Background and Problem Description,

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy[1]. However different algorithms and data can result in different results when machine learning is used to predict. This study as a result will emphasize on examining the performance of Random Forest, Nearest Neighbour, and Logistic Regression models on early-stage diabetes risk prediction data [2]. All the three models will be used for prediction of early stage diabetes in patients.

Methods,

We utilized python as one of the finest languages in machine learning to test the performance, notably the numpy, pandas, and matplotlib, sklearn packages. To begin, we utilized hyperparameter turning to determine the ideal settings for each algorithm. Furthermore, the models were created by these best parameters for hyperparameter tuning. Finally, each model's performance was evaluated using accuracy, recall, precision, f1-score based on percentage split, and cross validation.

Results

After performing all the performance tests for Forest, Nearest Neighbour, and Logistic Regression models the results were summarized under the following table (table1).

metrics	KNN	Random Forest	Logistic Regression
Accuracy	0.97	**0.98**	0.93
F1 score	0.97	**0.99**	0.94
Precision	0.98	**0.99**	0.94
Recall	0.97	**0.99**	0.94

Some discussion

According to Graph of three models accuracy using both test_split approach and Cross-validation, it is apparent that Random forest has been the best classified with approximate accuracy of 98% while Logistic regression on the other hand has been the worst model with 93% accuracy.

Considering the confusion matrix, random forest classification had correctly classified 155 patients(101 have diabetes and 54 are negative) with only one patient being incorrectly classified. Logistic regression on the other has managed to positively predict 97 as diabetic patients and 49 as non diabetic patients however 10 patients were incorrectly classified.

Conclusion

Prediction of **diabetes** at an **early stage** can lead to improved treatment, generally based on performance of 3 used models for classification, Random forest has been the best with accurate prediction of 98% whereas logistic regression has been the worst in terms of classifying patients.

References.

[1] I. Education, "What is Machine Learning?", *Ibm.com*, 2021. [Online]. Available: https://www.ibm.com/cloud/learn/machine-learning. [Accessed: 30- Nov- 2021].

[2]"UCI Machine Learning Repository: Early stage diabetes risk prediction dataset. Data Set", *Archive.ics.uci.edu*, 2021. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset. [Accessed: 30- Nov- 2021].