Please include the following in your report.

1. Abstract (summary of the project)
2. Introduction (Describe the problem and the summary of your contribution
3. Methods (Describe the dataset, the methods you followed, and the methods other people followed
4. Results (Results according to your method and the other methods)
5. Conclusion (Final remarks about the project)

Abstract

In this project, a dataset on metabolic syndrome was analyzed. Multiple machine learning models were created and optimized to see which one works best. The model that provided the best accuracy was saved. A website a was created that uses multiple saved machine learning models to provide the users a way to predict their likelihood of having a certain disease in the future. The models that were saved were the metabolic syndrome model, a breast cancer detection model that takes breast ultrasound images, and predicts whether it shows breast cancer or not, and a stroke prediction model. The current version of the website takes inputs from the user for which disease they are interested in and takes the input of the user’s personal health data that is required by the machine learning model. After that data is provided, and the user clicks submit button, an email is sent out to the user with their results to ensure privacy. A risk analysis was conducted on the dataset to figure out the prior probability of the metabolic syndrome, as well as the posterior probability and likelihood. For the metabolic syndrome dataset, the model that worked best on it was XGBoost classification model.

Introduction

Metabolic syndrome is a complex medical condition associated with multiple risk factors for cardiovascular diseases and type 2 diabetes. This project aims to develop and optimize machine learning models to predict the likelihood of metabolic syndrome using an advanced data analysis approach. By leveraging a dataset obtained from Kaggle, the research explores various machine learning techniques to create an accurate predictive model.

The primary objectives of this study were to: (1) preprocess and analyze the metabolic syndrome dataset, (2) compare multiple machine learning classification models, (3) identify the most effective predictive model, and (4) develop a web-based platform for disease risk assessment. The project goes beyond metabolic syndrome prediction, extending to additional health prediction models including breast cancer detection and stroke prediction.

A key innovative aspect of this research is the development of a web platform that ensures user privacy. The system takes user input, processes health data through machine learning models, and sends results via email. This approach provides a secure and accessible method for individuals to understand their potential health risks.

Methods:

The dataset that was used was obtained from Kaggle: <https://www.kaggle.com/datasets/antimoni/metabolic-syndrome>.

It contains information on individuals with metabolic syndrome. Metabolic syndrome is a complex medical condition associated with a cluster of risk factors for cardiovascular diseases and type 2 diabetes. The data includes demographic, clinical, and laboratory measurements, as well as the presence or absence of metabolic syndrome. There are a total of 15 columns including the response variable and the dataset contains 2402 data points. After reading the csv in python using pandas, two columns were removed. They were “seqn” and “marital”. These were removed because they did not affect the final outcome. The categorical values were converted into numerical values. For the “sex” column, “Male” was assigned a number 0, and “Female” was assigned a number 1. Similar mappings were generated for the “Race” column manually. The columns that contained missing values were filled out using the fillna() function using the default parameter which fills the null values with the mean of the entire dataset. The columns that contained missing values were columns “Income”, “WaistCirc” and “BMI”. The dataset was then manually separated into training and testing data to make sure both outcomes of the dependent variable were present in equal quantities. The training set was 1/3 the size of the total dataset. Random over sampling was performed in the training dataset. This marked the end of the data preprocessing step. After this step, multiple models were selected and tested. Those models included Random Forest Classifier, Decision Tree Classifier, XGBClassifier, Logistic Regression, MLP Classifier, and TabNet Classifier. Their accuracy, F1 score, precision and confusion matrix were printed out. Before finalizing the preprocessing strategy, there was also a comparison of different methods done. Accuracy was compared between models where Random over sampling was done before splitting the data, model where no random over sampling was done, random over sampling was done on both training and testing data after splitting the dataset, and when random over sampling was done only on the training set. The results are shown below.

