DEPARTMENT OF ELECTRICAL AND INFORMATION ENGINEERING

FACULTY OF ENGINEERING

UNIVERSITY OF RUHUNA

A PROJECT REPORT SUBMITTED IN PARTIAL FULFILMENT OF THE DEGREE OF THE BACHELOR OF THE SCIENCE OF ENGINEERING HONOURS

18th APRIL 2024

Predicting Diabetes Diagnosis Using Machine Learning

Group 12

Kheminda D.A.J.P (EG/2019/3636)
Kumarasinghe K.A.G.T.V (EG/2019/3645)

Contents

List of Figures

Acronyms

ML	-	Machine Learning
SVC	-	Support Vector Classifier
RF	-	Random Forest
ML	-	Machine Learning

Chapter 1

Predicting Diabetes Diagnosis Using Machine Learning

1.1 Introduction

Diabetes is a long-term, chronic metabolic disease marked by increased blood sugar levels.It poses a significant global health challenge, with its prevalence steadily increasing [1]. Early diagnosis and intervention are crucial in managing diabetes and preventing its complications. In this project, we want to use machine learning techniques to create a diabetes diagnosis prediction model using multiple clinical measurements.

The primary objective of this project is to construct a robust machine learning model capable of accurately predicting diabetes diagnosis in patients. By analyzing a comprehensive dataset comprising diverse medical features, including glucose levels, blood pressure, insulin levels, and body mass index (BMI) and etc. We aim to develop a predictive tool that can assist healthcare professionals in identifying individuals at risk of diabetes at an early stage.

1.2 Methodology

The methodology for this project involves the utilization of two machine learning algorithms. Such as Support Vector Classifier (SVC) and Random Forest. These algorithms will be trained on the dataset to learn the underlying patterns and relationships between the predictor variables and the target variable.

1. **Support Vector Classifier (SVC)**

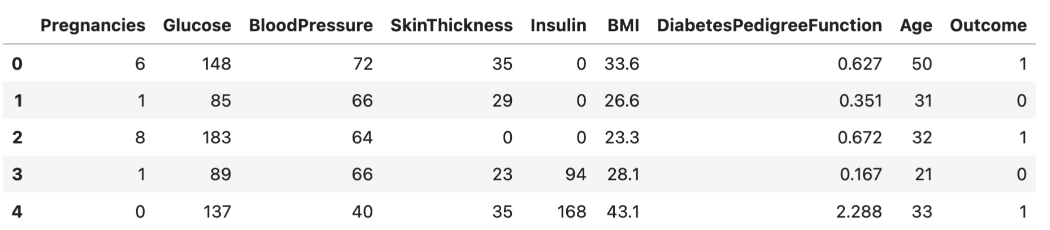
   Support Vector Classifier (SVC) is a powerful classification algorithm that works by finding the hyperplane that best separates the classes in the feature space. It aims to maximize the margin between classes, thereby improving generalization performance. By identifying the optimal hyperplane, SVC can effectively classify new data points based on their features.

2. **Random Forest**

   Random Forest is an ensemble learning technique that employs multiple decision trees to make predictions. Each decision tree in the forest is trained on a subset of the dataset, and predictions are made by aggregating the results from all trees. This ensemble approach helps mitigate over-fitting and improves the robustness of the model. Random Forest is one of the best algorithm for handling high dimensional and capture the complex relationships between the variable.

1.2.1 Dataset

This dataset is originally from the National Institute of Diabetes and Digestive Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset [2]. Figure 1.1 show the first 5 rows of the dataset.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 1.1: Diabetes dataset with first 5 rows

