Classificação de Doenças usando Machine Learning

Igor Teixeira Brasiliano[1]

[1]Departamento de Informática – Universidade Federal de Viçosa (UFV)

igor.brasiliano@ufv.br

Resumo. *Projeto final apresentado à disciplina de Inteligência Artificial I, de código INF 420, ministrada pelo professor Julio C. S. Reis, como requisito parcial para aprovação na disciplina.*

1. Introdução

O objetivo do trabalho é, dada uma série de sintomas, identificar a doença relacionada.

2. Metodologia

Para realização da classificação, foi escolhido trabalhar com a biblioteca scikit-learn, muito utilizada em projetos envolvendo aprendizagem de máquina, e o ambiente de desenvolvimento foi o Google Colab.
O dataset utilizado foi o 'Disease Symptom Prediction', retirado do Kaggle[1]. É composto por 4920 amostras de sintomas-doença, sendo 120 de 41 enfermidades diferentes. Entre as doenças é possível encontrar: Micose, Hepatites (A, B, C e D), Pneumonia, Aids, Gastroenterite, entre outras. Já entre sintomas, pode-se ter: arrepios, vômito, dor de cabeça, fadiga, dor de barriga, disartria (fala arrastada e lenta), inchaço nos olhos e rosto, crostas amareladas na pele, entre outros. No caso de estudo, os sintomas são as features, identificadas com 1 para presença de determinado sintoma e 0 para a ausência, e as doenças são as classes de interesse.
Foram avaliados três modelos de classificação: Naive Bayes, KNN e Decision Tree.
A escolha dos algoritmos citados se deu pelo fato de que, por uma primeira análise, o problema não parece ser complexo.

- Naive Bayes: visto que para cada doença há a presença ou não do sintoma, então o funcionamento esperado é: dado determinados sintomas, qual é a probabilidade de ser determinada doença.
- KNN: espera-se que a vizinhança de uma instância de uma determinada doença seja composta, em sua maioria, por outras instâncias da mesma classe.

- Decision Tree: o comportamento esperado é que haja uma boa relação entre sintoma-doença, sendo possível uma boa diferenciação pela árvore.

Todos os algoritmos foram utilizados com parâmetros já definidos como padrão, sem necessidade de realizar Grid Search para melhoramento.
Em relação às métricas, foram utilizadas: acurácia, precisão e revocação. E a avaliação foi realizada utilizando cross-validation com 10 divisões.

3. Resultados

Os resultados obtidos foram ótimos, com valores de 1.0 (100%) para todas as métricas em todos os classificadores estudados. O que já era esperado - boas métricas -, já que o problema é bem separável, ou seja, um conjunto de sintomas classifica bem uma determinada doença, ou pelo menos as que se encontram no dataset de estudo.

	Acurácia		Precisão		Revocação	
	Treino	Teste	Treino	Teste	Treino	Teste
Naive Bayes	1.0	1.0	1.0	1.0	1.0	1.0
KNN	1.0	1.0	1.0	1.0	1.0	1.0
Decision Tree	1.0	1.0	1.0	1.0	1.0	1.0

Tabela 1: resultados obtidos

Ao final do documento é possível encontrar a matriz de confusão para os três classificadores. Para a montagem da tabela, o dataset foi dividido em treino e teste em 25% (3690 para treino e 1230 para teste), o algoritmo foi treinado com a partição de treino e avaliado com a partição de teste.

Você deve apresentar as análises e respostas evidenciais que fundamentam o objetivo do seu trabalho. Para trabalhos do tipo 1), você pode inserir alguns *prints* de tela, etc, em um nível satisfatório para explicação do funcionamento geral do jogo desenvolvido. Já para trabalhos do tipo 2) você pode apresentar, por exemplo, uma análise comparativa de medidas de desempenho das abordagens exploradas no cenário de interesse. Neste caso, gráficos/figuras e tabelas são bem vindos :)

4. Código

O código e dataset podem ser encontrados em minha página no github

5. Conclusão

Como é possível uma boa separação de doenças dados seus sintomas, métricas altas já eram esperadas, e foram encontradas no decorrer do trabalho. A parte mais trabalhosa foi a manipulação do dataset para que fosse possível deixá-lo num formato de entrada para os classificadores. A parte de treinamento e análise de resultados foram feitas sem nenhuma

dificuldade e problema aparente. Logo, os três classificadores se mostram bem robustos para classificação de doenças.