Results:

Figure 1. Comparison of the accuracy of the models across different random over sampling strategies.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Accuracy with ROS before splitting** | **Accuracy without ROS** | **Accuracy with ROC after splitting** | **Accuracy with ROC after splitting only on training set** |
| Random Forest Classifier | 93.14% | 86.69% | 87.5% | 86.82% |
| Decision Tree Classifier | 91.03% | 83.77% | 82.83% | 83.91% |
| XGB Classifier | 94.09% | 87.24% | 88.15% | 88.63% |
| Logistic Regression | 79.22% | 80.86% | 77.93% | 77.67% |
| MLP Classifier | 77.64% | 54.51% | 78.59% | 76.84% |
| Tab Net | 82.17% | 67.27% | 80.87% | 76.84% |

|  |  |  |  |
| --- | --- | --- | --- |
| Models | Precision without ROS | Precision with ROC after Splitting | Precision with ROC after splitting only on the training set |
| Random Forest Classifier | 84.23% | 88% | 80.29% |
| Decision Tree Classifier | 78.8% | 87.55% | 78.16% |
| XGB Classifier | 83.66% | 91.1% | 85.09% |
| Logistic Regression | 76.17% | 77.63% | 66.23% |
| MLP Classifier | 43.6% | 76.56% | 64.4% |
| Tab Net | 52.72% | 74.06% | 61.75% |

Figure 2. Precision of the models across different random oversampling strategies.

|  |  |  |  |
| --- | --- | --- | --- |
| Models | Recall without ROS | Recall with ROC after Splitting | Recall with ROC after splitting only on the training set |
| Random Forest Classifier | 78% | 86.73% | 84.29% |
| Decision Tree Classifier | 75.4% | 79.56% | 78.16% |
| XGB Classifier | 80.46% | 84.56% | 83.14% |
| Logistic Regression | 68.5% | 78.47% | 78.16% |
| MLP Classifier | 88.1% | 82.39% | 80.45% |
| Tab Net | 92.72% | 95% | 94.63% |

Figure 3. Recall of the models across different random oversampling strategies.

|  |  |  |  |
| --- | --- | --- | --- |
| Models | F1 without ROS | F1with ROC after Splitting | F1 with ROC after splitting only on the training set |
| Random Forest Classifier | 80.87% | 87.40% | 82.24% |
| Decision Tree Classifier | 77.10% | 83.37% | 78.16% |
| XGB Classifier | 82.03% | 87.71% | 84.10% |
| Logistic Regression | 72.17% | 78.05% | 71.70% |
| MLP Classifier | 58.37% | 79.37% | 71.55% |
| Tab Net | 67.22% | 83.23% | 74.73% |

Figure 4. F1 score of the models across different random oversampling strategies.

Figure 5. Comparison of Accuracy of different models when ROS is done on the training data.

Figure 6. Comparison of Accuracy of different models when no ROS is done.

Conclusion:

The project successfully demonstrated the effectiveness of machine learning in predicting metabolic syndrome. Through careful data preprocessing and model optimization, the XGBoost classification model emerged as the most accurate approach, achieving the highest performance when random over-sampling was applied before data splitting.

The comparative analysis revealed significant insights into different machine learning models. The XGBoost classifier stood out with the highest accuracy of 94.09%, showcasing the potential of advanced statistical methods in healthcare prediction. By comparing various preprocessing strategies and model approaches, the research provided a comprehensive evaluation of predictive techniques.

The web platform developed as part of this project represents a significant step towards personalized healthcare. By integrating multiple predictive models – metabolic syndrome, breast cancer detection, and stroke prediction – the system offers a unique approach to health risk assessment. The email-based result delivery ensures user privacy while providing accessible and meaningful health insights.

This research contributes to the evolving field of predictive healthcare, demonstrating how machine learning can transform complex medical data into practical, actionable information. The project not only provides a robust method for metabolic syndrome prediction but also establishes a framework for developing similar predictive health models.

Future research could explore expanding the dataset, incorporating additional health parameters, and further refining the machine learning models to enhance prediction accuracy and generalizability.