Detailed Project Report on Predictive Models for Diabetes Detection

Introduction

The objective of this project is to develop and evaluate machine learning models that can predict the occurrence of diabetes based on health-related data. The models include a Random Forest Classifier, Support Vector Machine (SVM), and an Artificial Neural Network (ANN). This report provides an in-depth analysis of each model's performance based on accuracy, precision, recall, and F1-score. Additionally, confusion matrices are used to visually assess the performance.

Dataset Overview

The dataset comprises several health-related features such as age, gender, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level, which are utilized to predict the binary outcome (**diabetes** or **no diabetes**). The data was split into training (80%) and testing (20%) sets to ensure a robust evaluation of the models.

Models Overview

Three machine learning models were implemented and evaluated:

1. **Random Forest Classifier**: A robust ensemble method that uses multiple decision trees to make predictions.
2. **Support Vector Machine (SVM)**: A powerful classifier that finds the optimal hyperplane to separate different classes in the feature space.
3. **Artificial Neural Network (ANN)**: A deep learning model configured with dense layers to capture complex patterns in the data.

Model Evaluation Metrics

The performance of the models was evaluated based on the following metrics:

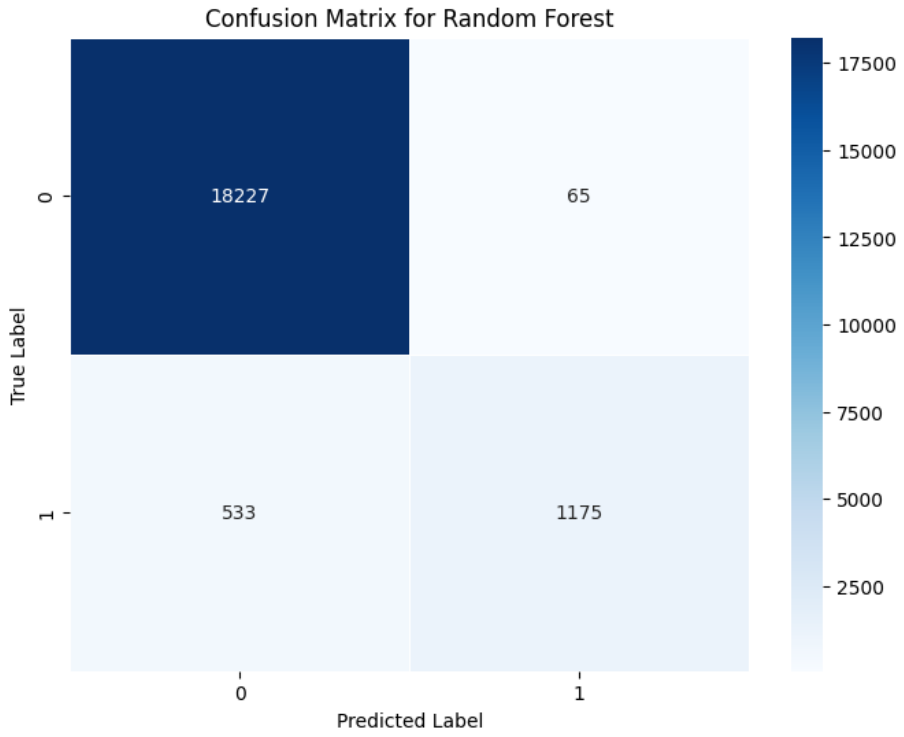
- **Accuracy**: The fraction of predictions our model got right.
- **Precision**: The number of true positive results divided by the number of all positive results.
- **Recall**: The number of true positive results divided by the number of positives that should have been retrieved.
- **F1-Score**: The harmonic mean of precision and recall.

Results and Discussion

Random Forest Classifier

- **Accuracy**: 97.01%
- **Precision for class 1**: 95%

- **Recall for class 1**: 69%
- **F1-Score for class 1**: 80%
- **Confusion Matrix**:
    - True Negative (TN): 18227
    - False Positive (FP): 65
    - False Negative (FN): 533
    - True Positive (TP): 1175