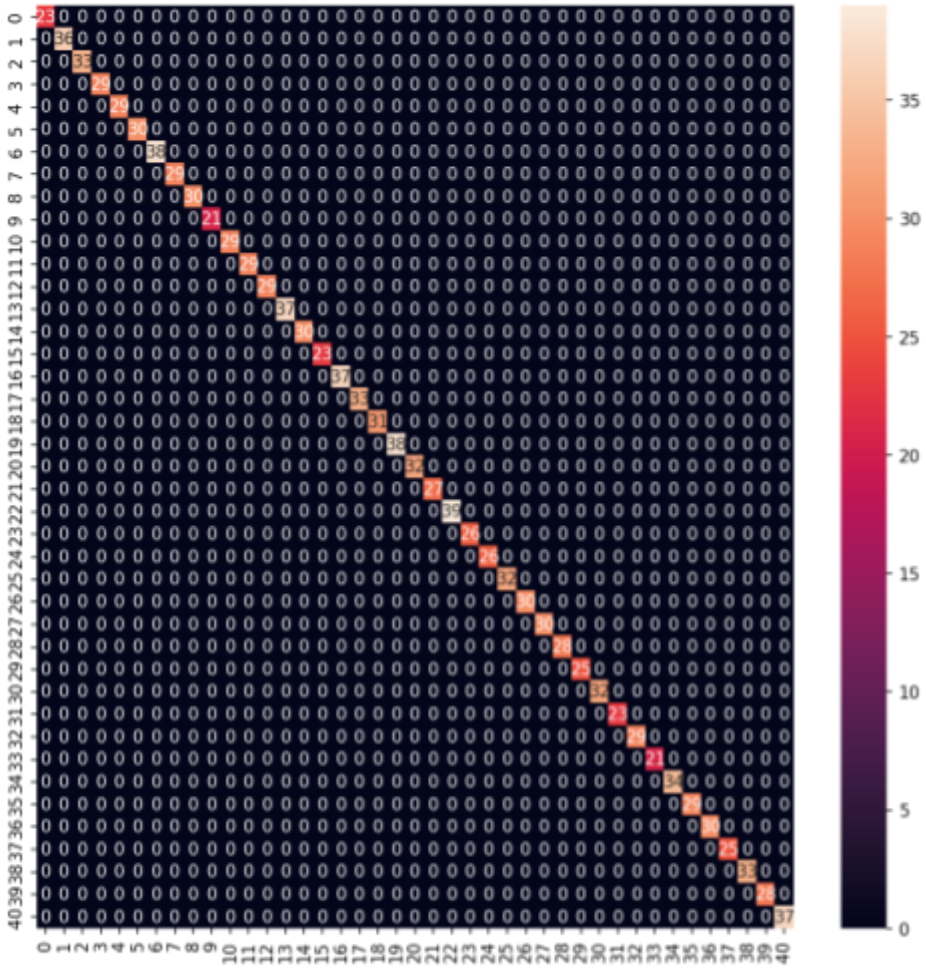

Figura 1: matriz de confusão para KNN

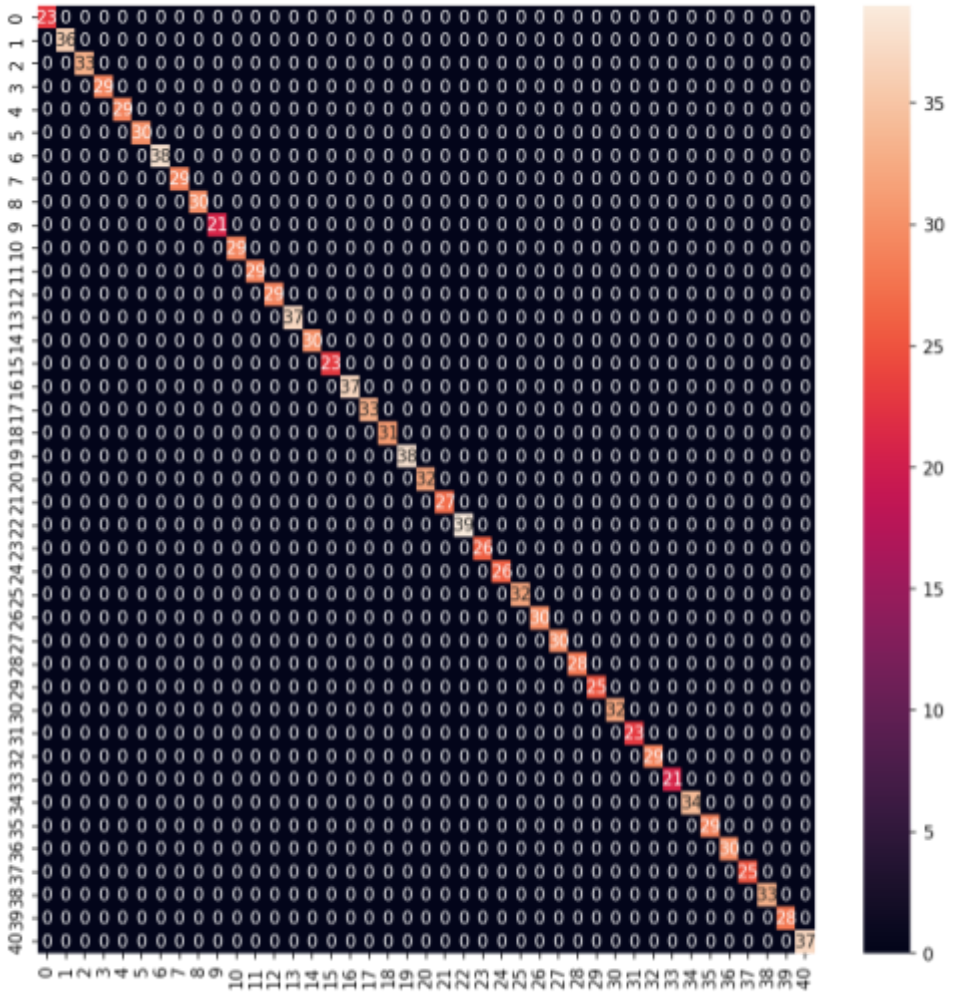

Figura 2: matriz de confusão para Naive Bayes

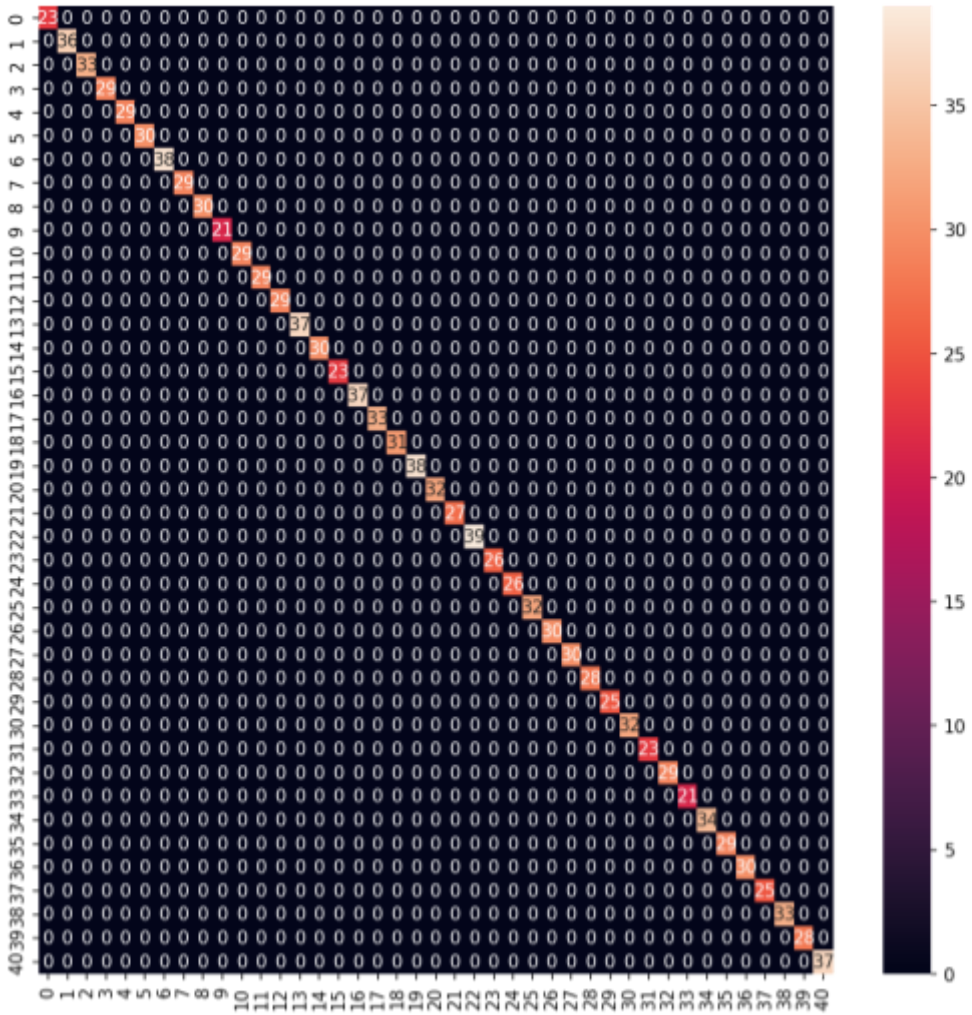

Figura 3: matriz de confusão para Decision Tree

Referências

[1] Disease Symptom Prediction, Kaggle. Disponível em <https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset?select=Symptom-severity.csv>