REPORT ON THE IMPLEMENTATION OF CLASSIFIERS FOR DIABETES PREDICTION

Principles of Data Mining and Machine Learning (2022 MOD007892)

SID: 2179798

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

Diabetes is a serious and chronic health condition affecting millions of people worldwide. Early detection and timely management of diabetes can significantly improve the quality of life of patients and prevent long-term complications. In this context, machine learning algorithms can serve as powerful tools for predicting the likelihood of diabetes in individuals. The diabetes dataset used in this study contains 768 instances and 9 attributes, including the outcome variable indicating whether the individual has diabetes or not This project aims to identify the best model among the implemented models to predict whether a patient is prone to diabetes within the next five years. The following models were implemented:

1. Logistic Regression
2. K- Nearest Neighbour
3. Support Vector Machine
4. Decision Tree
5. Random Forest

To evaluate and compare the models, several metrics extracted from the confusion matrix of each model will be used.

The confusion matrix is a table that summarizes the performance of a classification model by comparing the actual and predicted values of the target variable. From the confusion matrix, the following metrics were computed: Accuracy, Precision, Recall/Sensitivity, F1-score and Specificity.

EXPLORATORY DATA ANALYSIS (EDA)

The dataset contains various features such as age, body mass index (BMI), blood pressure, and glucose levels, among others. At this phase, the dataset was checked for duplicates, missing entries and incorrect dtypes. Some of the key EDA steps performed on the dataset include:

Descriptive Statistics: Descriptive statistics, such as mean, median, standard deviation, and quartiles, were computed for each feature in the dataset. This helps to summarize the data and identify any outliers or unusual observations.

Data Visualization: Various data visualization techniques, such as stacked bar charts, and heatmaps, were used to visualize the distribution of the features and identify any patterns or correlations.

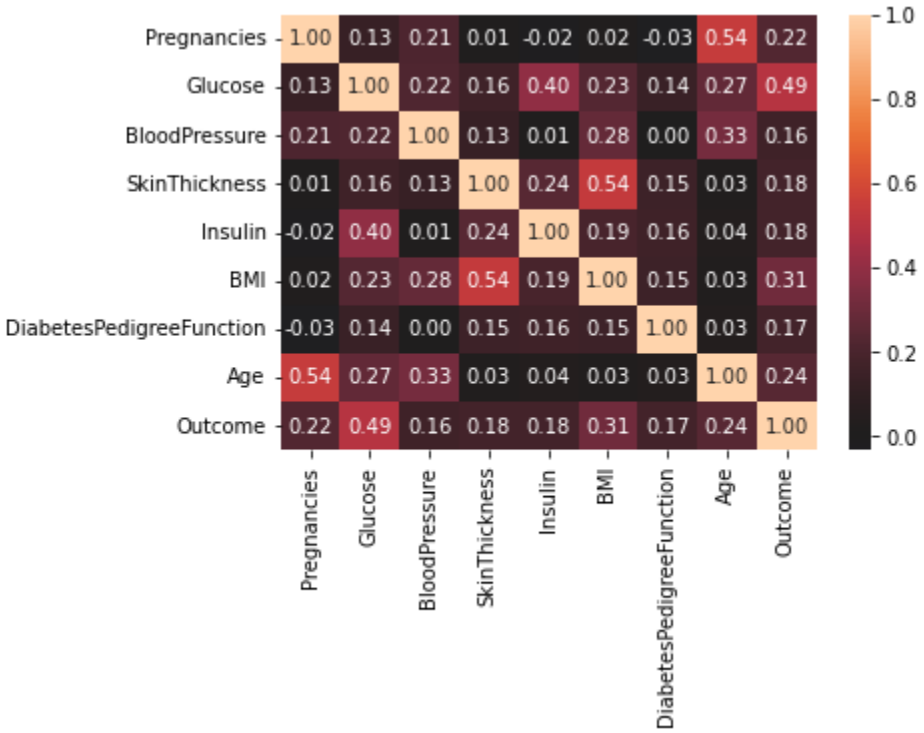

Figure 1: Correlation Matrix of All Variables

Overall, the EDA phase played a crucial role in understanding the dataset and preparing it for the machine learning models.

MODEL IMPLEMENTATION AND TRAINING

The six classification algorithms were implemented using the scikit-learn library in Python. The dataset was split into training and testing sets in an 80:20 ratio. Since the variables are numeric and on different scales, they were standardized using StandardScaler. Each model was then trained with a copy of the original training set and the confusion matrix was examined.

