RESEARCH ARTICLE OPEN ACCESS

RFE and SFS Feature Selection for Predicting Diabetes

Bhagyashree Dipak Panpatil*
*(UG student, Department of Information Technology, B.K.Birla College, Kalyan
Email: bhagyashreepanpatil19@gmail.com)

Abstract:
All ages of people can be affected by diabetes. There are basically three forms of diabetes; including type 1, type 2, and gestational diabetes in which type 2 is the most common. There are 422 million people worldwide who have diabetes according to World Health Organization (WHO). To prevent diabetes there are various machine learning and neural networks that are used to predict the diabetes. In this paper, the machine learning and neural network used are the MLP (Multilayer Perceptron) classifier and the RF (Random Forest) classifier. Among these classifiers, the MLP is the neural network classifier whereas Random forest machine learning classifiers. The Diabetes Prediction dataset and Pima Indian diabetes dataset are used in this research work which was collected from Kaggle. The work concluded by comparing both the dataset along the RFE and SFS feature selection techniques where Diabetes prediction dataset has better performance than Pima Indian Diabetes dataset Therefore Random forest and MLP models are used in this research paper.

Keywords — *Diabetes, RFE, SFS, MLP, RF*

I. INTRODUCTION

Diabetes is a disease that affects the human body's processes of glucose i.e. sugar where glucose is the main source of energy of human cells.Diabetes is a disease that has no permanent cure; hence early detection is required(1).Diabetes consists of two primary types Type 1 diabetes and Type 2 diabetes. Cause of Type 1 diabetes is caused by an autoimmune reaction in which the body's immune system mistakenly attacks and destroys insulin-producing bacteria it is usually developed in adolescence or childhood. Previously Type 1 diabetes was known as "juvenile diabetes". Frequent urination, excessive thirst, unexplained weight loss, fatigue, and blurred vision are the common symptoms of Type 1 diabetes. People with Type 1 diabetes need to pay close attention to their diet and engage in regular physical activity to prevent complications whereas Type 2 diabetes is caused due to lifestyle factors like poor diet, obesity, and lack of physical activity. To manage their blood sugar levels, people with Type 1 diabetes need to take insulin injections or use an insulin pump whereas treatment for Type 2 diabetes is to change their lifestyle, oral medications, and sometimes insulin. In Type 2 diabetes glucose is less effectively absorbed into the cell. During pregnancy, Gestational type of diabetes occurs which is also a type of diabetes. It develops during the second or third trimester and usually ends after childbirth. To manage diabetes effectively is to prevent complications such as kidney problems, heart disease, nerve damage, and vision issues.Machine learning is divided into four categories: Supervised Learning, Semi-Supervised Learning, Unsupervised Learning,
Reinforcement Learning. Supervised Learning is a machine learning technique that is used for machine learning with labeled datasets to identify input labels to make prediction and classification (2)

II. LITERATURE REVIEW

Recent literature has produced amount of research to predict diabetes based on symptoms by applying machine-learning and neural network techniques.

In 2021 comparison of machine learning algorithm for diabetes prediction where they found that Logistic regression and support vector machine works well. They performed Person's correlation method by normalizing the data from 0 to 1.They build NN model with different layer various epochs and found that the NN with two hidden layer provided 88.6% accuracy(Khanam & Foo, 2021)

In 2020, where they used ensemble technique for diabetes predication they performed six algorithm KNN, RF, Gradient boosting, DT, SVM and found that RF achieved highest accuracy. (Soni & Varma, n.d.)

In 2023, ensemble technique and hyper-tuning using grid search were they found that stack method achieved the best accuracy with an accuracy of 97.50% (Saihood & Sonuç, 2023)

In 2023, using supervised machine learning algorithm predicted diabeteswhere they used KNN and Naive Bayes and found that naïve Bayes outperform KNN. They performed Data integration, Data transformation, Data reduction, model testing and, model evaluation.(Febrian et al., 2023)

In 2023, two datasets were compared Iraqi society diabetes (ISD) and Pima Indian diabetes (PID) dataset were compared were they used RFE and GA feature selection technique with the KNN and RF model they found that ISD has higher accuracy than PID(Li et al., 2023)

In 2022, they performed logistics regression, Boost, gradient boosting, decision tree, Extra tree, light gradient boosting machine (LGBM) where they found LGBM shows highest accuracy(Ahamed et al., 2022)

