Graphical user interface, application

Description automatically generated

**MPEI 24/25**

Sistema de Deteção de Doenças

Relatório

Tiago Cepa (107578)

Daniel Simbe (110235)

**Introdução**

Este programa foi desenvolvido para ajudar na deteção de doenças com base nos sintomas fornecidos pelo usuário, utilizando abordagens de **filtros de Bloom**, **MinHash** e **Naive Bayes**. Ao iniciar o programa, o utilizador pode escolher entre duas opções:

1. **Introduzir sintomas manualmente**, onde o usuário fornece diretamente os sintomas para os quais deseja prever a doença associada.
2. **Utilizar inputs predefinidos**, onde o programa lê os sintomas de um arquivo CSV chamado **"sintomas.csv"**, que contém dados de doenças e seus respectivos sintomas.

Após a escolha da entrada, o programa irá realizar cálculos para prever a doença mais provável com base nos sintomas fornecidos. Ele aplica diferentes técnicas para análise, como **probabilidades condicionais** (Naive Bayes), **minhashing** para aproximação de semelhança entre sintomas e **filtro de Bloom** para otimizar a verificação de sintomas presentes.

**Funcionamento do Programa**

1. **Entrada de Dados**: O utilizador escolhe entre fornecer sintomas manualmente ou carregar os sintomas a partir do arquivo **"sintomas.csv"**.
   * **Sintomas Manuais**: Caso o utilizador escolha esta opção, ele insere diretamente os sintomas para os quais deseja fazer a previsão da doença associada.
   * **Leitura de Sintomas de CSV**: Quando o programa lê o arquivo CSV, ele processa os sintomas de cada doença e os organiza para posterior análise.
2. **Filtro de Bloom**: O filtro de Bloom é utilizado para otimizar a busca e filtrar rapidamente os sintomas presentes. Através de múltiplas funções hash, o filtro marca a presença ou ausência de sintomas e retorna um conjunto de sintomas filtrados.
3. **MinHash**: O algoritmo MinHash é usado para calcular as similaridades de Jaccard entre o conjunto de sintomas fornecidos pelo utilizador e as doenças presentes no dataset. Ele realiza isso por meio de permutações aleatórias e mapeamento de hashes para representar os sintomas de cada doença.
4. **Naive Bayes**: O modelo de Naive Bayes é empregado para calcular a probabilidade de cada doença com base nos sintomas fornecidos. Ele usa a frequência de ocorrência dos sintomas para calcular as probabilidades e fazer a previsão da doença.
5. **Main**: O programa permite ao utilzador escolher entre usar testes predefinidos (extraídos de Sintomas.csv) ou inserir sintomas manualmente. Após a escolha, a função Main é chamada, filtrando os sintomas com o Bloom Filter, calculando as probabilidades das doenças com Naive Bayes e as distâncias de Jaccard usando Minhash. Os resultados de cada técnica são exibidos no console. Após cada teste, o programa pergunta ao utilizador se deseja continuar ou encerrar a execução.
6. **Saídas**: Dependendo da técnica utilizada, o programa retorna:
   * Os sintomas que passaram pelo filtro de Bloom.
   * As similaridades de Jaccard para o.
   * As probabilidades preditivas de cada doença usando Naive Bayes.

**Utilização do Programa**

O programa pode ser utilizado de duas maneiras: fornecendo sintomas manualmente ou utilizando dados preexistentes do arquivo **"sintomas.csv"**. A seguir, temos uma explicação detalhada de como o programa pode ser executado e testado.

**Testes com Filtro de Bloom**

O Filtro de Bloom é uma estrutura de dados eficiente que permite verificar rapidamente se um elemento pertence a um conjunto. Ao passar sintomas para o filtro de Bloom, o programa verifica se os sintomas fornecidos pelo utilizador são compatíveis com os sintomas já registrados no sistema.

**Testes com MinHash**

O MinHash é usado para comparar a similaridade entre o conjunto de sintomas do utilizador e os sintomas registrados em doenças. Ele calcula uma assinatura de hash para o conjunto de sintomas de cada doença, e compara essas assinaturas para estimar a similaridade.

**Testes com Naive Bayes**

O Naive Bayes é utilizado para calcular a probabilidade de uma doença com base nos sintomas fornecidos. Ele avalia a probabilidade de cada doença, levando em consideração a frequência dos sintomas no dataset.

**Passo-a-Passo do Fluxo no Main**

A função MainLoop recebe como parâmetros o dataset (conjunto de dados de doenças e sintomas) e os sintomas fornecidos pelo utilizador (ou os sintomas de teste). Aqui está o que acontece dentro dessa função:

1. **Limpeza dos Sintomas de Entrada**:
   * Primeiramente, a função limpa os sintomas inseridos, removendo qualquer valor "missing" (ausente), para garantir que apenas dados válidos sejam processados.
2. **Bloom Filter**:
   * A função chama o BloomFilter para verificar quais sintomas fornecidos pelo utilizador realmente pertencem ao conjunto de sintomas registrados no dataset. Ela retorna os sintomas filtrados, ou seja, os sintomas que são reconhecidos pelo filtro de Bloom.
3. **Naive Bayes**:
   * A função então chama o método NaiveBayes para calcular a probabilidade de cada diagnóstico com base nos sintomas filtrados. O algoritmo de Naive Bayes é responsável por associar cada doença a uma probabilidade com base na presença dos sintomas.
4. **Minhash**:
   * Em seguida, a função chama o método Minhash para calcular as distâncias de Jaccard entre os sintomas inseridos e os diagnósticos possíveis. Isso é feito através de permutações aleatórias dos sintomas e comparação das assinaturas de Minhash.
5. **Exibição dos Resultados**:
   * Finalmente, os resultados de cada técnica são exibidos no console. O programa apresenta os sintomas filtrados, as probabilidades das doenças e as distâncias de Jaccard para ajudar o utilizador a entender quais diagnósticos são mais prováveis com base nos sintomas inseridos.

**Exemplo de Execução**

1. **Usando Testes Predefinidos**:
   * Se o utilizador optar por usar os testes predefinidos, o programa exibirá as probabilidades das doenças e distâncias de Jaccard para cada conjunto de sintomas no arquivo Sintomas.csv.
2. **Inserindo Sintomas Manualmente**:
   * Caso o utilizdor deseje inserir sintomas manualmente, o programa executará os algoritmos de processamento com os sintomas fornecidos, exibindo os resultados de forma similar.

No exemplo abaixo, o utilizador optou por usar os testes predefinidos e o programa calculou as probabilidades das doenças e as distâncias de Jaccard para os sintomas fornecidos no teste.

Interface gráfica do usuário

Descrição gerada automaticamente com confiança média

**Conclusão**

Com este projeto pudemos criar uma ferramenta para análise de sintomas e deteçao de doenças, utilizando técnicas avançadas como o filtro de Bloom, Naive Bayes e Minhash. Além de proporcionar uma análise eficiente e precisa para determinar as doenças mais associadas aos sintomas fornecidos, o programa também nos permite explorar e aprender mais sobre essas metodologias aplicadas à área da saúde. Profissionalmente, tais técnicas podem ser adaptadas e ampliadas, sendo de grande valor em sistemas de apoio a diagnósticos médicos, e outras áreas que envolvem análise de dados e inteligência artificial.