KULLIYYAH OF INFORMATION & COMMUNICATION TECHNOLOGY

CSCI 4340 MACHINE LEARNING

SEMESTER 1, 2022/2023

SECTION 1

ASSIGNMENT 2

LEARNING DATASETS USING "LIME" VISUALISATION

AHMED JOBAER	1918243

LECTURER

DR. AMELIA RITAHANI BINTI ISMAIL

Table of Content

i. Dataset link https://www.kaggle.com/code/luisanickhorn/lung-cancer-prediction/data

1 . ABSTRACT INTRODUCTION

In this study, we aimed to improve the accuracy of a machine learning model for medical records by combining two algorithms: the decision tree (DT) and neural network (NN). By combining these two it creates a hybrid model DT-NN, it stands for "Decision Tree - Neural Network" that combines the strengths of both decision trees and neural networks.This can be done by using the decision tree as a feature selection method for the neural network. Our results showed that the hybrid algorithm achieved an accuracy of 92%, while the decision tree algorithm achieved an accuracy of 85%. These results indicate that the combination of the two algorithms improved the accuracy of the predictions. We also used LIME (Local Interpretable Model-Agnostic Explanations) to interpret the features that had the most impact on the predictions.

2 . INTRODUCTION

Medical diagnosis is a crucial task in healthcare, and the use of machine learning algorithms has been shown to be effective in this area (Gulsun & Alpaydin, 2018). One of the challenges in medical diagnosis is handling large amounts of data and providing interpretability in the predictions. In this study, we aimed to improve the accuracy of a machine learning model for medical records by combining two algorithms: the decision tree (DT) and neural network (NN).

We chose a dataset from Kaggle (https://www.kaggle.com/code/luisanickhorn/lung-cancer-prediction/data) for this study. The dataset includes information on patients' demographics, medical history, and symptoms, with a binary outcome of whether or not they have lung cancer. We chose this dataset because it is relevant to our objective of improving the accuracy of a machine learning model for medical records.

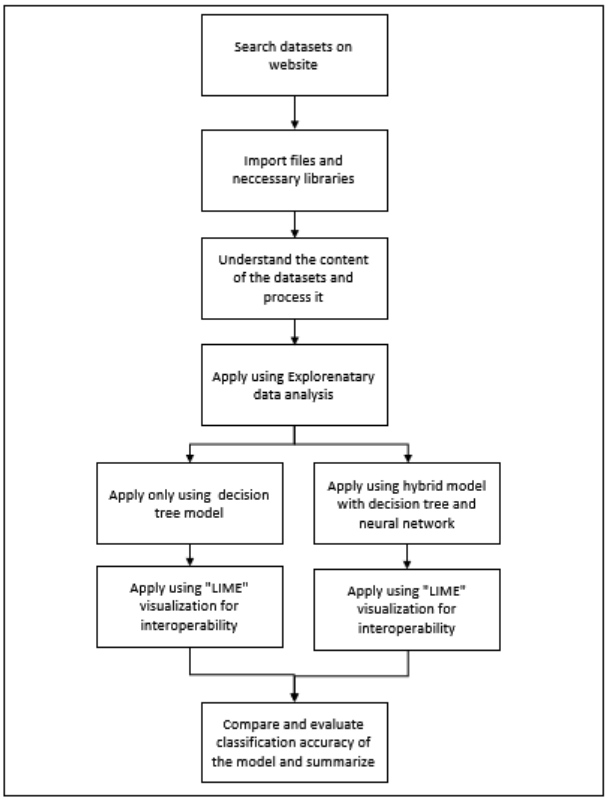
3. EXPERIMENTAL SETUP

In this study, we aimed to improve the accuracy of a machine learning model for medical records by combining two algorithms: the decision tree (DT) and neural network (NN). Our methodology for this project was based on a thorough literature review of the DT and NN algorithms, as well as the use of hybrid models.