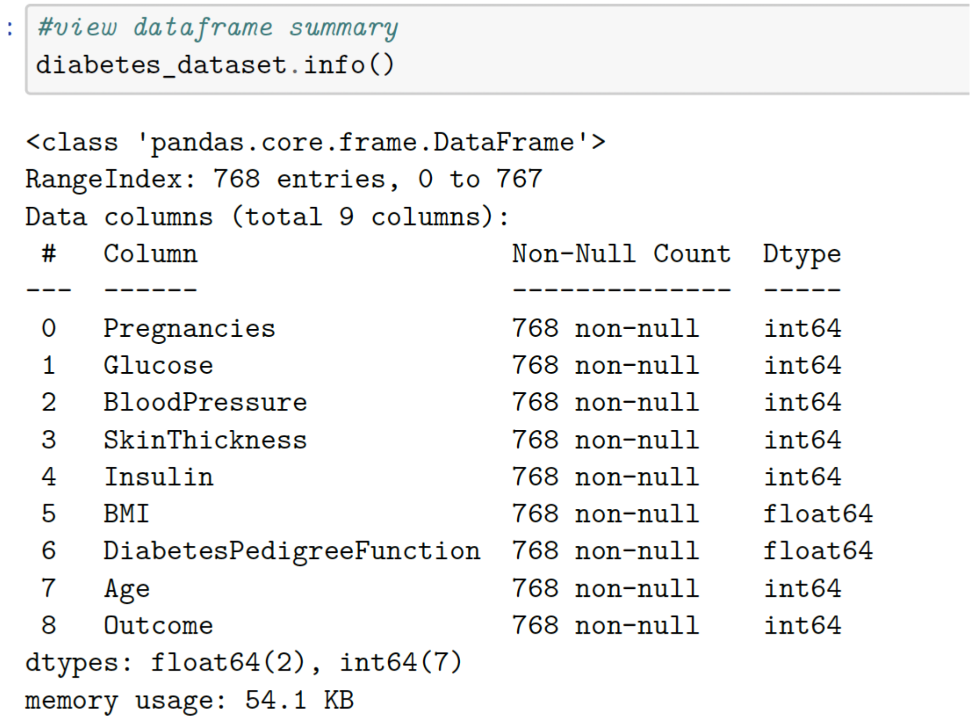
Meaning of the each features in the dataset.

1. **Pregnancies** - To express the Number of pregnancies
2. **Glucose** - To express the Glucose level in blood
3. **BloodPressure** - To express the Blood pressure measurement
4. **SkinThickness** - To express the thickness of the skin
5. **Insulin** - To express the Insulin level in blood
6. **BMI** - To express the Body mass index
7. **DiabetesPedigreeFunction** - To express the Diabetes percentage
8. **Age** - To express the age
9. **Outcome** - To express the final result 1 is Yes and 0 is No

