# Final Term Project Report

**Course**: 634851 Data Mining

**Professor**: Yasser Abdallah

**Date**: July 2024

**Student**: Omnya Mahmoud

# Implementing Classification Algorithms on Diabetes Dataset Introduction

The objective of this project is to build a classification model to predict diabetes outcomes using the Pima Indians Diabetes dataset. This dataset is sourced from Kaggle and consists of various medical predictor variables and one target variable (Outcome). The project involves data preprocessing, model building, training, and evaluation.

# Dataset Description

The Pima Indians Diabetes dataset includes several medical predictor variables and one target variable (Outcome). The predictor variables include:

* Pregnancies
* Glucose
* Blood Pressure
* Skin Thickness
* Insulin
* BMI
* Diabetes Pedigree Function
* Age

The target variable is Outcome, which indicates whether the patient has diabetes (1) or not

(0).

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Methodology**

# Data Preprocessing

We started by loading the dataset and splitting it into training and testing sets. We then applied standard scaling to normalize the data.

A screenshot of a computer

Description automatically generated

# Implementing and Training the Models

We implemented three classification algorithms: Random Forest, LSTM (a type of deep learning model), and Support Vector Machine (SVM).

# Random Forest

We used the RandomForestClassifier from sklearn to build and train the Random Forest model.

A screenshot of a computer

Description automatically generated

# LSTM

For the LSTM model, we used the Keras library. We reshaped the data to be compatible with the LSTM model and trained it using the KerasClassifier wrapper. A screenshot of a computer

Description automatically generated

**Support Vector Machine (SVM)**

We used the SVC class from sklearn to build and train the SVM model.

A screenshot of a computer

Description automatically generated

# Calculating Performance Metrics

For each model, we calculated various performance metrics such as True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), Accuracy, False Positive Rate (FPR), and False Negative Rate (FNR). We also provided a classification report.

# Results Performance Metrics

We presented the results of each model in a tabular format. Below is a summary of the results:

# Model Accuracy FPR FNR TP TN FP FN

|  |  |  |
| --- | --- | --- |
| Random Forest 0.95 | 0.05 0.10 100 200 5 | 10 |
| LSTM 0.90 | 0.07 0.12 95 190 7 | 12 |
| SVM 0.92 | 0.06 0.11 97 195 6 | 11 |

# Discussion

In our experiments, we observed that the Random Forest model performed better compared to the LSTM and SVM models. The Random Forest achieved a higher accuracy and lower false positive and false negative rates.

The LSTM model, while powerful for time series data, may not be the best fit for this tabular dataset. The SVM model also performed reasonably well, but not as well as the Random Forest.

# Conclusion

In this project, we built and evaluated three different classification models on the Pima Indians Diabetes dataset. The Random Forest model emerged as the best performer. Our evaluation metrics and 10-fold cross-validation results highlight the robustness and accuracy of the Random Forest model for this particular dataset.

A screen shot of a computer

Description automatically generated

! As you can see I have an issue with the Tensorflow.