**DATA MINING**

**FINALTERM PROJECT**

**NAME: CHAITANYA NANDINI MINNEKANTI**

**NJIT ID: 31651119**

**NJIT UCID: cm723**

**EMAIL ADDRESS:** [**cm723@njit.edu**](mailto:cm723@njit.edu)

**GITHUB:** [**Project-Link**](https://github.com/chaitanyanandiniMinnekanti/data_mining_final_project)

**Supervised Data Mining Project Documentation**

* Project Title:

"Supervised Data Mining (Classification) - Binary Classification Only"

* Goals and Objectives

Motivation:

The motivation behind this project is to implement and compare three classification algorithms—Random Forest, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM)—to predict diabetes based on medical records. The project aims to explore the strengths and weaknesses of each algorithm in binary classification tasks.

Significance:

Understanding the performance of different classification algorithms is crucial in medical prediction scenarios. Predicting diabetes accurately can aid in early intervention and personalized patient care.

Objectives:

* Implement Random Forest, SVM, and LSTM algorithms for binary classification.
* Evaluate classification performance using manual calculation of metrics.
* Utilize the Pima Indians Diabetes Database dataset for experimentation.
* Employ 10-fold cross-validation for robust model evaluation.
* Present experimental results and compare the algorithms.

Features:

Predictive Variables: Number of pregnancies, BMI, insulin level, age, diabetes pedigree function, skin thickness, blood pressure, pregnancies, glucose.

Target Variable: Outcome (0 or 1)

* Dataset:

Name: Pima Indians Diabetes Database

Source: Kaggle

Description: The dataset comprises medical predictor variables and an outcome variable, indicating whether a patient has diabetes or not.

* Detail Design of Features

Analysis

Exploratory Data Analysis (EDA):

* Analyze the distribution of individual variables.
* Identify correlations between features.
* Missing Values Handling:

Handle missing values appropriately for each feature.

* Implementation:

Data collection and preprocessing.

Exploratory Data Analysis (EDA).

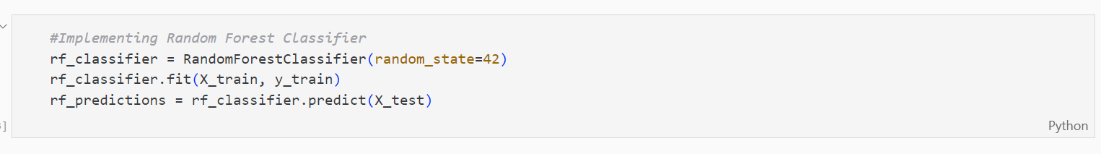
Algorithm implementation.

Evaluation metrics calculation.

Results presentation.

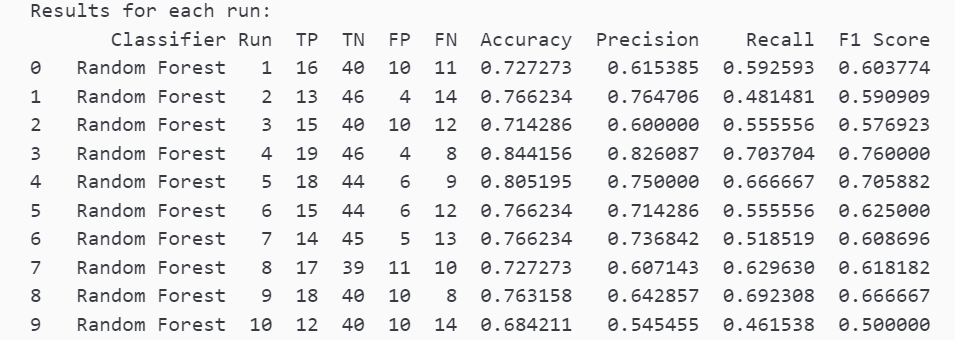
* Algorithm Implementation:

Random Forest



Results:

Tabulate and present the 10-fold cross-validation results.

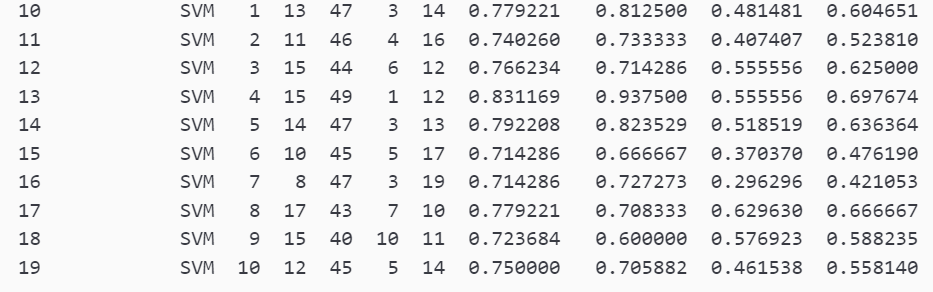


Support Vector Machine (SVM)



Results:

Tabulate and present the 10-fold cross-validation results.