1.2.2 Preprocessing

1. **Handling Null / Missing Values**

   Missing values are a common occurrence in real-world datasets and can significantly affect the performance of machine learning models if not addressed properly. In this project, there was no any null or missing values in dataset. It can conform by checking figure 1.2.



```
#view dataframe summary
diabetes_dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Figure 1.2: Handling null/missing values

2. **Handling Duplicated Values**

   Duplicated values are instances where one or more observations in a dataset are identical to others. In this project, we will implement strategies to identify duplicated values. According to the figure 1.3 there was no duplicated values in each columns.

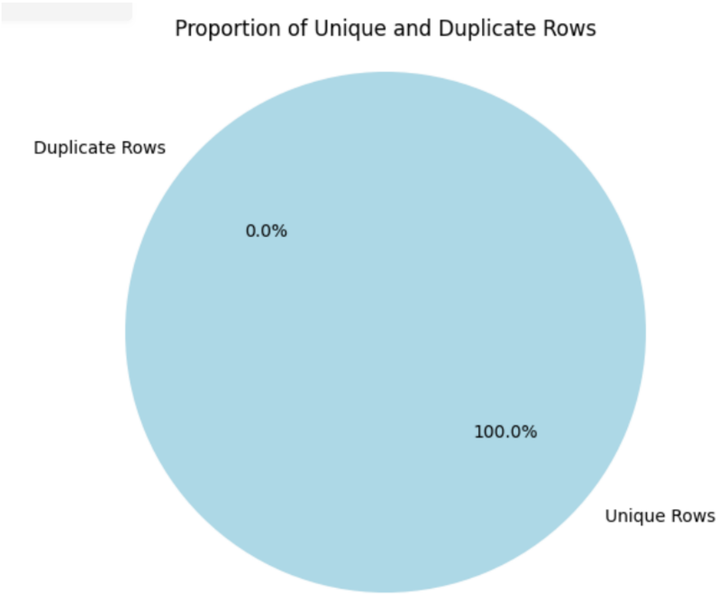


Figure 1.3: Duplicated values graph

3. **Handling Outliers**

   Outliers are data points that significantly deviate from the rest of the dataset. In our analysis to predict diabetes, we have chosen not to treat outliers in the dataset. Since outliers [figure 1.4] may contain valuable information relevant to the prediction task, such as DiabetesPedigreeFunction or insulin levels that could be indicative of diabetes, we have opted to retain them in our analysis.

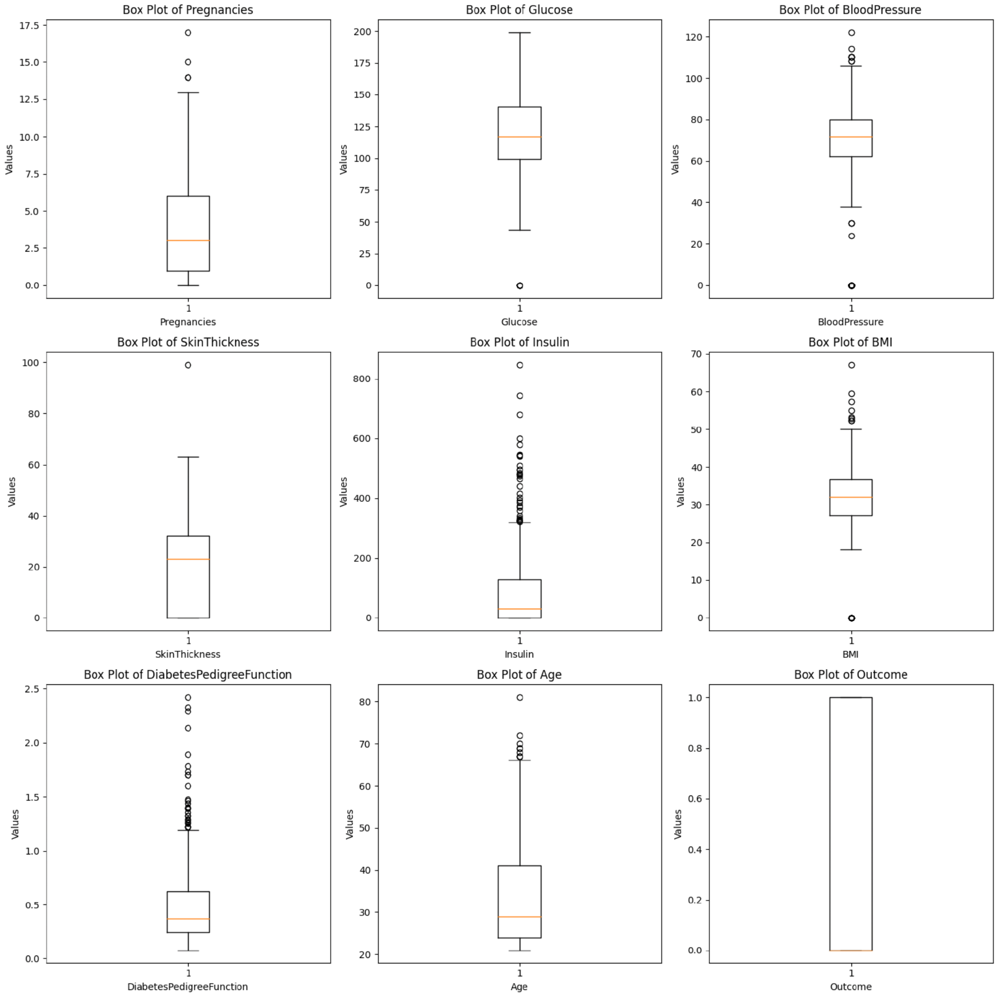
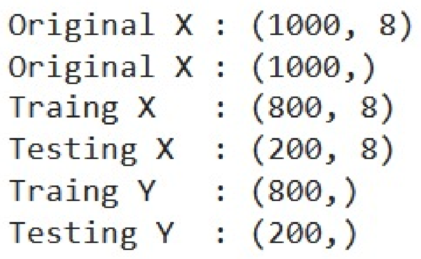


Figure 1.4: Box-plots for identify the outliers in the dataset

4. **Train Test Split**

   We will partition the dataset into two subsets, 80% for training set and 20% for test set [figure 1.5]. The training set will be used to train the machine learning models, while the test set will be used to evaluate their performance on unseen data.

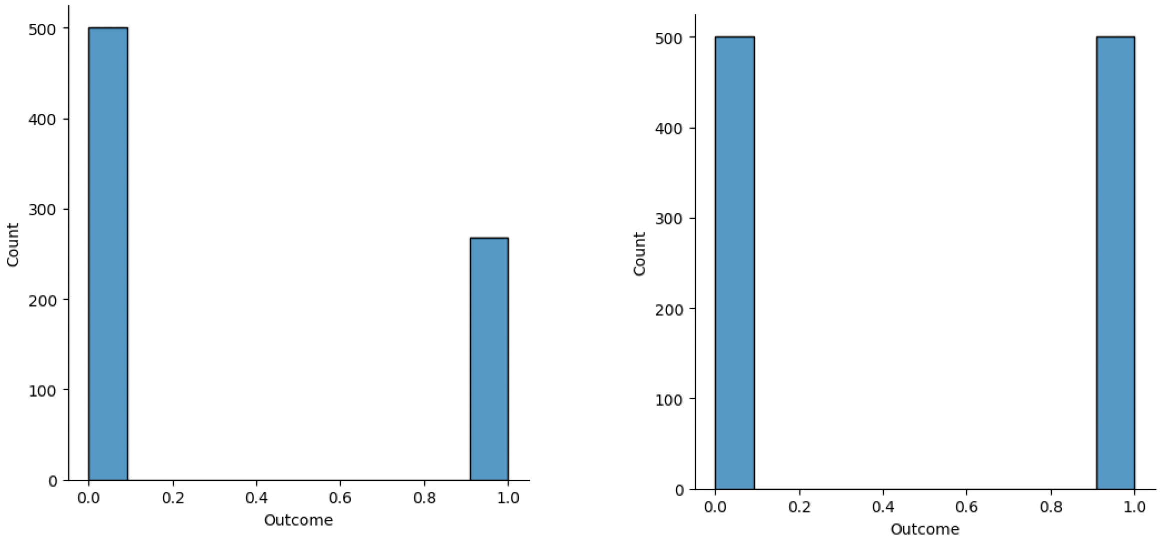