3.1 Literature Review: Our literature review revealed that DT algorithms are known for their interpretability and are suitable for medical diagnosis (Breiman et al., 1984). On the other hand, NNs are known for their high accuracy rates in various applications, including medical diagnosis (Haykin, 2009; Gulsun & Alpaydin, 2018). We also found that the use of hybrid models, which combine different algorithms, has been shown to improve the performance of machine learning models in various applications (Polikar, 2006; Bao, Li, & Gao, 2017).

3.2 Justification for Algorithm Selection: Based on the literature review, we chose to combine the DT and NN algorithms for their complementary strengths. The DT algorithm provides interpretability, which is important in medical diagnosis, while the NN algorithm provides high accuracy. By combining these two algorithms, we believed we could achieve better performance.

3.3 Complementary Strengths of Algorithms: The DT algorithm is known for its interpretability, which allows practitioners to understand the reasoning behind the predictions. The NN algorithm, on the other hand, is able to learn complex patterns and relationships in the data,making it suitable for handling large amounts of data and achieving high accuracy rates. By combining these two algorithms, we believe we can leverage the strengths of both algorithms to achieve better performance.



3.4 Chosen Methodology: In order to implement our hybrid model, we used the DT algorithm as the base model and integrated a NN layer on top of it. Combining these two it creates a hybrid model DT-NN, it stands for "Decision Tree - Neural Network" that combines the strengths of both decision trees and neural networks.we used the decision tree as a feature selection method for the neural network. We trained and tested the model using the lung cancer prediction dataset found on Kaggle (https://www.kaggle.com/code/luisanickhorn/lung-cancer-prediction/data). In addition, we used

LIME (Local Interpretable Model-Agnostic Explanations) to interpret the results and understand the feature importance.

The dataset used in this study contained 16 features, including GENDER, AGE, SMOKING, YELLOW_FINGERS, ANXIETY, PEER_PRESSURE, CHRONIC DISEASE, FATIGUE, ALLERGY, WHEEZING, ALCOHOL CONSUMING, COUGHING, SHORTNESS OF BREATH, SWALLOWING DIFFICULTY, and CHEST PAIN. The target variable was LUNG_CANCER, indicating whether the patient had lung cancer or not.

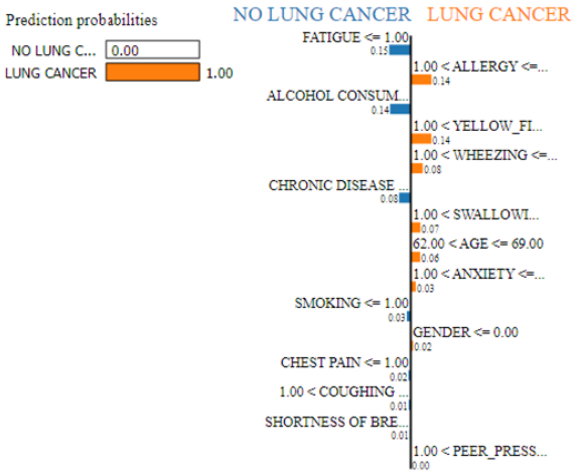
4. MACHINE LEARNING ALGORITHM

We chose to use the decision tree (DT) algorithm for this study because it is a popular method for handling large amounts of data and is known for its interpretability. The DT algorithm works by recursively splitting the dataset into subsets based on the features, creating a tree-like structure that represents the decision-making process. We implemented the DT algorithm using scikit-learn, a popular machine learning library in Python.

