**Relatório do MVP: Previsão de Anemia**

**Objetivo do Trabalho**

O objetivo deste trabalho é desenvolver um modelo preditivo para identificar anemia com base nos níveis de hemoglobina e nas distribuições de cores das imagens. O problema a ser resolvido envolve a construção de um pipeline de dados que inclui a busca, coleta, modelagem, carga e análise dos dados. As perguntas a serem respondidas são:

* Qual é a distribuição das variáveis do dataset, e existem outliers significativos?
* Qual modelo preditivo é mais eficaz para identificar anemia com base nas características fornecidas?
* Quais são os principais fatores que contribuem para a previsão de anemia?

**1. Busca pelos Dados**

Para atingir os objetivos definidos, foi escolhido o dataset de previsão de anemia disponível no Kaggle. O dataset contém informações sobre a distribuição percentual de cores em imagens e níveis de hemoglobina, além de indicar se o indivíduo é anêmico ou não.

O dataset selecionado é o "Anaemia Prediction using Hb, Sex, %RBC prints", e possui as seguintes colunas:

Number: Identificador único.

Sex: Sexo do indivíduo (M/F).

%Red Pixel: Percentual de pixels vermelhos na imagem.

%Green Pixel: Percentual de pixels verdes na imagem.

%Blue Pixel: Percentual de pixels azuis na imagem.

Hb: Nível de hemoglobina em g/dL.

Anaemic: Indicador de anemia (Yes/No).

**2. Coleta**

O dataset foi coletado do Kaggle e carregado para o ambiente de análise. Para garantir que as informações estejam disponíveis para análise, foram criadas tabelas no MySQL Database, especificamente a tabela de fato fact\_anaemia e a tabela de dimensão dim\_sex. As informações foram importadas para essas tabelas e estão prontamente acessíveis para processamento.

**3. Modelagem**

Foi construído um modelo de dados no esquema estrela, com as seguintes tabelas:

1. Dimensão dim\_sex:

* sex\_id: Identificador único do sexo.
* sex: Valor do sexo (M/F).
* sex\_name: Valor do sexo em extenso (“Male”, “Female”)

1. Fato fact\_anaemia:

* case\_id: Identificador único.
* sex\_id: Identificador do sexo.
* red\_pixel\_percentage: Percentual de pixels vermelhos.
* green\_pixel\_percentage: Percentual de pixels verdes.
* blue\_pixel\_percentage: Percentual de pixels azuis.
* hemoglobin\_level: Nível de hemoglobina.
* anaemic\_status: Indicador de anemia.

Essas tabelas foram projetadas para suportar análises detalhadas e consultas eficientes.

**4. Carga**

Os dados foram carregados no MySQL Database através de uma pipeline ETL, envolvendo a transformação dos dados do formato CSV para o formato das tabelas no banco de dados. O código foi implementado para automatizar a carga dos dados e garantir a integridade dos dados durante o processo.

**5. Análise**

a. Qualidade dos Dados

Foi realizada uma análise de qualidade dos dados para identificar e tratar outliers.

A análise envolveu:

* Distribuição dos Dados: As colunas de percentual de pixels e nível de hemoglobina foram analisadas para identificar possíveis anomalias. Os outliers foram detectados utilizando o método IQR (Interquartile Range).
* Tratamento de Outliers: Após a identificação dos outliers, os dados foram filtrados para garantir que as análises subsequentes não fossem afetadas por valores extremos.

A figuras 1 mostra a distribuição de pacientes com anemia, onde podemos ver que aproximadamente 70% da base de dados após a filtragem dos outliers são pacientes sem anemia.

Figura 1 - Distribuição de pacientes com anemia

Gráfico, Gráfico de barras

Descrição gerada automaticamente

A figura 2 mostra a distribuição dos percentuais de pixels, indicando a frequência de cada percentual de pixel para cada cor.

Figura 2 - Distribuição dos percentuais de pixels

Gráfico, Histograma

Descrição gerada automaticamente

E, na figura 3, podemos ver um mapa de calor para a matriz de correlação, indicando como cada cor de pixel influencia no nível de hemoglobina. Nota-se que o percentual de pixel verde apresenta uma maior correlação (negativa) com o nível de hemoglobina.

Figura 3 – Mapa de calor para a matriz de correlação

Gráfico

Descrição gerada automaticamente

b. Solução do Problema

O problema foi solucionado através das seguintes etapas:

1. Análise Exploratória: Realizou-se uma análise detalhada das variáveis para entender sua distribuição e a relação entre elas. Foram identificados padrões e insights sobre a presença de anemia com base nas características fornecidas.
2. Construção do Modelo Preditivo: Diversos modelos de machine learning foram treinados utilizando os dados. Foram aplicados algoritmos como Regressão Logística, Random Forest e Support Vector Machine para prever a anemia. Cada modelo foi avaliado quanto à sua precisão, recall e F1-score para determinar o melhor desempenho.
3. Os dados foram carregados do banco de dados MySQL e normalizados para garantir que todas as variáveis estejam na mesma escala. A normalização é essencial para algoritmos de machine learning, como a regressão logística, que são sensíveis às escalas das variáveis.
4. Os dados foram divididos em conjuntos de treino e teste. O conjunto de treino foi usado para treinar o modelo e o conjunto de teste foi usado para avaliar seu desempenho.
5. Foi treinado um modelo de regressão logística para prever a condição de anemia. A regressão logística é um algoritmo de classificação que é adequado para prever uma variável binária, como a condição de anemia (0 para não anêmico e 1 para anêmico).
6. O modelo foi avaliado usando as métricas de acurácia, precisão, recall e f1-score.

Os resultados foram os seguintes:

Acurácia: 0.9375

precision recall f1-score support

0 0.93 1.00 0.96 25

1 1.00 0.71 0.83 7

accuracy 0.94 32

macro avg 0.96 0.86 0.90 32

A análise revelou que existem relações significativas entre as variáveis de distribuição de cores e os níveis de hemoglobina. A distribuição dos níveis de hemoglobina e as porcentagens de pixels de diferentes cores são fatores importantes na construção de um modelo preditivo eficaz para detectar anemia. O modelo de regressão logística desenvolvido mostrou um desempenho promissor, com uma acurácia de 93.75% e bons valores de precisão e recall.

**Conclusão**

Este trabalho demonstrou a criação de um pipeline de dados na nuvem utilizando Databricks e MySQL. Desde a coleta até a análise, cada etapa foi documentada e discutida. A análise inicial indica que os dados têm potencial para desenvolver um modelo preditivo eficaz para identificar anemia com base nos níveis de hemoglobina e distribuições de cores das imagens. O modelo de regressão logística treinado mostrou-se eficaz, mas há espaço para melhorias futuras.

**Autoavaliação**

Ao final do trabalho, foi possível atingir os objetivos traçados inicialmente. As dificuldades encontradas incluíram a necessidade de ajustar a modelagem dos dados e lidar com a qualidade dos dados. Para trabalhos futuros, recomenda-se aprofundar a análise dos modelos preditivos e explorar técnicas adicionais de pré-processamento e validação para melhorar ainda mais a precisão das previsões.