Essa estratégia de divisão foi escolhida para maximizar o uso eficiente dos dados disponíveis e assegurar que cada conjunto desempenhe seu papel de forma clara e funcional. O conjunto de validação foi utilizado para avaliar e ajustar o desempenho do modelo durante o treinamento, enquanto o conjunto de teste foi reservado exclusivamente para medir a capacidade do modelo de generalizar para novos dados.

Além disso, foi garantida a manutenção das proporções das classes nos rótulos entre os subconjuntos, minimizando o risco de desbalanceamentos que poderiam distorcer os resultados. A reprodutibilidade das divisões foi assegurada, permitindo que os experimentos fossem replicados de forma consistente.

- **Quais modelos escolhidos?**

Optamos por escolher os modelos RandomForest, KNN e Logistic Regression para fazermos a aplicação do nosso caso.

- **Por que esses modelos?**

RandomForest e KNN, foram modelos que tivemos uma indicação para aplicarmos, visto que, precisávamos extrair um padrão de resultado esperado. O Logistic Regression foi uma escolha a parte, onde o escolhes por conta da sua funcionalidade matemática de reta das distribuições dos dados, onde conseguimos pegar o valor mais provável. Ambos foram escolhidos pelo retorno de valores binários.

- **Como foi a implementação?**

Como 65% dos nossos dados são de diagnóstico negativo de Alzheimer, aconteceu que os nossos modelos estavam mais enviesados a dizer negativo muito mais vezes, isso fez com que embora pareça uma acurácia alta, muitos casos positivos estavam erroneamente sendo classificados como negativos. Para resolver isso aplicamos uma técnica conhecida como oversampling, basicamente baseado nos dados anteriores novos dados são criados, nesse caso dados positivos de Alzheimer até que a quantidade de diagnósticos positivos e negativos se iguale e acabe com o viés.

- **Quais os resultados obtidos?**

Logistic Regression e KNN, não tiveram um resultado tão adequado por conta do formato do nosso dataset ser boa parte com valores categóricos, logo o RandomForest teve uma melhor performance em cima da nossa aplicação, logo abaixo vou deixar os resultados obtidos:

1. Random Forest:
   - Precisão: 94%
   - Recall: 93,5%

2. KNN
   - Precisão: 81%
   - Recall: 49%

3. Logistic Regression
   - Precisão: 82%
   - Recall: 61%