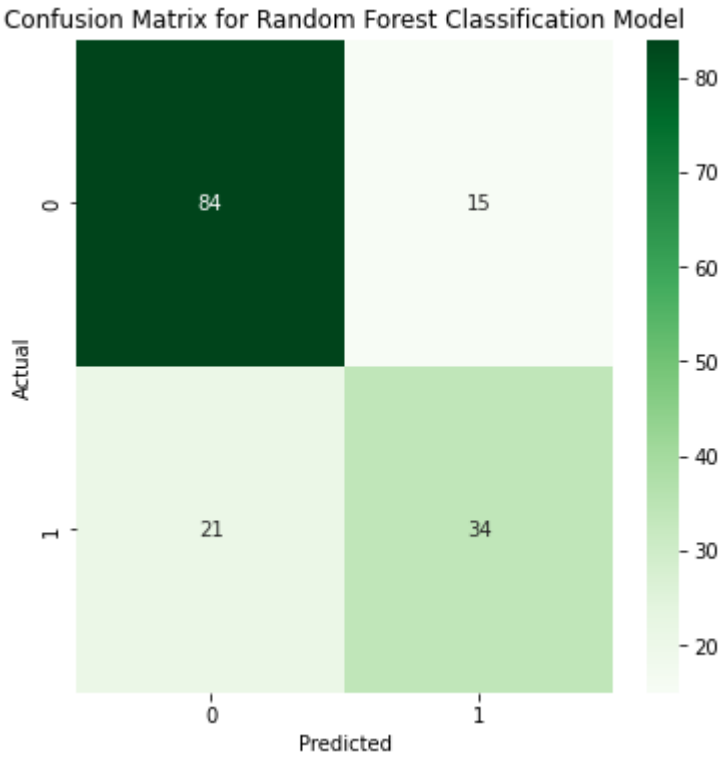

Figure 2: Confusion Matrix for Random Forest Classifier

The Logistic Regression algorithm achieved an accuracy of 77.3%, with a precision of 70.8%, recall/sensitivity of 61.8%, F1-score of 66%, and specificity of 85.9%.

The KNN algorithm achieved an accuracy of 70%, with a precision of 57.9%, recall/sensitivity of 60.0%, F1-score of 58.9%, and specificity of 75.8%.

The SVM algorithm achieved an accuracy of 77.3%, with a precision of 70.8%, recall/sensitivity of 61.8%, F1-score of 63.9%, and specificity of 85.9%.

The Decision Tree algorithm achieved an accuracy of 71.4%, with a precision of 60.8%, recall/sensitivity of 56.4%, F1-score of 58.5%, and specificity of 79.8%.

The Random Forest algorithm achieved an accuracy of 76.6%, with a precision of 69.4%, recall/sensitivity of 61.8%, F1-score of 65.4%, and specificity of 84.8%.

EVALUATION

Evaluation metrics are essential in assessing the performance of machine learning models. In classification tasks, there are several metrics used to evaluate the performance of a model. However, it's important to note that a single metric alone is not always sufficient to evaluate the performance of a model, and a combination of metrics should be used to gain a more comprehensive understanding of a model's performance.

Accuracy: This metric measures the overall correctness of the predictions. It is the ratio of the number of correct predictions to the total number of predictions. However, accuracy alone is not always a reliable measure of model performance, especially in cases of imbalanced classes.

Precision: Precision measures the proportion of true positives (correctly identified positives) out of all predicted positives. It helps to assess the model's ability to minimize false positives, which are cases where the model predicted a positive outcome when the actual outcome was negative. High precision indicates that the model is good at identifying only true positives and has few false positives.

Recall (or Sensitivity): Recall measures the proportion of true positives out of all actual positives. It helps to assess the model's ability to minimize false negatives, which are cases where the model predicted a negative outcome when the actual outcome was positive. High recall indicates that the model is good at identifying most of the true positives and has few false negatives.

Specificity: Specificity measures the proportion of true negatives (correctly identified negatives) out of all actual negatives. It helps to assess the model's ability to minimize false positives. High specificity indicates that the model is good at identifying only true negatives and has few false positives.

F1-score: The F1-score is the harmonic mean of precision and recall. It is a measure of the model's accuracy that considers both precision and recall. A high F1-score indicates that the model has both high precision and high recall.

By evaluating these metrics, we can get a better understanding of the model's performance and determine which model is best suited for the given problem.

Out[69]:

	Models	Sensitivity/Recall	Specificity	Precision	Accuracy	F1_Score
0	Logistic Regression	0.618182	0.858586	0.708333	0.772727	0.660194
1	KNN	0.600000	0.757576	0.578947	0.701299	0.589286
2	SVM	0.618182	0.858586	0.708333	0.772727	0.660194
3	Decision Tree	0.563636	0.797980	0.607843	0.714286	0.584906
4	Random Forest	0.618182	0.848485	0.693878	0.766234	0.653846

Figure 3: Evaluation Metrics for the Machine Learning Algorithm

The results showed that all five algorithms could predict diabetes with varying levels of accuracy and precision. Among them, the Logistic Regression algorithm performed the best as it has the best values across all metrics particularly the sensitivity and specificity metrics.