```
Original X : (1000, 8)
Original X : (1000,)
Traing X   : (800, 8)
Testing X  : (200, 8)
Traing Y   : (800,)
Testing Y  : (200,)
```

Figure 1.5: Split dataset into two part

5. **Treat for Imbalance dataset**

   By employing SMOTE, we aim to mitigate the effects of class imbalance and develop machine learning models that are more effective in accurately predicting diabetes diagnosis across both classes. This ensures that the models are not biased towards the majority class and can provide reliable predictions for both positive and negative instances. Figure 1.9 show the dataset after and before treating.



(a) Imbalanced label

(b) Balanced label

Figure 1.6: dataset after and before treating

6. **Correlation Matrix**

   In the context of this project, a correlation matrix heatmap provides a visually intuitive way to explore the relationships between features and the target variable. By observing the colors in the heatmap [figure 1.7], we can quickly identify which features are strongly positively or negatively correlated with the target variable. This allows for easy identification of potentially important features for predicting diabetes diagnosis.We observed that all features in our dataset exhibit some degree of relationship with the target variable. Therefore, we have made the decision to retain all features for training our model.

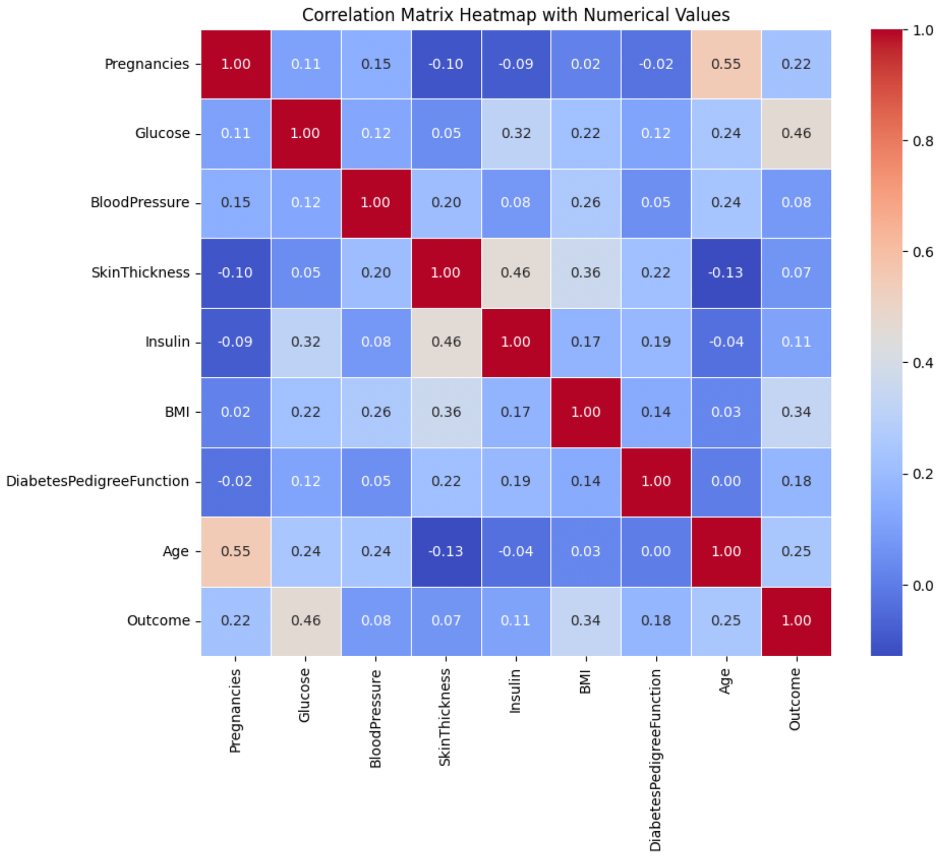
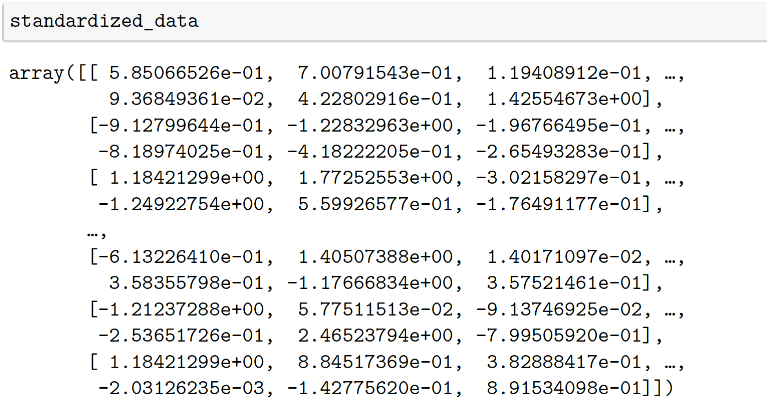


Figure 1.7: Correlation Matrix Heatmap with Numerical Values

7. **Data Standardization**

   Data standardization is a preprocessing technique used to rescale the features of a dataset to have a common scale or distribution. This process ensures that all features contribute equally to the analysis and prevents features with larger scales from dominating the modeling process. Figure 1.8 show the sample of standardized data in our project.



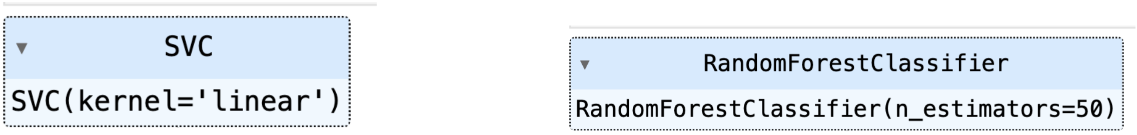
```
standardized_data
```

