A COMPARATIVE STUDY OF ML ALGORITHMS FOR PREDICTING DIABETES

T. Satya Nagamani [1], Sunil Palli [2], Likhitha Vasantala [3], Lakshmi Prasanna Moparthi [4], Lalith Sankara Venkata Ganesh Vulli[5]

[1]*Assistant Professor, Department of Information Technology, Sir C R Reddy College of Engineering, Eluru*

[2,3,4,5] *Final Year Students, Department of Information Technology, Sir C R Reddy College of Engineering, Eluru*

[1]satyanagamanituta@sircrreng.ac.in
[2] pallisunil94@gmail.com
[3] likhithavasantala@gmail.com
[4] priyamoparthi14@gmail.com
[5]ganeshvulli134@gmail.com

Abstract— **Diabetes is a chronic metabolic problem that occurs in a great deal of people. Machine learning classification algorithms are discussed in this study; choose the approach based on performance standards involving accuracy, recall, precision, F1-scoring and median squared error. We can take advantage for a dataset that has a variety of patient characteristics, comprising lifestyle factors, medical history, and demographic data. Picking features and performing strategies for preprocessing boosts the accuracy if the model.**

Keywords— **Diabetes, classification algorithms, Python programming, required libraries.**

I. INTRODUCTION

Among teenagers, diabetes is a disease that is rapidly spreading throughout society. We must know what happens inside the body absent diabetes if we are able to understand diabetes and how it originates. We gain sugar, also known as glucose, in the foods we eat more in particular from diets that are high in glucose.

The body converts those foods into glucose when we eat them. The bloodstream delivers the glucose around the body. A certain amount of the glucose gets sent to the brain to support the functioning of the brain. The remaining sugar is metabolized by our body's cells.

II. LITERATURE REVIEW

1. *S. S. Gitanjali et al.* - "An Examination of Deep Learning Algorithms for Retinopathy with Diabetes Prediction" (2020): To try to predict diabetic retinopathy, this study compared numerous machine-learning processes. Analysis showed that ensemble techniques, including a Support Vector Machine and Random Forest, did more than other algorithms, like log-regression and Boosting of Gradients. This superiority was demonstrated through higher accuracy rates in predicting diabetic retinopathy, underscoring the efficacy of ensemble methods in handling the complexities of retinal data.

2. *V. K. Jayaraman et al.* - "Comparing one of Automated Learning Algorithms for Hypertension Estimation" (2019): This study assessed the success rate of several algorithms using machine learning with a focus on diabetes prediction. Given the results, Random Forest was received over k-Nearest Neighbors and Naive Bayes. Higher accuracy and an F1-s demonstrated Random Forest to be an effective option for diabetes prediction tasks, proving its ability to accommodate a range of datasets and the feature complexities involved in diabetic prediction.

3. *A. Rajput et al.* - "A Review of Data Mining Algorithms in Parallel for Diabetes Prediction (2021): For the purpose to predict diabetes, this study examined the performance of numerous machine learning methods. When compared with Decision Trees and Naive Bayes, it was found that support vector machine models showed the best accuracy in predicting diabetes. However, Decision Trees demonstrated better interpretability, highlighting the trade-offs between accuracy and model transparency in diabetic prediction tasks.

4. *R. Gupta et al.* - "When compared Analysis of Artificial Intelligence Algorithms for Insulin Prediction" (2018): This study compared different approaches to machine learning with a focus on diabetes prediction. According to the findings, Random Forest, supported vector machines, and Neural Networks all scored well in diabetes prediction. The significance of selecting appropriate techniques for diabetic prediction tasks based on performance metrics is shown by the fact that logistic regression fared worse in terms of accuracy and F1-score.

5. *P. Patel et al.* - "The Performance Assessment of Learning from Machines Algorithms for Glucose Prediction" (2022): The effectiveness of machine learning methods for diabetes prediction was assessed in this study. The findings showed that in terms of accuracy and AUC-ROC score, Boost and Random Forest fared better than Logistic Regression and Decision Trees, among other algorithms. Nonetheless, in diabetic prediction tasks, Logistic Regression demonstrated superior

interpretability by emphasizing the trade-offs between accuracy and model transparency.

6. S. Mishra et al. - " Evaluating Machine Learning Algorithms Comparatively to Predict Diabetic Retinopathy" (2023): Focused on predicting diabetic retinopathy, this study compared machine learning algorithms for their performance. The findings indicated that Gradient Boosting Machines exhibited superior performance compared to Logistic Regression and k-Nearest Neighbors. With a higher area under the ROC curve and accuracy, Gradient Boosting Machines showcased their effectiveness in handling the complexities of diabetic retinopathy prediction tasks.