They performed there research with three different classifiers RBF, IBK, and JRip these classifiers were implemented to estimate the performance of the algorithms a comparative feature selection methods Chi-square method, Information gain method, Cluster variation method, and correlation method were used. The information gain method showed the highest accuracy of the 98.78% whereas correlation method denoted the least error rate of 0.0013(Mishra et al., 2016)

III. METHODS

A. *Dataset Description and software tool*

In this research, two datasets are used Diabetes prediction (DP) dataset and the Pima Indian diabetes (PID) dataset where the Pima Indian diabetes (PID) dataset is the most used dataset. The PID and DP datasets both are collected from Kaggle. The DP dataset is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative).The DP dataset consists of features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, blood sugar level, and diabetes. DP dataset contains a total of 9 features. DP dataset includes information about 100,000 patients along with 9 different attributes. In the DP dataset, diabetes is taken as a target variable. Secondly, the PID dataset is originally, from the National Institute of Diabetes and Digestive and Kidney Disease. All the patients in the PID dataset are of Pima Indian heritage females at least 21 years old. The PID dataset consists of features such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome. The PID dataset contains a total of 9 features. The PID dataset includes information about 768 patients along with 9 different attributes. In PID, the Outcome is taken as the target variable. In this paper, the software tool used is jupyter notebook the MLP classifier and Random forest both are implemented in JupyterNotebook and the language used in Python. Fig. 1 represents the PID dataset Fig. 2 represents the PD dataset

Graphical representation of the PID and DP dataset

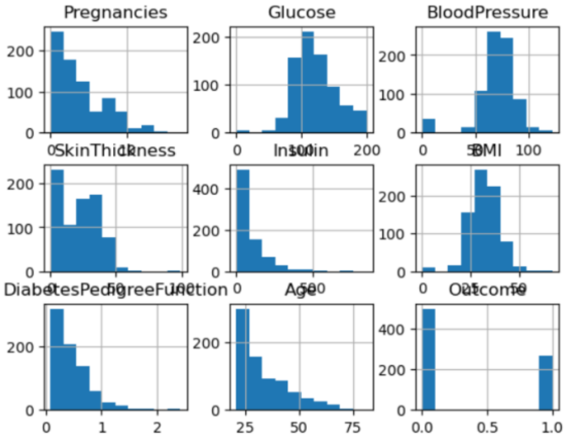


Fig. 1 Graphical representation of PID dataset

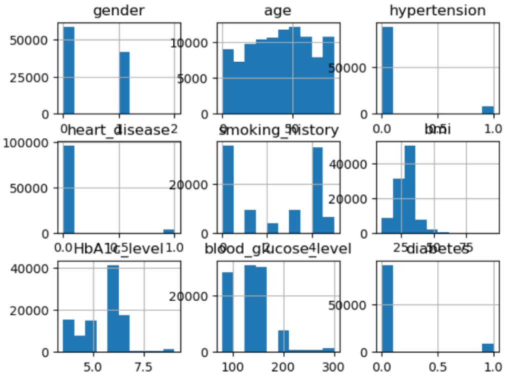


Fig. 2 Graphical representation of DP dataset

B. *Data Pre-processing*

Pre-processing helps transform data so that a better machine learning model can be built, providing higher accuracy (1).

In the DP dataset, data needed to be cleaned because the columns smoking history and gender contained string values that needed to be converted into integers. Therefore, in the gender column male and female converted into 0 & 1, and smoking history column never, no info, current converted into 4, 0 & 1. In the PID dataset every column contained an integer so no data conversion was needed. In both the dataset null values were checked and removed.

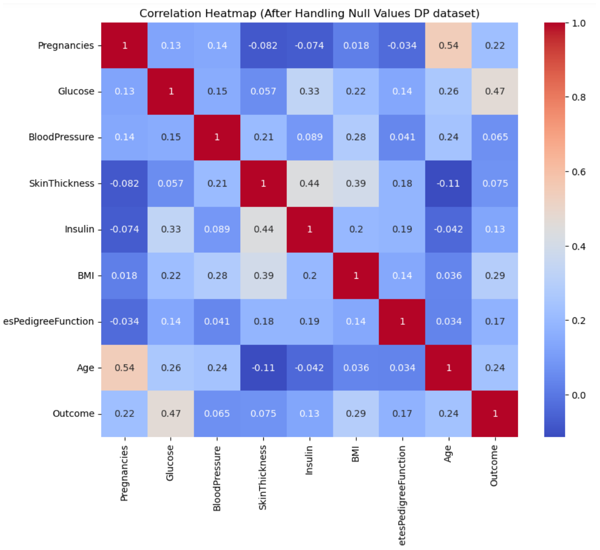


Fig. 3 Correlation Heat map after handling null values (PID)

