Diabetes Prediction

Team 03

B.Sai Abhishek
BL.EN.U4AIE21015
Artificial Intelligence,
Amrita School of Engineering,
Banglore, India
abhishekbusetty@gmail.com

Balam Ruchith balaji
BL.EN.U4AIE21017
Artificial Intelligence,
Amrita School of Engineering,
Banglore, India
bl.ruchith@gmail.com

Chillakuru Hari
BL.EN.U4AIE21038
Artificial Intelligence,
Amrita School of Engineering,
Banglore, India
Hari291010@gmail.com

Dr. Manju Venugopalan
Assistant Professor (Sr.Gr),
Amrita School of Engineering,
Bengaluru, India

***Abstract*—**Diabetes is a common, long-term illness that affects millions of individuals globally. Patient outcomes can be greatly enhanced by diabetes early identification and accurate prediction. This research uses PySpark, a potent distributed computing platform for big data processing, to investigate the use of many machine learning methods, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Naive Bayes, for the prediction of diabetes.The dataset used in this study is an assortment of pertinent health and lifestyle data collected from people, such as age, blood pressure, BMI, and glucose levels. The goal is to create and evaluate prediction models in order to determine who is at risk of developing diabetes. For binary classification, the logistic regression model serves as a baseline, while models like as decision trees and random forests are investigated for their capacity to represent intricate relationships in the data. Naive Bayes is chosen for its ease of use and success in probabilistic classification, whereas Support Vector Machines (SVM) is picked for its ability to handle high-dimensional datasets. PySpark is used for the implementation, which makes effective use of its distributed computing features to handle large-scale datasets. Scaling features, dealing with missing values, and encoding categorical variables are all part of data preparation. The models' performances are assessed using metrics like accuracy, precision, recall, and F1 score after they have been trained, validated, and adjusted using the proper methods. This study adds to the increasing body of research on machine learning-based diabetes prediction and shows that these models may be implemented in a distributed computing environment like PySpark. The results may help medical practitioners identify those who are at risk of diabetes, which would enable early intervention and individualized treatment plans.

***Keywords*—Diabetes Prediction,Machine Learning PySpark, Logistic Regression,Decision Tree, Random Forest,Support Vector Machine [SVM**

I. INTRODUCTION

Millions of people worldwide suffer with diabetes, a chronic metabolic disease that is still a serious global health concern. For the purpose of putting preventative measures and individualized healthcare interventions into practice, early identification of those at risk of diabetes is essential. Algorithms for machine learning have become useful tools for healthcare predictive modeling in recent years. This study focuses on implementing and contrasting multiple machine learning methods for the prediction of diabetes by utilizing the capabilities of the distributed computing platform PySpark. Diabetes is becoming more and more common, which raises serious concerns about world health and calls for creative methods of early identification and risk assessment. Machine learning presents a viable path forward for predictive healthcare analytics because of its capacity to examine intricate patterns in large datasets. Using a suite of machine learning algorithms developed in PySpark, a distributed computing framework renowned for its scalability and effectiveness in handling huge datasets, we explore the field of diabetes prediction in this paper. Diabetes, which is defined by elevated blood glucose levels, necessitates the early detection of high-risk patients and the prompt implementation of therapies. Using PySpark into our analysis not only puts our work in the context of big data analytics in healthcare, but it also enables us to leverage distributed computing for increased efficiency. The collection of algorithms selected, which includes Naive Bayes, Random Forest, Decision Tree, Support Vector Machine (SVM), and Logistic Regression, covers a wide range of machine learning approaches. Because of their diversity, it is possible to thoroughly examine how well-suited they are for diabetes prediction using a variety of health indicators. The findings of this study are presented in the following parts, which also cover the methodology used, the nuances of data pretreatment, and the specific execution

of each algorithm in the PySpark environment. The objective of this comparison analysis is to identify the advantages and disadvantages of each algorithm in the particular context of diabetes prediction by comparing performance metrics including accuracy, precision, recall, and F1 score.

II. LITERATURE SURVEY

A. An Efficient Diabetes Prediction Model using Machine Learning

Several machine learning algorithms were examined for the purpose of classification. Logistic Regression, KNN, CART, Random Forest, SVM, and LightGBM achieved accuracy rates of 84.8%, 84%, 85.7%, 88.1%, 85.3%, and 88.2%, respectively, when the accuracy of each algorithm was assessed. The accuracy was further increased to 90.2% by selecting the LightGBM model following hyper-parameter adjustment. Compared to previously available datasets, this study shows a significant improvement in the accuracy and precision of diabetes prediction with the use of the LightGBM model. Future research could examine the likelihood that those who are not already diabetics will become so in the future.