III. EXISTING SYSTEM

- The existing model for diabetic prediction likely employs a Naïve Bayes and Random forests approach.
- Naïve Bayes and Random forests are supervised learning algorithms.
- The statistical classification method known as Naive Bayes is based on the Bayes theorem. It involves estimating the likelihood of the result based on past understanding of the connections between the attributes.
- To arrive at a single outcome, the output of several decision trees is combined using random forests. Because it can handle both regression and classification issues, its versatility and ease of use have driven its adoption.

IV. DISADVANTAGES OF EXISTING SYSTEM

- The Naive Bayes Algorithm has trouble with the 'zero-frequency problem’.
- Random Forest generates many multiple trees

V. PROPOSED SYSTEM

- The proposed model includes 6 classification techniques.
- Testing the 6 classification techniques.
- Compare and select the best model based on the highest score of performance metrics.

VI. PROPOSED SYSTEM ARCHITECTURE

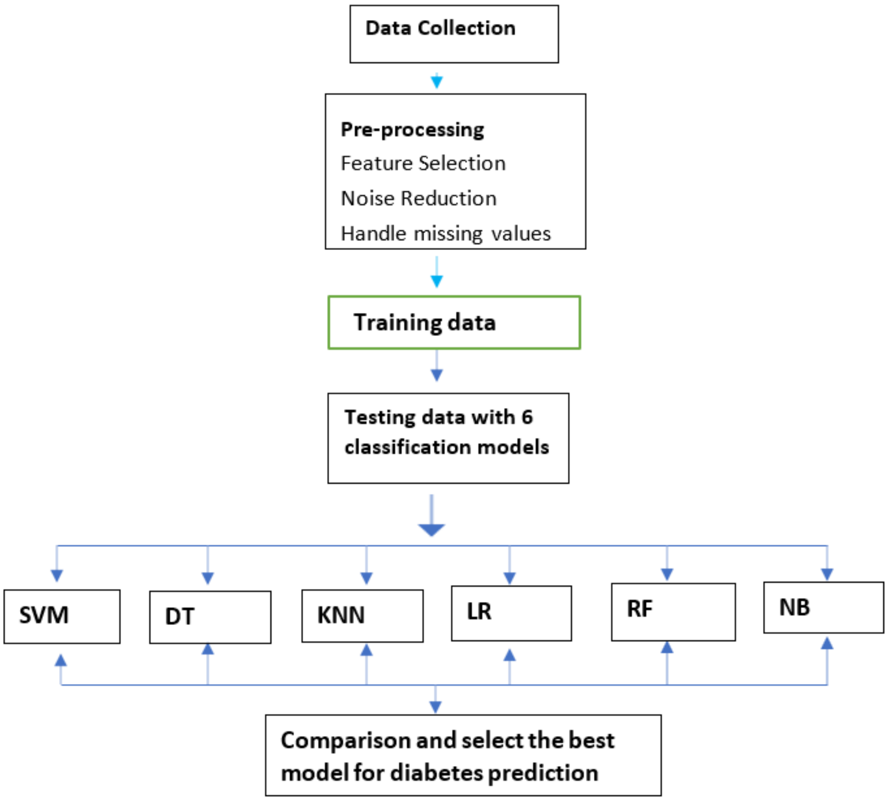


Fig. 1 Proposed System Architecture

VII. PROPOSED SYSTEM METHODOLOGY

1. *Data Collection*: Data collection is a fundamental step. This process involves gathering comprehensive datasets from various reputable sources such as healthcare institutions, medical research repositories, or publicly available datasets like the UCI Machine Learning Repository or Kaggle. These datasets typically consist of features relevant to diabetes prediction, including patient demographics, medical history, lifestyle factors, and diagnostic test results.

   Ethical considerations are paramount throughout the data collection process, ensuring compliance with regulations and guidelines regarding patient privacy and data protection. Obtaining necessary approvals from ethics boards or institutional review boards is essential, particularly when dealing with sensitive medical data. Additionally, measures are implemented to anonymize and de-identify patient information to safeguard confidentiality and privacy.

