Sima_Masoumi_And_Ansar_Atiq_FinalProject

August 19, 2023

1 Title: Cardiovascular Disease Prediction Using Deep Learning

Group Member Names : Sima Masoumi, Ansar Atiq

1.0.1 INTRODUCTION:

AIM : In The original paper that I have selected Ensemble framework for cardiovascular disease prediction tree-based models were used, in this paper we are exercising deep learning to make improvement in the model performance. At the end, the evaluation metrics of the Deep Learning model will be compared to the original tree based models to see if the Deep Learning model can be more accurate in predicting heart disease.
****

Github Repo: https://github.com/Simamsm/ml_project
****

DESCRIPTION OF PAPER: The paper presents a machine learning-based framework for predicting heart diseases, outlining its structure including literature review, algorithms, experiments, and conclusions. The framework aims to enhance heart disease prediction, benefiting medical practices and patient care.
****

PROBLEM STATEMENT : * Try to replicate the results given in paper on the binary class classification to predict heart disease. * Instead of the original tree based models, use a Deep Learning approach which can be trained on a new data set. * To compare the performance of the Deep Learning model with the original tree based models
****

CONTEXT OF THE PROBLEM: * The tree based models usually perform very well, however, Deep Learning models can also be very accurate in other cases, usually when there is more data available. * In this work, deep learning models will be used and if the performance is better or even equal to the tree based models, it will be a better alternative. In case of equal performance, deep learning models will perform better in future once there is more data available.
****

SOLUTION: * There will be Exploratory Data Analysis on the data to understand the features impacting heart disease * Then there will be a deep learning model trained to be able to predict the heart disease.

2 Background

In the original paper, they divided their work into 4 steps: 1. Data understanding and EDA 2. Training of various models 3. Feature selection 4. Final model training by stacking all the models. 5. Evaluating the final model

No.	Reference	Explanation	Dataset Used	Weakness
1	Modak et al.(2022)	Multilayer perceptron for heart disease classification.	Cleveland, Hungarian, Switzerland, Long Beach, Statlog	The approach's performance on datasets other than the specified ones is unknown. Additionally, the interpretability of the multilayer perceptron model is limited.
2	Sarah et al.(2022)	Logistic regression for heart disease classification.	Cleveland	Logistic regression might struggle if there are complex nonlinear relationships in the data.
3	Nguyen et al.(2021)	Naive Bayes, Logistic Regression, SVM, and Decision Trees for heart disease classification.	Cleveland	Naive Bayes and Logistic Regression could underperform on datasets with high dimensionality. SVM and Decision Trees might require extensive hyperparameter tuning for optimal results.
4	Latha et al.(2019)	Random Forest, Multilayer Perceptron, Bayes Net, Naive Bayes for heart disease classification.	Cleveland	Random Forest and Multilayer Perceptron might struggle with imbalanced datasets. The Bayes Net and Naive Bayes might struggle if independence assumptions are violated.
5	Atallah et al.(2019)	SGD, KNearest Neighbor, Random Forest, Logistic Regression for heart disease classification.	Cleveland	SGD Classifier's convergence might be sensitive to learning rate and scaling. KNearest Neighbor's performance could be affected by noisy or irrelevant features. Random Forest's interpretability might be limited compared to simpler models.
6	Pawlovsky (2018)	Weighted k-nearest neighbour for heart disease classification.	Cleveland	The performance of weighted k-nearest neighbor might degrade in high-dimensional spaces. Sensitivity to the choice of distance metric and k-value.
7	Bialy et al.(2016)	Ensemble of FDT, C4.5, MLP, SVM, and Naive Bayes for heart disease classification.	Cleveland	Ensemble methods might require additional computational resources. The performance gain from combining models might plateau if base models are similar.

3 Implement paper code :

- The code from the paper was implemented and there were some errors from differences in library versions and some other matters. They were taken care of. The paper code after the updates can be found in here

3.0.1 Contribution Code :

- The notebook containing the contribution code can be found here Steps taken in the code:

1. Data wrangling and understanding.
2. Exploratory data analysis
3. Data pre-processing
4. Model building using deep learning
5. Applying early stopping
6. Finding the best threshold for the positive and negative classes using the Geometric Mean (G-Mean) method
7. Model performance evaluation

3.0.2 Results :

Some of the factors contributing heart disease: 1. Patients with heart disease tend to have lower cholesterol level 2. With the data we have in hand, it seems like the patients suffering from heart disease tend to be younger Patients with downsloping have the highest rate of heart disease, around 75 percent. 3. Patients with resting electrocardiographic results being ST-S Abnormality have the highest rate of heart disease, about 60 percent of them had a heart disease 4. IT seems like fasting blood sugar $>$ 120 mg/dl does not have a direct impact on heart disease.

Observations :

- The deep learning model has a recall score of 97 percent on the test data

3.0.3 Conclusion and Future Direction :

Conclusion: In this project, we found the factors contributing into heart disease (mentioned above) and also a deep learning model was trained to predict heart diseas. The deep learning model, although its performance is very well, it can benefit from having more training data. As we know, neural network models will have higher performance when more data is available, so as a future work to this project, more training data can be used and then the subsequent model's performance can be compared to the model trained in this project. #### Learnings :

We learned how deep learning methods can be used on a data set with both categorical and numerical features. Also we learned how early stopping method can stop the training at the right time to prevent from overfitting. We also learned how we can set athreshold for the predictions to have positive and negative classes in the predictions. #### Results Discussion :

Limitations : There was limitation on the amount of data available to train the model. There were around 1.2K samples which is typically not enough for deep learning models, however in this case, the model performance was still very good.

Future Extension : In future, hyperparameter tuning can be done to tune the parameters available in deep learning models. Also collecting more training data will be very important to have an improvement in model performance.

4 References: