Heart Attack Risk Prediction

Dragomir Božoki Faculty of Technical Sciences Novi Sad, Serbia Biomedical Engineering

I. INTRODUCTION

The Heart Attack Risk Prediction Dataset serves as a valuable resource for delving into the intricate dynamics of heart health and its predictors. Heart attacks, or myocardial infarctions, continue to be a significant global health issue, necessitating a deeper comprehension of their precursors and potential mitigating factors. This dataset encapsulates a diverse range of attributes including age, cholesterol levels, blood pressure, smoking habits, exercise patterns, dietary preferences, and more, aiming to elucidate the complex interplay of these variables in determining the likelihood of a heart attack. By employing predictive analytics and machine learning on this dataset, researchers and healthcare professionals can work towards proactive strategies for heart disease prevention and management.

II. DATA BASE AND DATA ANALYSIS

This dataset encompasses diverse attributes, including age, cholesterol level, blood pressure, smoking habits, exercise patterns, dietary preferences, and more, aiming to shed light on the complex interaction of these variables in determining the likelihood of a heart attack. The database contains 8763 samples and 26 different features. After data preprocessing, which involved removing irrelevant features and adjusting the remaining attributes, the dataset comprises 11 categorical and 11 numerical features. The target feature indicates whether a patient is at risk of a myocardial infarction. The proportion of samples showing a risk of a heart attack is 36%, while the proportion of samples not showing this risk is 64%, roughly in a 2:1 ratio. Such a disproportion can pose a challenge when applying certain machine learning algorithms, thus necessitating strategies to address this issue.

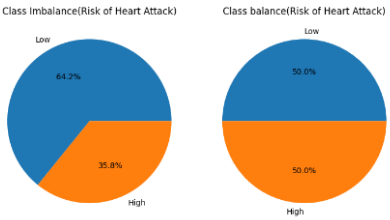

Figure 2.1 Original and Balanced Class Distribution

The target feature, "Heart attack risk," was extracted from the original database and placed into a separate vector "y." To equalize the percentage of members of both classes to 50:50, oversampling was performed, increasing the number of samples in the less represented class by replicating its members. This operation was carried out due to the k-Nearest Neighbors (kNN) classification algorithm, which is sensitive to unequal sample distributions across classes.

III. MODEL SELECTION AND HYPERPARAMETERS

Three machine learning methods were selected to address the classification problem: k **Nearest Neighbors** (kNN), **Random Forest Classifier** (RFC), and **Neural Networks** (NN). Each of these methods has its characteristics and advantages that make them suitable for data analysis and classification. The data were then divided into training and test sets, with the test data comprising 10% of the original dataset. Subsequently, data standardization was performed through ***z-normalization***, providing a zero mean and unit standard deviation. For the Random Forest Classifier and Neural Network algorithms, the original class distribution was retained, while for the kNN algorithm, the modified distribution through oversampling was applied. The combination of these methods can provide a comprehensive solution for solving classification problems. All three algorithms were tested on datasets processed with both **Principal Component Analysis** (PCA) for dimensionality reduction and without PCA. The top Fhree performing algorithms were selected based on their performance. The selection of the best **hyperparameters** for the

chosen models was achieved using the ***GridSearchCV*** method.

IV. Results of algorithms on the test set with and without PCA algorithm

	precision	recall	f1-score	support
0	0.740260	0.498252	0.595611	572.000
1	0.612162	0.819168	0.700696	553.000
accuracy	0.656000	0.656000	0.656000	0.656
macro avg	0.676211	0.658710	0.648154	1125.000
weighted avg	0.677293	0.656000	0.647266	1125.000

Figure 4.1 report of kNN without PCA

In *Figure 4.1*, the depicted scores represent the performance metrics of the optimal algorithm, specifically kNN (k-Nearest Neighbors), when applied to the dataset without the utilization of Principal Component Analysis (PCA). It is noteworthy that the algorithm was trained on an oversampled database. This implies that the training dataset was augmented through duplication of minority class samples, thereby balancing class distribution. The efficacy of kNN, trained on this oversampled dataset, provides a solid assumption regarding its performance on the original sample instances. This assumption stems from the inherent nature of kNN, which relies on the proximity of instances in feature space for classification. By ensuring a balanced representation of class instances through oversampling, the algorithm is better equipped to discern patterns and relationships within the data, thus enhancing its predictive capabilities.

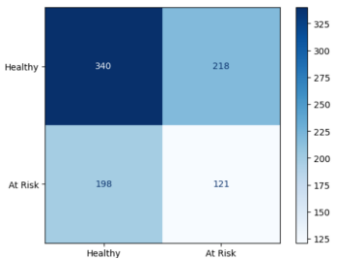

Figure 4.2 ConfusionMatrix of NN

The confusion matrix of the second-best algorithm reveals a recall rate of 37%, indicating its performance on an imbalanced database. It's evident that there's a significant degradation in performance compared to the leading kNN algorithm. This decline could be attributed to either inherent limitations of the algorithm itself or the absence of a balanced database. Notably, the performance of other methods showcased similar results, with performance fluctuating between suboptimal outcomes. This variability can be attributed to algorithms encountering local minima during optimization, which may either surpass or approach the global minimum. In *V. **conclusion***, we delve into potential reasons behind the observed low performance.

Conclusion

In this scientific study, models tested exhibited accuracy below 0.67 without oversampling. With 3169 positive risk indicators of heart attack, roughly 35% of the data, even the simplest model predicting all instances as negative would achieve around 0.65 accuracy. This suggests the dataset is practically useless for prediction due to lack of clear patterns and real-world reflection. Such observations hold true for models tested in this study and others. These results likely stem from dataset characteristics, possibly stemming from data generation algorithms like ChatGPT, emphasizing the need for critical examination of data quality and representativeness. Additionally, there's a call for improved data generation methods to ensure model reliability and real-world applicability.

V. References

[1] **‘’Practicum for MachineLearning‘’** - *Tijana Nosek, Branko Brkljač, Danica Despotović, Milan Sečujski, Tatjana Lončar Turukalo*

[2] **‘’Machine learning PyTorch and Scikit-Learn‘’** -

Sebastian Raschka, Yuxi (Hayden) Liu, Vahid Mirjalili

[3] **Kaggle DataSet NoteBooks for the DataBase**

[4] **DataBase**