B. .Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data

This study uses pertinent data mining techniques to forecast diabetes. Knowledge extraction from the data contained in the system is the aim of data mining. dataset in addition to looking for trends. People in this civilization suffer from a variety of health issues and are often unaware of the symptoms or the underlying causes of their illnesses. Diabetic Mellitus is one of the health issues. The bulk of the population suffers from diabetes. Younger individuals nowadays are also experiencing this issue. In this work, we have employed HUE predictive analysis to anticipate the kind of behaviors that will be expected as well as the chronic disorders and their associated details.

C. Diabetes Mellitus Prediction using Supervised Machine Learning Techniques

Early diabetes prediction could potentially save a life. We have worked very hard to build a system and obtain the best possible outcomes in this task. In this study, two machine learning algorithms are compared using various criteria. The experimental work in this case uses a Kaggle dataset that is publicly available. The Random Forest classification algorithm yields an accuracy of 99.03% in the experimental results of the developed model. You can also use logistic regression classification, which has a good 94.23% accuracy rate. The prediction or diagnosis of additional diseases using different machine learning algorithms is part of the future work. As a result, the work can be improved and extended to automate the diabetes analysis.

III. METHODOLOGY

To conduct this research, we followed a systematic procedure outlined in

A. Dataset and Features Description:

We utilized diabetes data for our prediction model, covering 1 lakh instances. The dataset comprises of Nine columns.
The dataset includes the following fields:

Age,Gender,Hypertension,heart_disease,smoking_history,bmi,HbA1c_level,blood_glucose_level,diabetes.

B. Dataset Preparation:

Imported necessary libraries (e.g., NumPy, pandas, matplotlib, seaborn). Loaded the dataset for the

Prepared the data by removing the string values(smoking history) and also the gender.

C. Applying Correlation:

Investigated correlations among variables to identify relationships. Analyzed the correlation matrix to guide feature selection.

D. Applying Machine Learning Techniques:

Imported the pyspark library for model development. Selected features and target for X and Y axes. Divide the data in an 80:20 ratio between training and testing sets.

E. Developing Classification Model:

Utilized Logistic Regression, Random Forest Regressor, Navie bayes, SVM, and Decision Tree for Diabetes prediction. Conducted a comparative analysis of the prediction models.

F. Model Performance Analysis:

Calculated errors in the prediction models using Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error.

By following these steps, our research aimed to provide insights into sales prediction models for Walmart stores.

TABLE I
DATASET

gender	age	hypertens	heart_dis	smoking_	bmi	HbA1c_le	blood_glu	diabetes
Female	80	0	1	never	25.19	6.6	140	0
Female	54	0	0	No Info	27.32	6.6	80	0
Male	28	0	0	never	27.32	5.7	158	0
Female	36	0	0	current	23.45	5	155	0
Male	76	1	1	current	20.14	4.8	155	0
Female	20	0	0	never	27.32	6.6	85	0
Female	44	0	0	never	19.31	6.5	200	1
Female	79	0	0	No Info	23.86	5.7	85	0
Male	42	0	0	never	33.64	4.8	145	0
Female	32	0	0	never	27.32	5	100	0
Female	53	0	0	never	27.32	6.1	85	0
Female	54	0	0	former	54.7	6	100	0
Female	78	0	0	former	36.05	5	130	0
Female	67	0	0	never	25.69	5.8	200	0
Female	76	0	0	No Info	27.32	5	160	0
Male	78	0	0	No Info	27.32	6.6	126	0
Male	15	0	0	never	30.36	6.1	200	0
Female	42	0	0	never	24.48	5.7	158	0

Fig 1: Features of Dataset

```
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   gender               100000 non-null  object
 1   age                  100000 non-null  float64
 2   hypertension         100000 non-null  int64
 3   heart_disease        100000 non-null  int64
 4   smoking_history      100000 non-null  object
 5   bmi                  100000 non-null  float64
 6   HbA1c_level          100000 non-null  float64
 7   blood_glucose_level  100000 non-null  int64
 8   diabetes             100000 non-null  int64
dtypes: float64(3), int64(4), object(2)
memory usage: 6.9+ MB
```

Fig 2: Dataset Description

A. Dataset Preparation

Dataset preparation is a critical step in the data mining process, emphasizing data pre-processing. Our primary focus was on the training dataset, where we executed essential transformations. To facilitate model development, we converted categorical data into numerical values. Specifically, we assigned 0 for females and 1 for males, ensuring appropriate representation. After these transformations, the dataset was carefully processed and made ready for analysis. In the final steps, we opted to allocate 75% of the data for training purposes and reserved the remaining 25% for testing. This decision ensures a balanced and effective approach to model evaluation.

Flow Diagram

The objective of this study was to predict the diabetes using a variety of supervised machine learning techniques, such as Support Vector Machine, Random Forest, Decision Trees, Logistic Regression, and Naïve Bayes(Fig. 3). The dataset used to train the predictive models.To guarantee data quality, the dataset underwent extensive cleaning using Python. Important variables that were taken into account After that, the dataset was divided into training and testing sets while keeping a predetermined ratio.