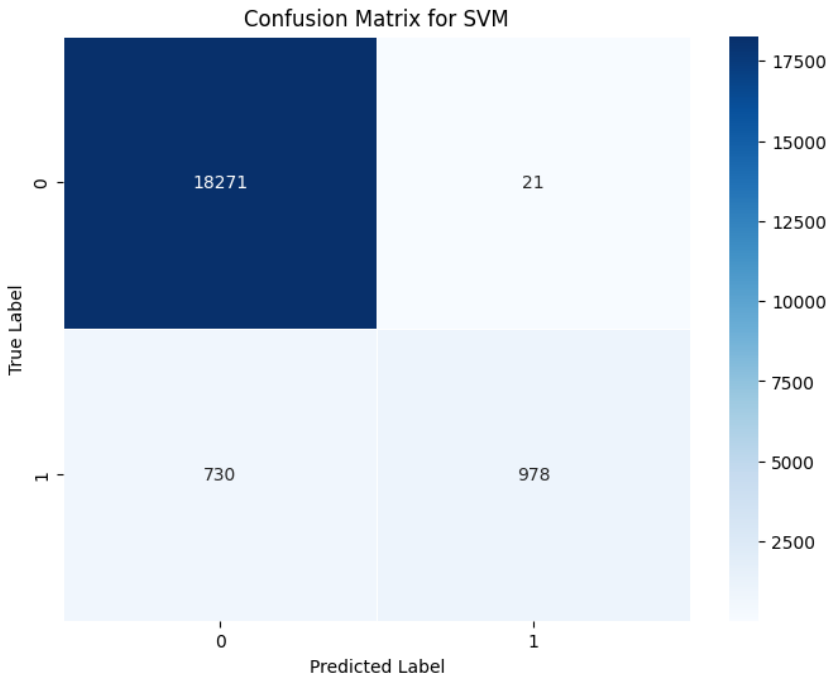


This Figure shows confusion matrix for Random Forest.

Support Vector Machine (SVM)

- **Accuracy**: 96.245%
- **Precision for class 1**: 98%
- **Recall for class 1**: 57%
- **F1-Score for class 1**: 72%
- **Confusion Matrix**:
    - TN: 18271
    - FP: 21

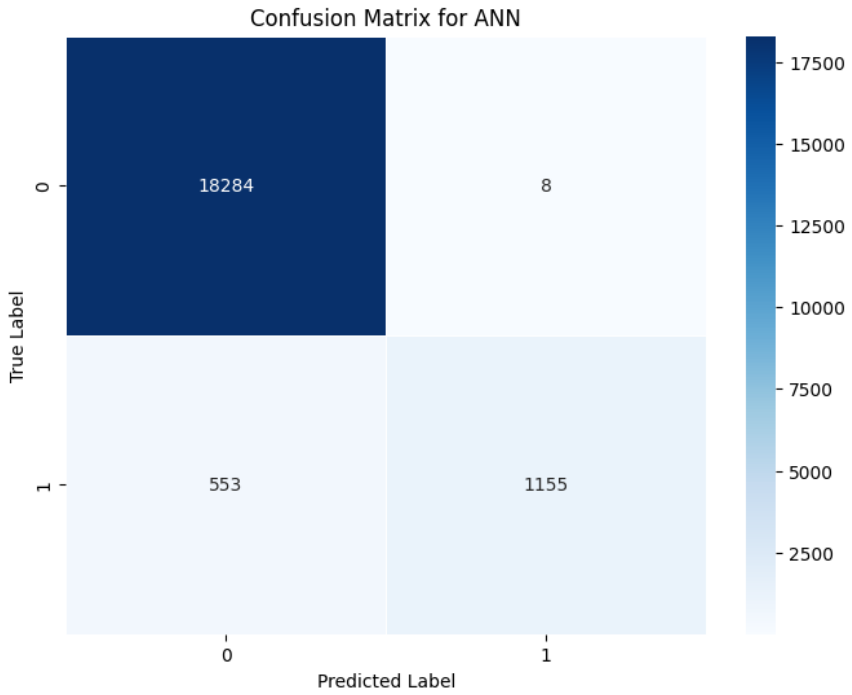
- FN: 730
- TP: 978



This Figure shows confusion matrix for SVM.

Artificial Neural Network (ANN)

- **Accuracy**: 97.195%
- **Precision for class 1**: 99%
- **Recall for class 1**: 68%
- **F1-Score for class 1**: 80%
- **Confusion Matrix**:
  - TN: 18284
  - FP: 8
  - FN: 553
  - TP: 1155



This Figure shows confusion matrix for Artificial Neural Network.

Conclusion

All three models demonstrated high accuracy with the ANN and Random Forest showing similar performance and slightly outperforming the SVM in terms of recall for the positive class. Thus, the problem is evaluated using three models. Hence, the models performed robust.