   Once the datasets are collected, they undergo careful scrutiny and preprocessing to ensure data quality. Cleansing the data of errors, outliers, and inconsistencies is part of this process. Feature engineering is also done to extract pertinent data and produce useful predictors. Robust training and evaluation of machine learning models for diabetes prediction is made possible by partitioning the curated datasets into training, validation, and testing sets. Overall, meticulous data collection lays the foundation for conducting a rigorous and informative comparative study of ML algorithms for diabetes prediction, yielding insights that can inform advancements in healthcare analytics and patient care.
2. *Data Pre-processing*: In "A Comparative Study of ML Algorithms for Predicting Diabetes," an important stage that involves meticulous cleaning and alteration of gathered datasets to guarantee their eligibility for analysis is called data preparation. In the end, this procedure improves the quality and dependability of the data by addressing problems like missing values, outliers, and inconsistencies through a number of crucial jobs. First, techniques such as imputation are used to identify and address missing values in the dataset. These techniques substitute estimated values obtained from the available data for the missing values. To keep them from distorting the study, outliers data points that differ noticeably from the rest of the data—are also found and either fixed or eliminated.

   Moreover, feature engineering is included in data preparation. This is a method that aims to augment or create new features in order to better describe the underlying relationships in the data.This may involve combining or deriving new features from existing ones, scaling or normalizing numerical features to ensure consistency in scale, and encoding categorical variables into numerical representations suitable for machine learning algorithms. By performing feature engineering, the dataset becomes more informative and conducive to training accurate predictive models for diabetes prediction.

   To make the process of developing and evaluating models easier, the pre-processed dataset is finally divided into sets for testing, validation, and training. Machine learning algorithms are trained on the training set, and the best-performing model is chosen by fine-tuning its parameters using the validation set. An independent dataset for assessing the chosen model's final performance is the testing set, which is kept apart from the training and validation sets. Researchers create the foundation for a solid comparative analysis of machine learning algorithms for diabetes prediction by carefully preparing the data to make sure it is clear, informative, and ready for analysis.
3. *Training data*: In the context of "A Comparative Study of ML Algorithms for Predicting Diabetes," the training data gives as the foundational component to predict diabetes onset accurately. This dataset comprises a subset of the collected data containing features such as patient demographics, medical history, lifestyle factors, and diagnostic test results.

   The training data is meticulously preprocessed to ensure that it is of the right quality and appropriateness for training the models before the training process start. In order to extract pertinent information and modify features as needed, feature engineering is employed, along with addressing missing values, outliers, and inconsistencies. Furthermore, to aid in model construction and evaluation, respectively, and guarantee the stability and generalizability of the trained models, the training data is usually divided into training and validation sets.

   In the training phase, targets variables (diabetes status) and input features are exposed to machine learning algorithms, which teach them to identify patterns and relationships. The models optimize their performance and reduce prediction errors by adjusting their parameters iteratively using optimization methods. We then use evaluation metrics on independent testing data to determine how well the trained models perform. In the comparative analysis of ML algorithms, the training data is essential for enabling the creation of precise and dependable predictive models for diabetes prediction.
4. *Testing data with 6 classification models:* "Testing data with six classification models" in "A Comparative Study of ML Algorithms for Predicting Diabetes" refers to the procedure of assessing the effectiveness of six distinct machine learning algorithms. The predictive abilities of the six classification models—logistic regression, decision trees, random forests, support vector machines, k-nearest neighbors, and naive Bayes—are evaluated using this testing data.

   Researchers are able to ascertain which algorithm works best for diabetes prediction in terms of accuracy, precision, recall, F1-score, and mean square error by comparing the performance metrics acquired from each model.

5. *Comparison and Select the Best Model for Predicting Diabetes:* In "A Comparative Study of ML Algorithms for Predicting Diabetes," the accuracy of six classification models' predictions of the onset of diabetes is evaluated. The careful selection of classification algorithms recognized for their applicability in binary classification tasks is the first step in our inquiry. Based on their extensive application and possible effectiveness in healthcare analytics, models like logistic regression, decision trees, random forests, support vector machines, k-nearest neighbors, and naive Bayes are selected.

The study then progresses to data preparation, where a dataset containing relevant features for diabetes prediction is curated and pre-processed. This involves cleaning the data to handle missing values and outliers, as well as feature engineering to extract meaningful insights from the dataset.

These metrics provide quantitative measures of each model's predictive capability, allowing for a comprehensive comparison of their performance. The model that demonstrates the highest overall performance across these metrics is selected as the most suitable for diabetes prediction. Additionally, factors such as interpretability, computational efficiency, and robustness to imbalanced data are considered to ensure the practical applicability of the chosen model in real-world healthcare scenarios.