```
array([[ 5.85066526e-01,  7.00791543e-01,  1.19408912e-01, ...,
          9.36849361e-02,  4.22802916e-01,  1.42554673e+00],
       [-9.12799644e-01, -1.22832963e+00, -1.96766495e-01, ...,
         -8.18974025e-01, -4.18222205e-01, -2.65493283e-01],
       [ 1.18421299e+00,  1.77252553e+00, -3.02158297e-01, ...,
         -1.24922754e+00,  5.59926577e-01, -1.76491177e-01],
       ...,
       [-6.13226410e-01,  1.40507388e+00,  1.40171097e-02, ...,
          3.58355798e-01, -1.17666834e+00,  3.57521461e-01],
       [-1.21237288e+00,  5.77511513e-02, -9.13746925e-02, ...,
         -2.53651726e-01,  2.46523794e+00, -7.99505920e-01],
       [ 1.18421299e+00,  8.84517369e-01,  3.82888417e-01, ...,
         -2.03126235e-03, -1.42775620e-01,  8.91534098e-01]])
```

Figure 1.8: Sample of standardized data

1.2.3 Training the Model

In this phase of the project, we train machine learning models using preprocessed data with two different algorithms. Support Vector Classifier (SVC) and Random Forest algorithm are the used algorithms for our project.