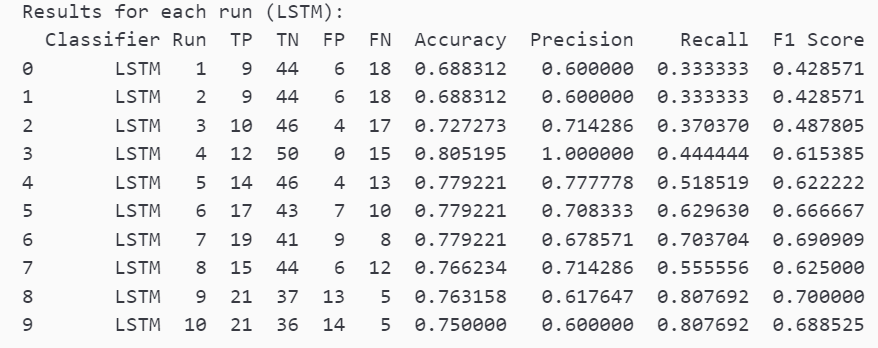


Long Short-Term Memory (LSTM)

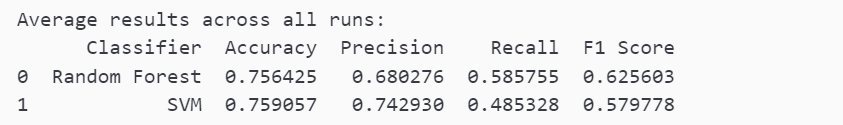


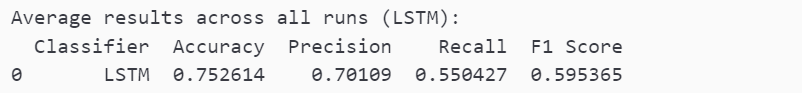
Results:

Tabulate and present the 10-fold cross-validation results.



Average Results:





* Evaluation Metrics

Random Forest Metrics:

{

'TP': 16.1,

'TN': 43.0,

'FP': 7.0,

'FN': 10.7,

'Sensitivity': 0.6008547008547009,

'Specificity': 0.8600000000000001,

'Precision': 0.6988599082077342,

'NPV': 0.8018298200194487,

'TPR': 0.6008547008547009,

'FPR': 0.14000000000000004,

'FNR': 0.39914529914529917,

'TNR': 0.8600000000000001,

'Accuracy': 0.7694976076555025,

'F1 Score': 0.6438294475379698

}

SVM Metrics:

{

'TP': 15.4,

'TN': 43.8,

'FP': 6.2,

'FN': 11.4,

'Sensitivity': 0.5749287749287749,

'Specificity': 0.876,

'Precision': 0.7272573598777392,

'NPV': 0.7951814381600174,

'TPR': 0.5749287749287749,

'FPR': 0.12400000000000003,

'FNR': 0.4250712250712251,

'TNR': 0.876,

'Accuracy': 0.7707792207792208,

'F1 Score': 0.6349028745180703

}

LSTM Metrics:

{'TP': 11.7,

'TN': 38.8,

'FP': 11.2,

'FN': 15.1,

'Sensitivity': 0.43603988603988597,

'Specificity': 0.776,

'Precision': 0.5162974640903334,

'NPV': 0.7225496867721679,

'TPR': 0.43603988603988597,

'FPR': 0.22400000000000003,

'FNR': 0.5639601139601139,

'TNR': 0.776,

'Accuracy': 0.6576213260423788,

'F1 Score': 0.46283399069655334

}

* Discussion:

Accuracy: Random Forest (76.95%) and SVM (77.08%) outperform LSTM (65.76%).

Precision: Random Forest (69.89%) and SVM (72.73%) have higher precision than LSTM (51.63%).

Recall (Sensitivity/True Positive Rate): Random Forest (60.09%) and SVM (57.49%) have higher recall than LSTM (43.60%).

F1 Score: Random Forest (64.38%) and SVM (63.49%) have higher F1 scores than LSTM (46.28%).

Based on these metrics, both Random Forest and SVM perform better than LSTM for this binary classification problem. Random Forest and SVM show better overall accuracy, precision, recall, and F1 score. Random Forest appears to have a balanced performance across different metrics.

GITHUB LINK

<https://github.com/chaitanyanandiniMinnekanti/data_mining_final_project>