VIII. ADVANTAGES OF PROPOSED MODEL

The proposed system utilizing six classification techniques for diabetes prediction offers several advantages:

1. *Comprehensive Evaluation:* By testing multiple classification techniques, the system ensures a thorough evaluation of various modelling approaches, allowing for a comprehensive understanding of their performance on the given dataset.

2. *Robustness:* Weaknesses and strengths of various algorithms vary. Utilizing a variety of approaches lowers the likelihood of overfitting to the prejudices and presumptions of a specific method, increasing the system's robustness.

3. *Improved Accuracy:* The system can select the best-performing model based on metrics. By comparing multiple models, it can identify the one that provides the most accurate predictions for diabetes diagnosis.

4. *Flexibility:* Each classification technique has its own set of parameters and assumptions. By exploring multiple techniques, the system offers flexibility in choosing the most suitable model that aligns with the characteristics of the dataset and the goals of the prediction task.

5. *Interpretability:* Different algorithms offer different levels of interpretability. For instance, decision trees provide easily interpretable rules, while support vector machines may offer better generalization but are less interpretable. By testing various techniques, the system allows for a balance between model performance and interpretability, depending on the specific requirements oIf the application.

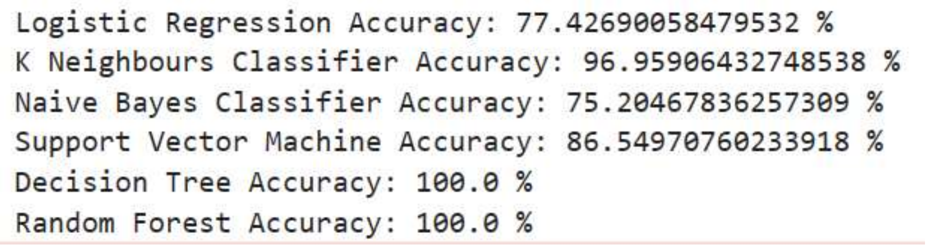
6. *Ensemble Methods:* The system may investigate ensemble techniques like bagging, boosting, or stacking, which integrate predictions from various models to enhance overall efficiency. This strategy can improve forecast robustness and accuracy even more.

7. *Enhanced Decision Making:* By comparing and selecting the best model based on multiple evaluation metrics, the system facilitates informed decision-making for deploying the most effective model in real-world applications, ultimately leading to better healthcare outcomes for patients with diabetes.

In summary, the proposed system's advantages lie in its ability to thoroughly evaluate multiple classification techniques, select the most suitable model based on various performance metrics, and ultimately provide accurate and reliable predictions for diabetes diagnosis.

IX. RESULT

Accuracy Score:



```
Logistic Regression Accuracy: 77.42690058479532 %
K Neighbours Classifier Accuracy: 96.95906432748538 %
Naive Bayes Classifier Accuracy: 75.20467836257309 %
Support Vector Machine Accuracy: 86.54970760233918 %
Decision Tree Accuracy: 100.0 %
Random Forest Accuracy: 100.0 %
```