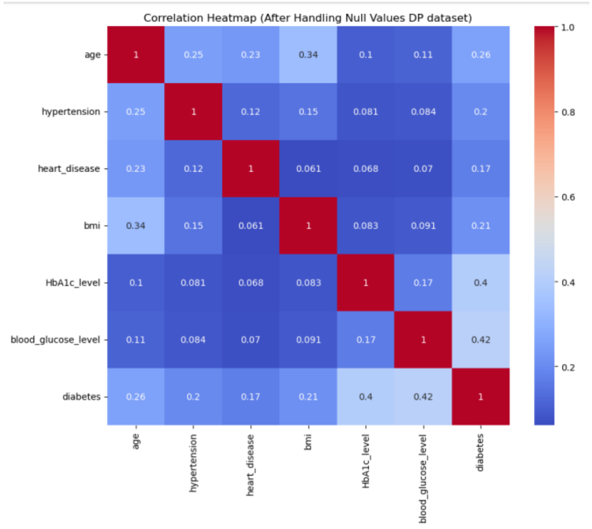


Fig. 4 Correlation Heat map after handling null values (DP)

C. *Feature Selection*

Recursive feature elimination (RFE) and Sequential feature selection (SFS) techniques are applied in both datasets. The goal is to compare the performance of both datasets and also to check which feature selection technique gets better accuracy compared to the other. RFE is a backward selection method that starts with all the features and iteratively removes the least imported features one by one. Until a predetermined number of features or desired level is not achieved RFE continues its process. SFS is a forward selection method that begins with an empty set of features and adds the most important features to the set. SFS evaluates the performance of a model with each added feature. To reduce the dimensionality of the feature space both SFS and RFE are used.

D. *Train and test*

After dataset preprocessing and cleaning the dataset, both datasets are split to train and test. The training set contains 60% of the data whereas the test set contains 40% of the data. Cross-Validation is also performed with the value of CV=3 means 3-fold cross-validation it also means the dataset is divided into 3 equal parts in both datasets.

E. *Classification*

The Random forest (RF) classifier and Multilayer perceptron (MLP) classifier are used in both datasets. The RFE is a popular machine learning algorithm used for classification as well as regression tasks. Multiple decision trees combined to make the prediction using the ensemble method. Multilayer perceptron is a type of artificial neural network that is also used for classification tasks. Multilayer perceptron is also called a feed-forward neural network or simply a neural network. Multilayer perceptron contains three layers input, hidden, and output. Random forest is a supervised machine learning algorithm. RFE is applied directly to both the classifiers but SFS is directly applied to Random forest and not Multilayer perceptron because the library ‘mlxtend’ has no direct support for MLP classifier therefore Random forest is used as a feature selector and the SFS technique is applied to the selected feature to train and test the MLP classifier. Parameters were added to increase the accuracy of the model. In the MLP classifier activation was added as hyperbolic tangent function, solver as Adam, and hidden layer size to 1000 whereas in the Random forest classifier n-estimator was added as 100 and max-depth as 5.

IV. RESULTS

In this research, it is found that the Diabetes Prediction (DP) dataset is better than the Pima Indian diabetes (PID) dataset. The Diabetes Prediction (DP) dataset has better accuracy than the Pima Indian diabetes (PID) dataset. Comparing both the feature selection techniques in the Random forest classifier RFE performance is better than SFE but in the MLP classifier SFS performance is better than RFE as shown in Table 1.

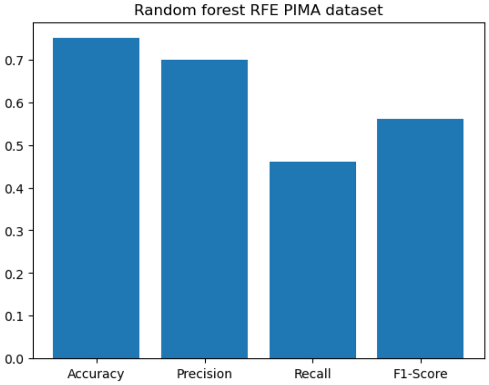
Accuracy = (TP+TN) / (TP+TN+FN+FP)

