Project 2

Octavio Villalaz

10/17/2024

CSE-632

The goal of this project is to use a dataset of health and lifestyle characteristics to create prediction models for diabetes detection. Three machine learning methods were put into practice and assessed using metrics like accuracy, F1-score, and AUC-ROC: Random Forest, Support Vector Machine (SVM), and Logistic Regression. Following extensive preprocessing of the data and a comparison of the models, Logistic Regression showed the best performance with the highest AUC-ROC, suggesting that it is the best model for this dataset. The findings emphasize how crucial it is to use the best model and preprocessing methods in order to guarantee accurate predictions in healthcare applications.

Diabetes Prediction Using Machine Learning

The dataset used for this project contains health and lifestyle-related attributes, including features such as BMI, Blood Pressure, Physical Activity, Smoking Status, and demographic information like Age, Education, and Income. It contain 23 columns, and the goal variable is "Diabetes," which is a binary response that indicates whether or not a person has the disease. The dataset shows a little imbalance across classes, and in order to assure dependable model performance, handling missing values and encoding categorical variables needs to be done.

Several software tools were used to ensure efficient data analysis, model building, and performance evaluation. Python served as the primary programming language due to its versatility and rich ecosystem for data science. Pandas was employed for data manipulation, handling missing values, and feature engineering. Scikit-Learn provided essential tools for implementing and evaluating machine learning models, including Logistic Regression, Random Forest, and SVM, as well as for preprocessing tasks like scaling and encoding. For visualizing ROC curves and comparing model performances, Matplotlib was used. All development and experimentation were conducted in Jupyter Notebook, enabling quick debugging and iterative testing.

In order to ensure that the dataset is suitable for modeling first we must handle missing values, encode categorical variables, and scale numerical features. To handle the missing values first the dataset is separated between numerical values and categorical values. For example, BMI, GeneralHealthRating and NumberOfBadMentalHealthDays are all numerical values and the mean is used when imputing the data. BloodPressure, Cholesterol and CholesterolCheck are all categorical values which are imputed using the mode. The results of this preprocessing teqnique show that none of the columns with numerical values where missing any values, the categorical colums did contain missing values. The categorical columns with the highest values missing where Cholesterol, Stroke and BloodPressure but after preprocessing the all had 0 missing values.

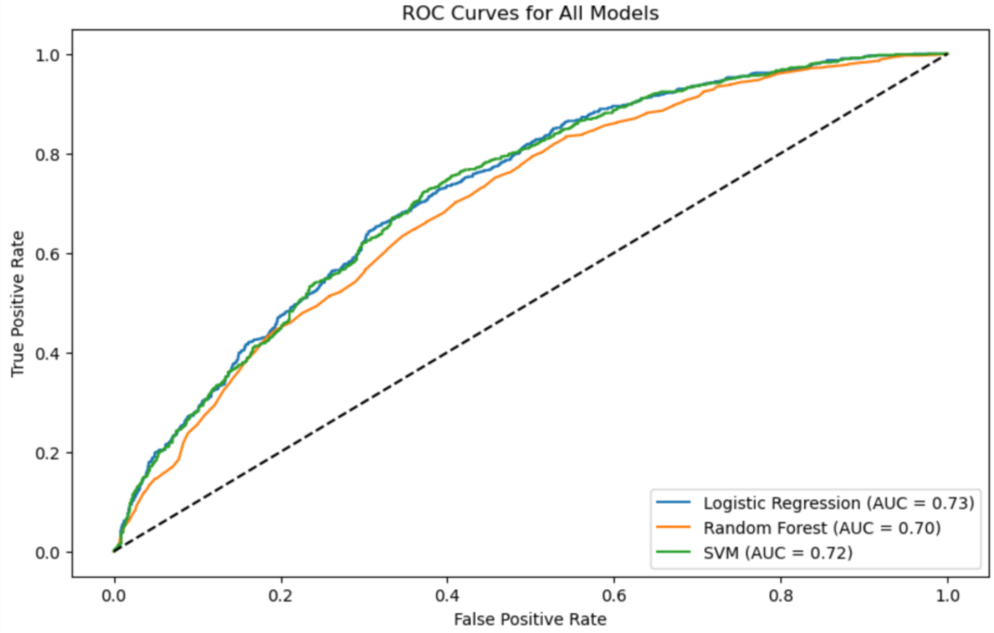
After removing all missing values the categorical values are then encoded in order to normalize the values into numerical data. One-hot encoding is applied to nominal data which includes columns like Sex and FruitInDiet, and ordinal encoding is applied to columns like Education and GeneralHealthRating where natural order exists.

Models like SVM and KNN are sensitive to feature scales, making it necessary to standardize or normalize features. This technique helps the machine learning algorithm to improve its performance by ensuring the features have similar scales. After applying Scaling, values will have summary statistics where the mean are approximately 0, and the standard deviation are approximately 1.

After processing the dataset, it is then trained using three predictive models Logistic Regression, Random Forest Classifier, and Support Vector Machine (SVM). Logistic Regression is a linear model used for binary classification. Instead of predicting continuous values, it estimates the probability that a given input belongs to a particular class. It uses the logistic (sigmoid) function to convert linear outputs into probabilities between 0 and 1. Key parameters include the regularization strength (C), Optimization algorithm, Penalty that is a type of regularization, and Tuning Strategy.

Random Forest Classifier is an ensemble model that builds multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting. For each input, every tree makes a prediction, and the forest selects the class with the majority vote. Key parameters include Number of decision trees in the forest, max_depth, max_features and Tuning Strategy.

Support Vector Machine is a supervised learning algorithm that finds the optimal hyperplane to separate data points from different classes in a high-dimensional space. The algorithm maximizes the margin between the closest data points from each class. Key parameters include the regularization strength (C), Kernel, gamma and Tuning Strategy.



The ROC curve comparison shows the performance of Logistic Regression, SVM, and Random Forest in predicting diabetes. Logistic Regression achieved the highest AUC (0.73), indicating it is the most effective at distinguishing between positive and negative cases. SVM followed closely with an AUC of 0.72, performing almost as well as Logistic Regression. Random Forest had the lowest AUC (0.70), suggesting it struggled slightly, possibly due to overfitting or ineffective feature selection. All models performed better than random guessing (represented by the diagonal line). Based on the ROC curves and AUC values, Logistic Regression emerges as the best model for this dataset, balancing simplicity and performance.