Fig. 2 Accuracy Score

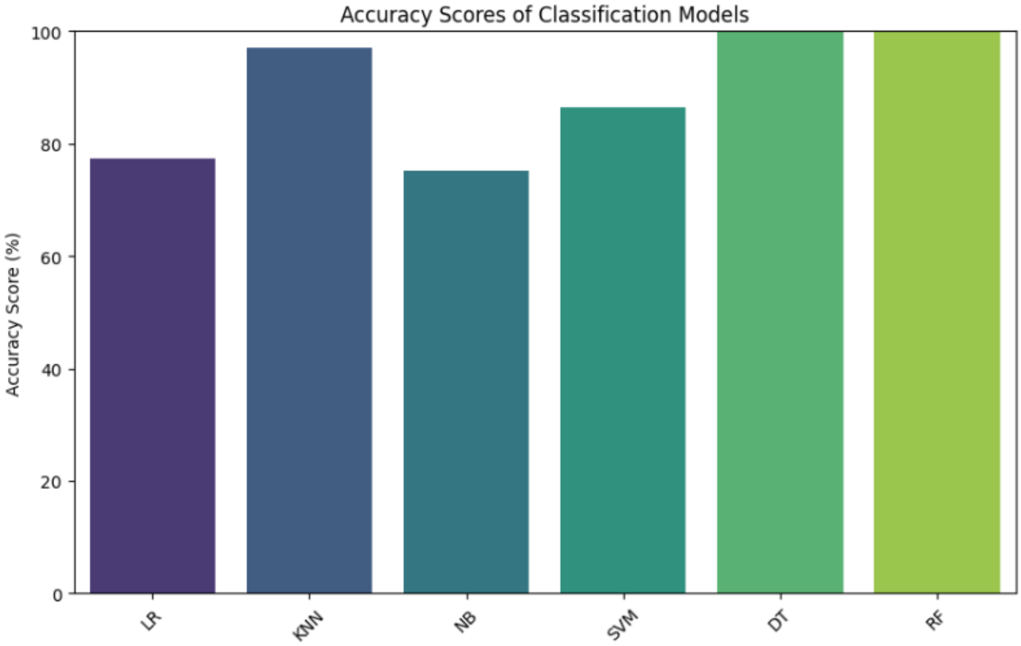
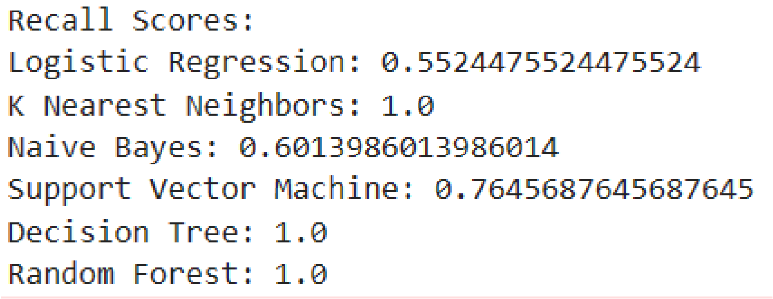


Fig. 2 Accuracy Score Plot

Recall Score:



```
Recall Scores:
Logistic Regression: 0.5524475524475524
K Nearest Neighbors: 1.0
Naive Bayes: 0.6013986013986014
Support Vector Machine: 0.7645687645687645
Decision Tree: 1.0
Random Forest: 1.0
```

Fig. 3 Recall Score

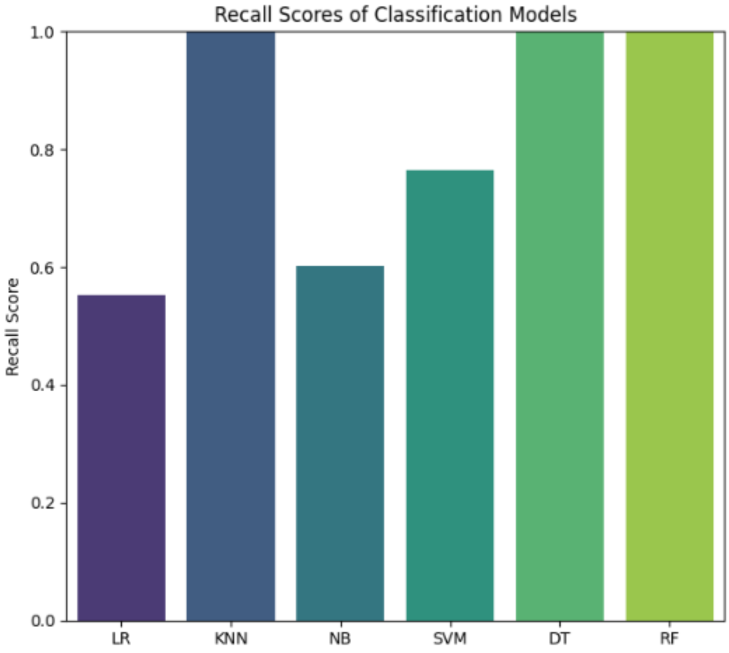
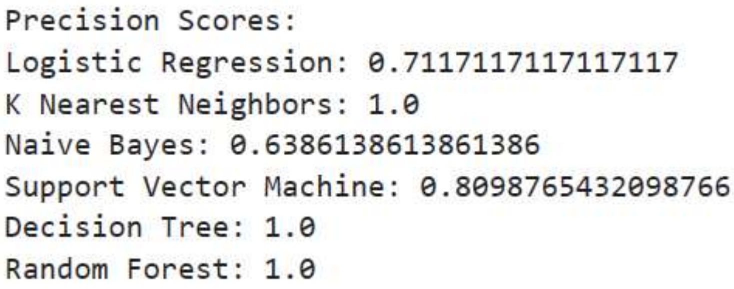