```
▾ SVC
SVC(kernel='linear')
```

```
▾ RandomForestClassifier
RandomForestClassifier(n_estimators=50)
```

(a) Model training using SVC algorithm

(b) Model training using RF algorithm

Figure 1.9: Training the model using SVC and RF

1.3 Results

Here we present the final results obtained from Support Vector Classification and Random Forest Classification methods.In comparing the two models, we've utilized evaluation metrics including the confusion matrix, classification report, and accuracy score to comprehensively evaluate their performance.

1.3.1 Performance Metrics Summary

Comparison of Confusion Matrix

When comparing the confusion matrices of Support Vector Classification (SVC) and Random Forest (RF), we notice a few differences.The SVC model demonstrates a slightly higher number of false negatives compared to RF, potentially indicating a higher sensitivity towards identifying non-diabetic cases. Meanwhile, RF exhibits a lower false negative rate, suggesting a stronger capability in correctly identifying diabetic cases. Both models maintain comparable true positive and true negative rates, underlining their effectiveness in predicting both diabetic and non-diabetic cases accurately.

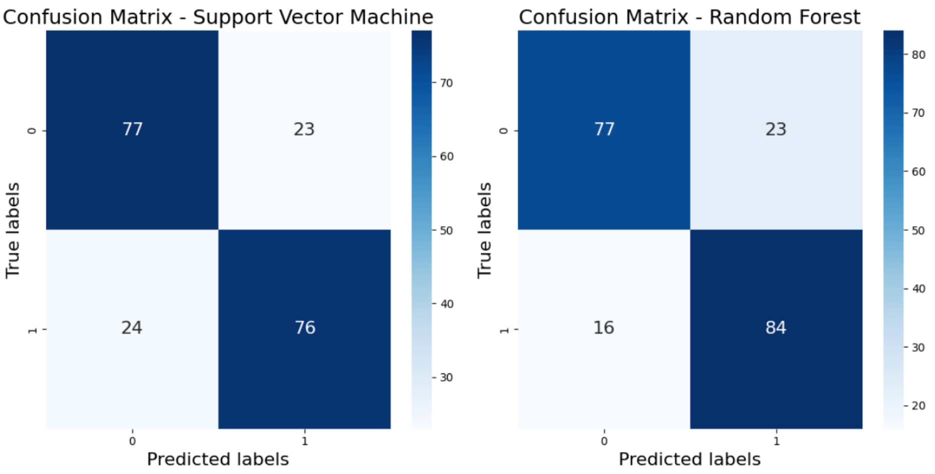
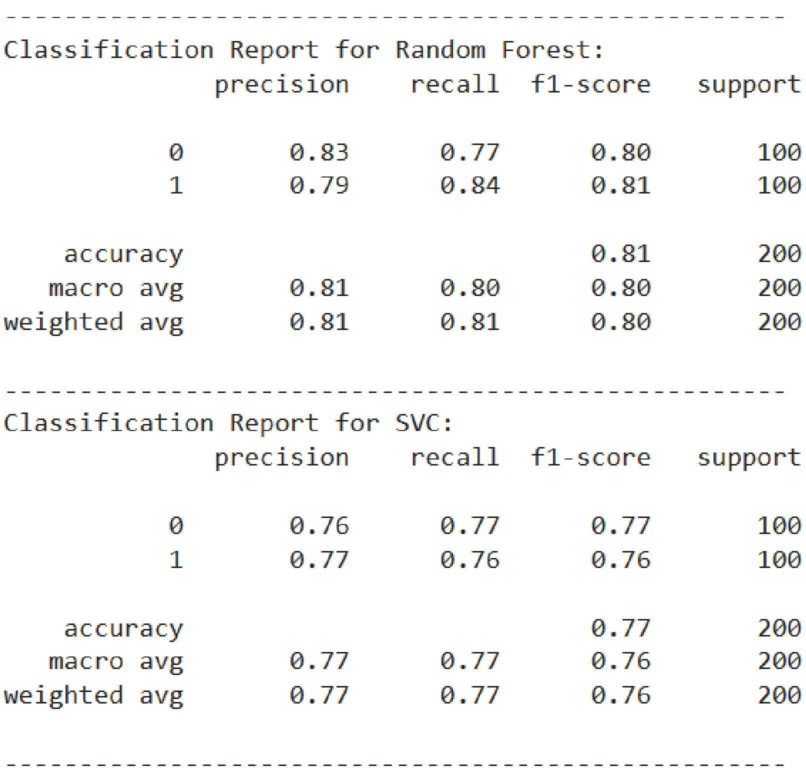


Figure 1.10: Comparative Confusion Matrices: SVC vs. Random Forest

Comparision of Classification Report

The classification reports for Random Forest (RF) and Support Vector Classification (SVC) models reveal some interesting insights. RF shows slightly higher precision and recall for class 1 (diabetic) compared to SVC, indicating that RF is better at correctly identifying diabetic cases. On the other hand, SVC demonstrates marginally higher precision and recall for class 0 (non-diabetic) compared to RF, suggesting that SVC is slightly better at correctly identifying non-diabetic cases. Overall, both models exhibit comparable performance in terms of F1-score and support, with RF leaning towards better performance in diabetic identification and SVC showing a slight advantage in non-diabetic identification.