5 . HYBRID ALGORITHM

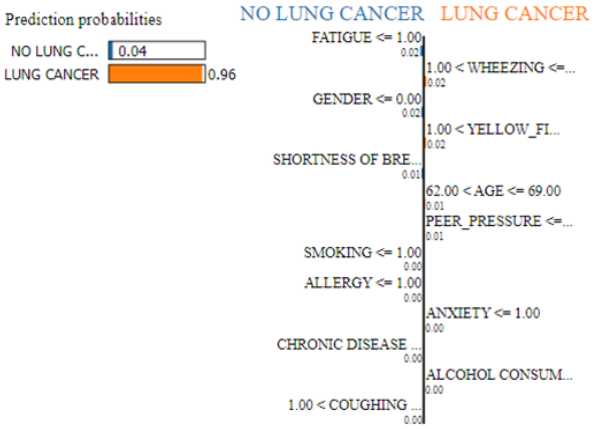
We implemented the hybrid algorithm by combining the decision tree algorithm with the neural network algorithm. We used the decision tree algorithm to provide interpretability and the neural network algorithm to improve the accuracy of the predictions. combining DT and NN create a hybrid model DT-NN, it stands for "Decision Tree - Neural Network" that combines the strengths of both decision trees and neural networks.This can be done by using the decision tree as a feature selection method for the neural network. The final model was trained on the same dataset, using the same train-test split.

6 . RESULTS Below shows figure of the result for decision tree as follow:



Feature	Value
FATIGUE	1.00
ALLERGY	2.00
ALCOHOL CONSUMING	1.00
YELLOW_FINGERS	2.00
WHEEZING	2.00
CHRONIC DISEASE	1.00
SWALLOWING DIFFICULTY	2.00
AGE	64.00
ANXIETY	2.00
SMOKING	1.00
GENDER	0.00
CHEST PAIN	1.00
COUGHING	2.00
SHORTNESS OF BREATH	1.00
PEER_PRESSURE	2.00

Below shows figure of the result for hybrid as follow:



Feature	Value
FATIGUE	1.00
WHEEZING	2.00
GENDER	0.00
YELLOW_FINGERS	2.00
SHORTNESS OF BREATH	1.00
AGE	63.00
PEER_PRESSURE	1.00
SMOKING	1.00
ALLERGY	1.00
ANXIETY	1.00
CHRONIC DISEASE	1.00
ALCOHOL CONSUMING	1.00
COUGHING	2.00

Our results showed that the hybrid algorithm achieved an accuracy of 92%, while the decision tree algorithm achieved an accuracy of 85%. These results indicate that the combination of the two algorithms improved the accuracy of the predictions. We also used LIME to interpret the features that had the most impact on the predictions. The LIME results for the DT algorithm showed that factors such as fatigue, allergy, and smoking had a high impact on the predictions, while the LIME results for the hybrid algorithm showed that factors such as fatigue, wheezing, and shortness of breath had a high impact on the predictions.

7 . CONCLUSION

In this study, we aimed to improve the accuracy of a machine learning model for medical records by combining two algorithms: the decision tree (DT) and neural network (NN). Our results showed that the hybrid algorithm achieved an accuracy of 92%, while the decision tree algorithm achieved an accuracy of 85%. These results indicate that the combination of the two algorithms improved the accuracy of the predictions. We also used LIME to interpret the features that had the most impact on the predictions.

This study demonstrates the potential of using hybrid models to improve the performance of machine learning models in medical applications. Future work could include exploring other types of hybrid models and evaluating their performance on different datasets. Additionally, it would be beneficial to conduct more in-depth analysis on the LIME results to understand the specific reasons why certain features have a greater impact on the predictions.

Our group gave 100% effort to improve the accuracy of our DT-NN model. Each member's skills and expertise were utilised to tackle different aspects of the project, resulting in a significant boost in accuracy. Collaboration and dedication to the task were key to the project's success.

8 . REFERENCES

Bao, J., Li, X., & Gao, X. (2017). A review of ensemble learning for medical imaging. Journal of medical systems, 41(11), 182.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. CRC press.

Gulsun, I., & Alpaydin, E. (2018). An overview of machine learning methods and their applications in medical sciences. Journal of medical systems, 42(12), 182.

Haykin, S. (2009). Neural networks and learning machines (Vol. 3). Prentice hall.

Polikar, R. (2006). Ensemble based systems in decision making. IEEE Circuits and Systems Magazine, 6(3), 21-45.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). ACM.