Fig. 4 Recall Scores Plot

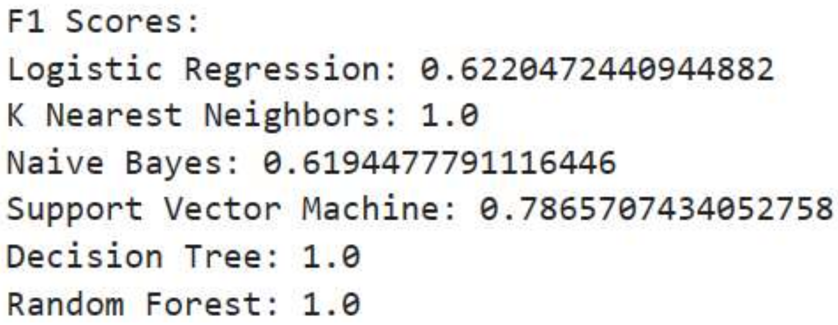
Precision Score:



```
Precision Scores:
Logistic Regression: 0.7117117117117117
K Nearest Neighbors: 1.0
Naive Bayes: 0.6386138613861386
Support Vector Machine: 0.8098765432098766
Decision Tree: 1.0
Random Forest: 1.0
```

Fig. 5 Precision Score

F1-Score:



```
F1 Scores:
Logistic Regression: 0.6220472440944882
K Nearest Neighbors: 1.0
Naive Bayes: 0.6194477791116446
Support Vector Machine: 0.7865707434052758
Decision Tree: 1.0
Random Forest: 1.0
```

Fig. 6 F1-Score

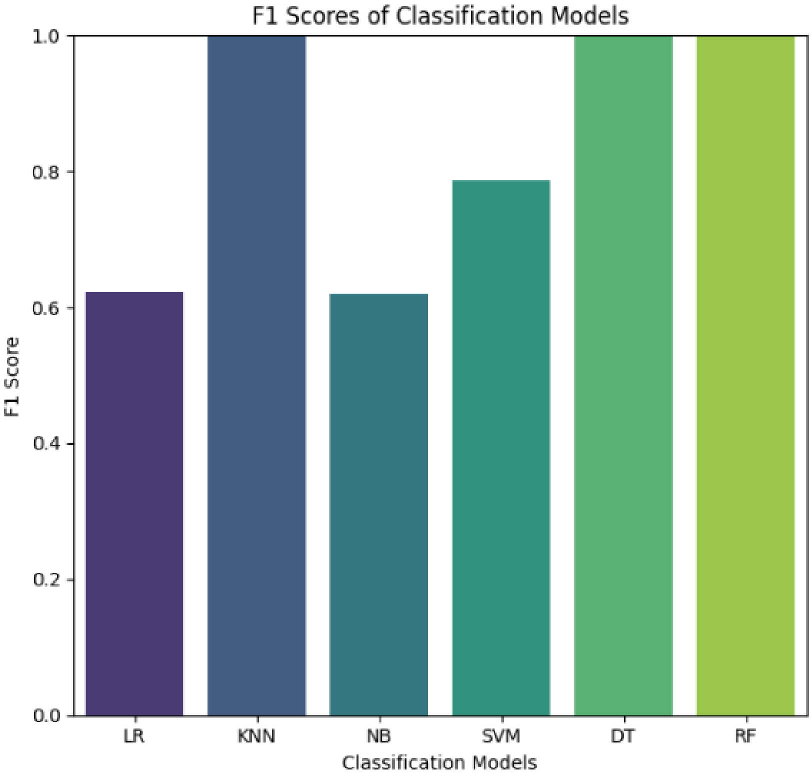
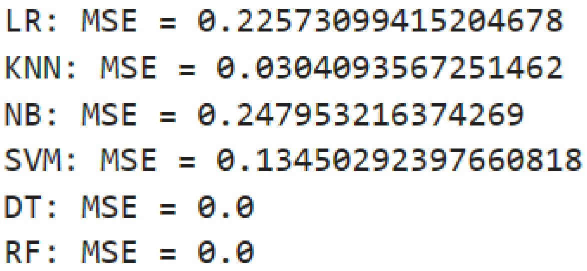