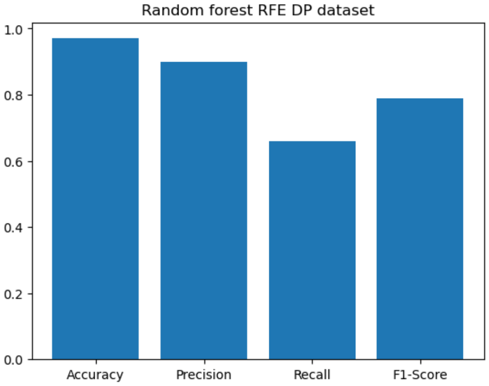
TABLE I

COMPARISON OF BOTH DATASET

	ACCURACY%	
	PIMA INDIAN DIABETES DATASET	**DIABETES PREDICTION DATASET**
Random forest (RF) using RFE	76%	97.17%
Multilayer Perceptron (MLP) using RFE	75.32%	94%
Random forest (RF) using SFS	75%	97%
Multilayer Perceptron (MLP) using SFS	75.57%	96.42%

Bar graph of the Random Forest using RFE just to compare both the dataset


Random forest RFE PIMA dataset


Random forest RFE DP dataset

V. CONCLUSION

In this research Recursive feature elimination (RFS) techniques and sequential feature selection (SFS) which are feature selection techniques were applied in the Random forest classifier and Multilayer perceptron classifier. The Pima Indian diabetes (PID) dataset and diabetes prediction (DP) dataset were compared with each other and along with RFE and SFS techniques cross validation using 3 folds was also used which split the dataset into 3 equal parts and along with cross- validation parameters were added in both the model to improve the accuracy. In the Pima diabetes (PID) dataset accuracy was greater than 70% whereas in the Diabetes prediction (DP) dataset accuracy was greater than 90% in both random forest classifier and MLP classifier. The accuracy of the random forest using the RFE technique for the PID dataset was found as 76% and the DP dataset 97.17% and MLP classifier accuracy for the PID dataset was found as 75.32% and the DP dataset 94%. The accuracy of the random forest using the SFS technique for the PID dataset was found as 75% and the DP dataset 97% and the MLP classifier accuracy for the PID dataset was found as 75.57% and the DP dataset 96.42%. Comparing these 2 datasets Diabetes prediction (DP) dataset is better than the Pima diabetes prediction (PID) dataset. Hence Diabetes Prediction (DP) dataset was found to have better performance than the Pima Indian diabetes (PID) dataset. Further research can be extended by using other machine learning and neural network models or taking new dataset and comparing the performance with these research paper.

Data Availability

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

REFERENCES

1. *Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. ICT Express, 7(4), 432–439.* *https://doi.org/10.1016/j.icte.2021.02.004*

2. *Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., &Yunanda, R. (2023). Diabetes prediction using supervised machine learning. Procedia Computer Science, 216, 21–30.* *https://doi.org/10.1016/j.procs.2022.12.107*

3. Ahamed, B. S., Arya, M. S., & Nancy V, A. O. (2022). Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques. *Frontiers*

in Computer Science, *4*, 835242. https://doi.org/10.3389/fcomp.2022.835242

4. Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A. T., Ghazal, T. M., & Ahmad, M. (2022). Prediction of Diabetes Empowered With Fused Machine Learning. *IEEE Access*, *10*, 8529–8538. https://doi.org/10.1109/ACCESS.2022.3142097

5. Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., & Yunanda, R. (2023). Diabetes prediction using supervised machine learning. *Procedia Computer Science*, *216*, 21–30. https://doi.org/10.1016/j.procs.2022.12.107

6. Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, *7*(4), 432–439. https://doi.org/10.1016/j.icte.2021.02.004

7. Li, X., Curiger, M., Dornberger, R., & Hanne, T. (2023). Optimized Computational Diabetes Prediction with Feature Selection Algorithms. *Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, 36–43. https://doi.org/10.1145/3596947.3596948

8. Mishra, S., Chaudhury, P., Mishra, B. K., & Tripathy, H. K. (2016). An implementation of Feature ranking using Machine learning techniques for Diabetes disease prediction. *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, 1–3. https://doi.org/10.1145/2905055.2905100

9. Saihood, Q., & Sonuç, E. (2023). A practical framework for early detection of diabetes using ensemble machine learning models. *Turkish Journal of Electrical Engineering and Computer Sciences*, *31*(4), 722–738. https://doi.org/10.55730/1300-0632.4013

10. Soni, M., & Varma, D. S. (n.d.). Diabetes Prediction using Machine Learning Techniques. *International Journal of Engineering Research*, *9*(09).