```
--------------------------------------------------
Classification Report for Random Forest:
              precision    recall  f1-score   support

           0       0.83      0.77      0.80       100
           1       0.79      0.84      0.81       100

    accuracy                           0.81       200
   macro avg       0.81      0.80      0.80       200
weighted avg       0.81      0.81      0.80       200

--------------------------------------------------
Classification Report for SVC:
              precision    recall  f1-score   support

           0       0.76      0.77      0.77       100
           1       0.77      0.76      0.76       100

    accuracy                           0.77       200
   macro avg       0.77      0.77      0.76       200
weighted avg       0.77      0.77      0.76       200

--------------------------------------------------
```

Figure 1.11: Comparative of Classification Reort: SVC vs. Random Forest

1.4 Discussion and Conclusion

In our project we train and evaluated the SVC and Random Forest models to detect whether patient have diabetes or not.

our conclusion can be classify as following,

- **Pre Process**

  Diabetes Dataset had 768 data rows and 9 features. we used some pre-processing technique to treat the dataset. Such as handling missing values and duplicated values, handling outliers, generate balanced dataset and etc. Finnaly we genered good dataset for traing our two models.

- **Model Performance**

  We have utilized evaluation metrics including the confusion matrix, classification report, and accuracy score to comprehensively evaluate their performance. RF shows slightly higher precision and recall for class 1 (diabetic) compared to SVC, indicating that RF is better at correctly identifying diabetic cases. On the other hand, SVC demonstrates marginally higher precision and recall for class 0 (non-diabetic) compared to RF, suggesting that SVC is slightly better at correctly identifying non-diabetic cases.

- **Model Comparison**

  The SVC model demonstrates a slightly higher number of false negatives compared to RF, potentially indicating a higher sensitivity towards identifying non-diabetic cases. Meanwhile, RF exhibits a lower false negative rate, suggesting a stronger capability in correctly identifying diabetic cases. Both models maintain comparable true positive and true negative rates, underlining their effectiveness in predicting both diabetic and non-diabetic cases accurately.

Finally, this project demonstrates the potential of machine learning in healthcare, particularly in the field of medical diagnosis. By using diagnostic and accurate ML model, we can improve the health of the patient and contribute to the advancement of medical science.

References

[1] . WHO, “Diabetes,” 2024. Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar), which leads over time to serious damage to the heart, blood vessels, eyes, kidneys and nerves.

[2] A. D. KHARE, “Diabetes dataset,” 2022. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.