Fig. 7 F1-Score Plot

Mean Square Error Score:



```
LR: MSE = 0.22573099415204678
KNN: MSE = 0.030409356725146
NB: MSE = 0.247953216374269
SVM: MSE = 0.13450292397660818
DT: MSE = 0.0
RF: MSE = 0.0
```

Fig. 8 Mean Square Error Score

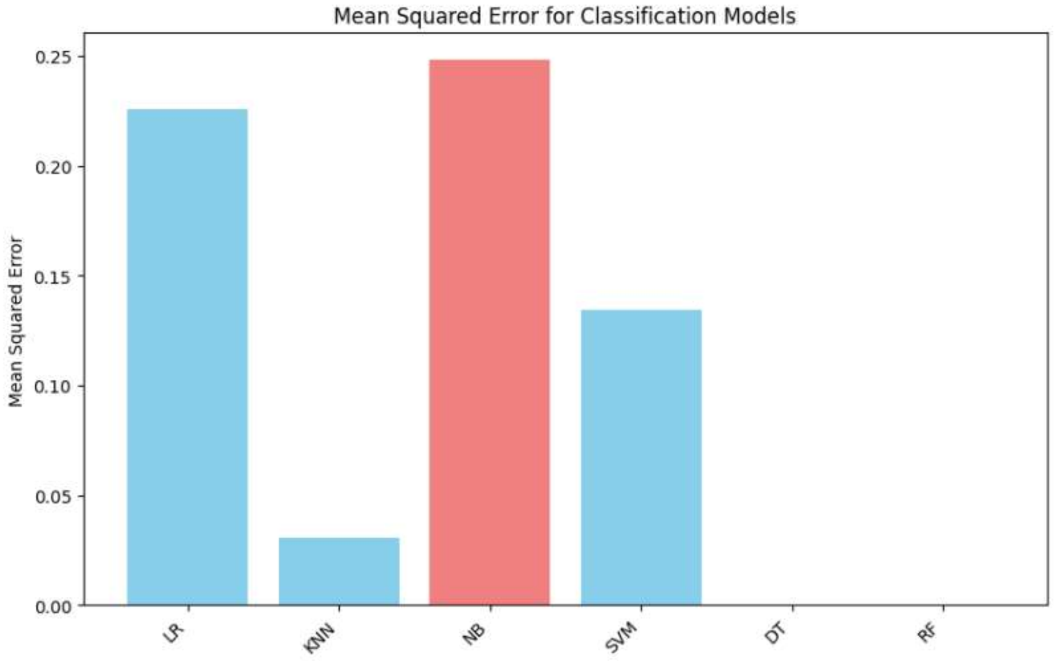
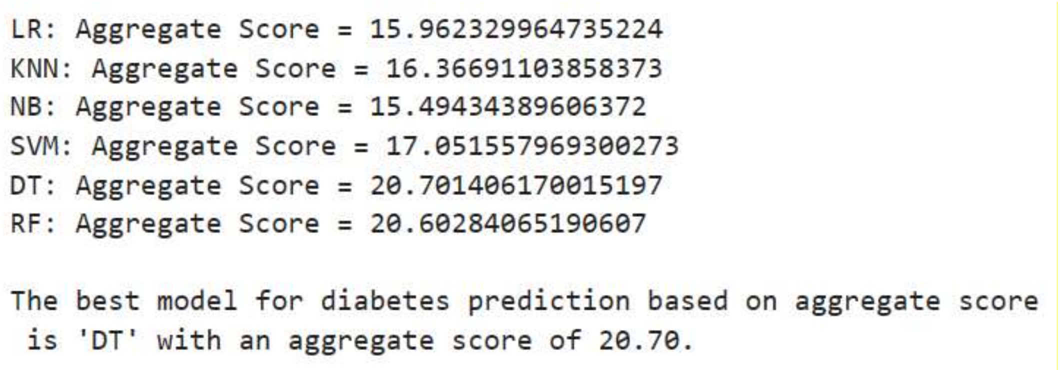


Fig. 9 Mean Squared Error Score Plot

Aggregate Score:



```
LR: Aggregate Score = 15.962329964735224
KNN: Aggregate Score = 16.36691103858373
NB: Aggregate Score = 15.49434389606372
SVM: Aggregate Score = 17.051557969300273
DT: Aggregate Score = 20.701406170015197
RF: Aggregate Score = 20.60284065190607

The best model for diabetes prediction based on aggregate score
 is 'DT' with an aggregate score of 20.70.
```