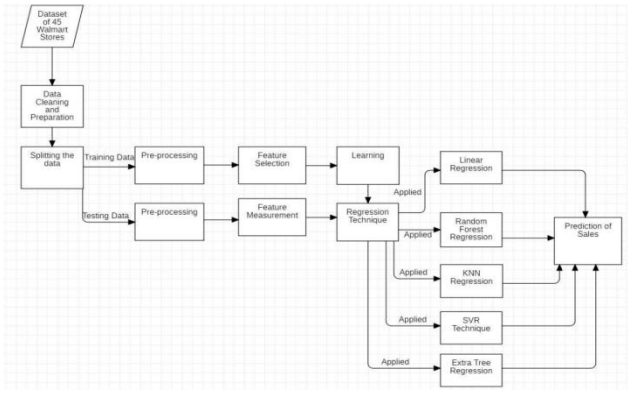
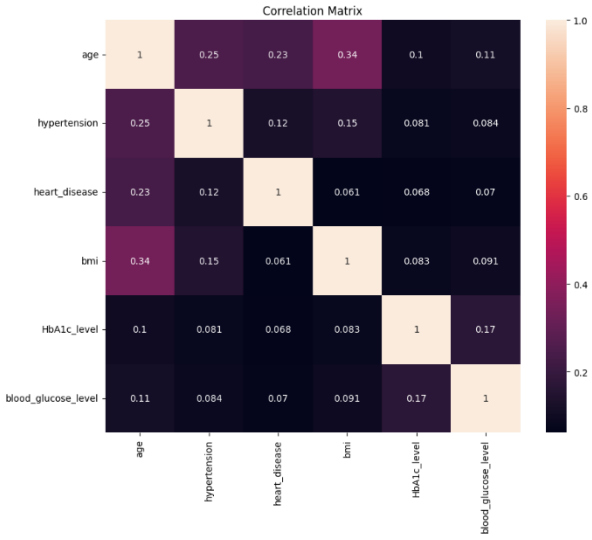


Fig 3: Flow Diagram

B. Correlation

Understanding the relationships among variables is a crucial aspect of our analysis, and we focused on exploring correlations within the dataset. This step involved examining the correlation matrix to identify patterns and connections between different features. Our approach to correlation involved assessing the strength and direction of relationships between variables. By doing so, we gained valuable insights into how various factors interact within the dataset. This information guided our feature selection process, ensuring that we considered the most relevant variables for our analysis. In summary, the correlation step in our research played a pivotal role in shaping the subsequent stages of model development, offering a foundation for informed decision-making regarding feature inclusion.

Fig 4: Correlation between features of dataset



C. Applying Machine Learning Techniques

For the model-building phase, we employed four classification algorithms: Logistic Regression, Random Forest, and Decision Tree. Our analysis aimed at predicting diabetes the factors contributing to attacks on humans by products.

1) Logistic Regression: A statistical technique called logistic regression is applied to binary classification issues in which the outcome variable is categorical and has two classes, usually denoted by the numbers 0 and 1. It is frequently used in machine learning to forecast, given one or more independent variables, the likelihood that an instance will belong to a specific class. The logistic function, also called the sigmoid function, is used in logistic regression to restrict the output between 0 and 1, in contrast to linear regression, which predicts a continuous outcome.

$$P(Y=1)=1+1/(e^{-(b0+b1X1+b2X2+...+bnXn)}$$

2) Random Forest: Random Forest, an ensemble classifier, leverages decision tree algorithms in a randomized manner.

Developed by Leo Breiman, it excels in both regression and classification tasks within supervised machine learning. Notably, Random Forest avoids tree pruning, demonstrating randomness in creating a bootstrap dataset and constructing decision trees from it. The algorithm yields fast results with high prediction accuracy, making it suitable for handling diverse input data effectively. The combination of subsurface randomization and bagging enhances its performance by replacing the training dataset for each new tree.

3) Naïve Bayes: The base of the probabilistic machine learning technique Naive Bayes is the Bayes theorem, which measures the probability of an occurrence given prior knowledge of conditions that may be relevant to the event. It is well-known for being straightforward, quick, and effective—especially when working with high-dimensional datasets—and is especially well-suited for classification tasks.The premise of feature independence—that is, the idea that every characteristic separately affects the probability of a given class—gives rise to the "naive" component of Naive Bayes. Although this may not always be the case in real-world situations, the method frequently exhibits remarkably good performance and robustness in actual use.