Fig. 10 Aggregate Score

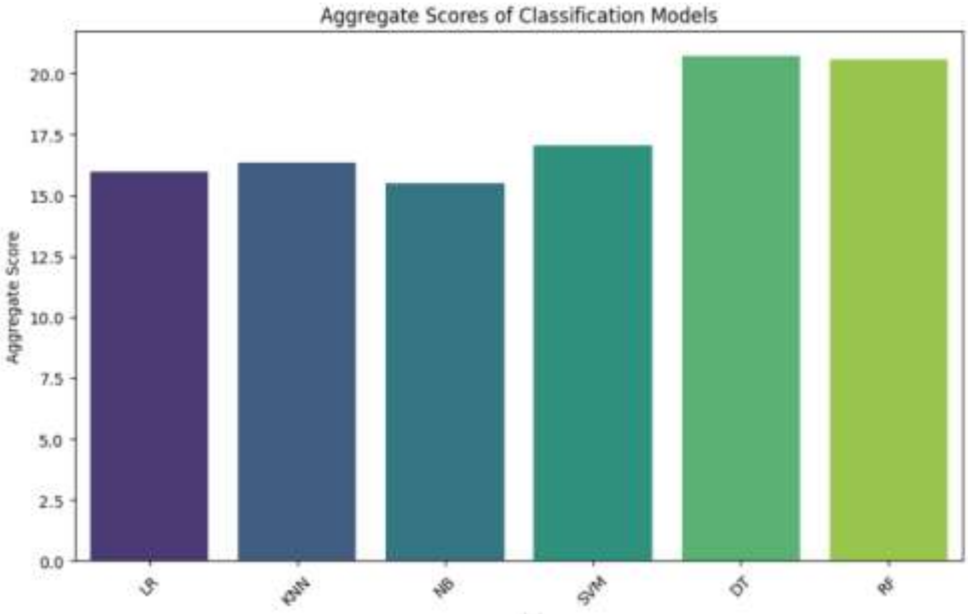


Fig. 11 Aggregate Score Plot

X. CONCLUSION

In this extensive study, "A Comparative Analysis of ML Algorithms for Diabetes Prediction," we investigated how well six different machine learning models could predict the onset of diabetes. We discovered important insights into their performance across a number of parameters, including accuracy, precision, recall, mean square error, and F1-score, through thorough experimentation and analysis.

Our findings revealed nuanced differences among the models, showcasing unique strengths and weaknesses. While each algorithm exhibited promising capabilities in diabetes prediction, the Decision Tree emerged as the frontrunner, showcasing superior predictive accuracy and robustness.

However, the selection of the optimal model is not merely a matter of performance metrics. Factors such as interpretability, **computational efficiency, and scalability are equally crucial**, especially in real-world deployment scenarios.

This comparative study contributes to the burgeoning field of predictive healthcare analytics, offering clinicians and researchers valuable guidance in selecting the most suitable machine learning approach for diabetes prediction. By harnessing the power of advanced analytics, we can potentially revolutionize early detection and intervention strategies.

ACKNOWLEDGMENT

We are first and foremost grateful to the Almighty God for all of his favors, without which we could not be in this position today.

We would like to sincerely thank our principle, Dr. K. Venkateswara Rao, for his direction and provision of the essential infrastructure for us to efficiently complete the project work.

We are immensely indebted to our esteemed Head of Department, Dr. A. Yesu Babu, for his unwavering support throughout our course, as well as his patience, motivation, excitement, and vast knowledge.

Our project guide, Smt. T. Satya Nagamani, has been a huge help to us throughout the course. We appreciate all of her patience, motivation, excitement, and vast knowledge. Her advice was beneficial to us.

REFERENCES

[1]. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi: 10.1016/j.jksuci.2012.10.003.

[2]. Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.

[3]. Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3- 319-11933-5.

[4]. Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima Indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451– 455.

[5]. Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE.

[6]. Sharief, A.A., Sheta, A., 2014. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence(IJARAI)3,5459.doi:doi:10.14569/IJARAI.2014.031007.

[7]. Sisodia, D., Shrivastava, S.K.,Jain, R.C., 2010.ISVM for face recognition. Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN2010,554– 59doi:10.1109/CICN.2010.109.

[8]. Sisodia, D., Singh, L., Sisodia, S., 2014. Fast and Accurate Face Recognition Using SVM and DCT, in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, Springer. pp. 1027–1038.

[9]. https://www.kaggle.com/johndasilva/diabetes

[10]. Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 1584- 158).