4) Decision Tree: The decision tree is a versatile and simple machine learning technique for applications including regression and classification. The approach works by recursively partitioning the dataset into subsets based on the most significant attribute at each node. Every split is established by assessing a criterion, which is frequently based on metrics such as variance reduction for regression jobs and Gini impurity or information gain for classification tasks. The tree structure is made up of nodes, each of which symbolizes a decision made in response to a certain aspect. The branches that branch off of a node indicate potential outcomes or more options. The procedure keeps on until a predetermined stopping criterion—like a minimum number of instances in a leaf node or a predetermined tree depth—is satisfied.

5) Random Forest: Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputs the mode of the classes (classification) or the average prediction (regression) of the individual trees. It is a powerful and widely used algorithm known for its high accuracy, robustness, and ability to handle complex datasets.

D. Developing Regression Model:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

IV. RESULTS

Tables I, II, and III, which present the outcomes of the prediction models fed with datasets from the Kaggle are summarized below It can be shown from Tables. Accuracy precision, Recall F1Score is mentioned in the table.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	95.18	85.63	61.82	71.88
SVM	94.52	99.67	36.02	52.92
Decision Tree	97.05	100	65.49	79.15
Random Forest	97.28	100	68.44	81.26
Naïve Bayes	90.21	38.65	24.91	30.29

Out of all the models that have been described thus far, the Random Forest Model performed the best at predicting sales. While the graph and Random Forest Technique appear to be quite similar, the accuracy of the former is higher, with all three cases achieving an accuracy of 97.28% because all data points nearly fall on the best of fir. This is because, in order to deliver superior performance, both of these strategies use an ensemble model of learning that averages the results from multiple decision trees.

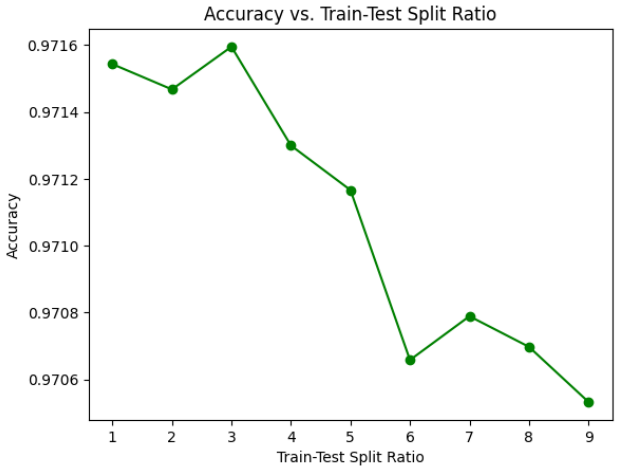


Fig 5: For Different Ratios of Train-Test(Random Forest)

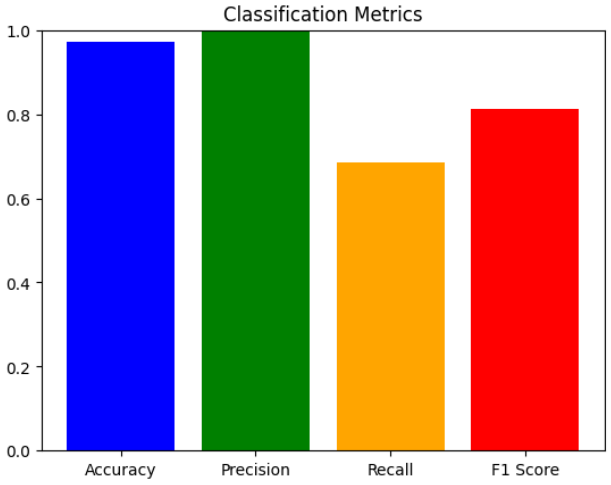


Fig 6: Metrics

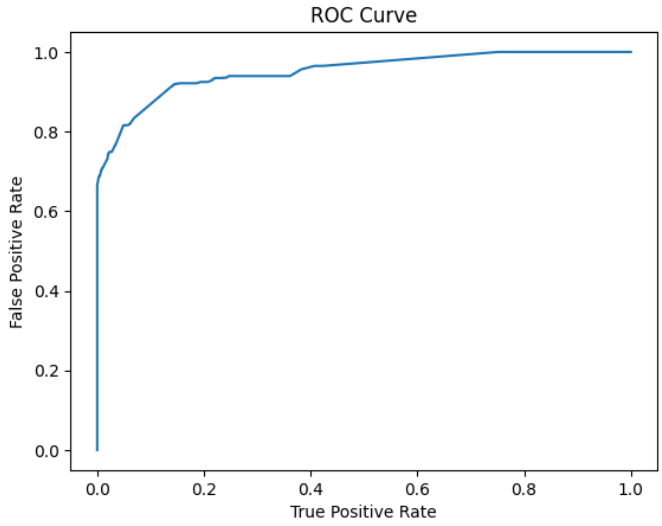


Fig 7: ROC Curve(Random Forest)

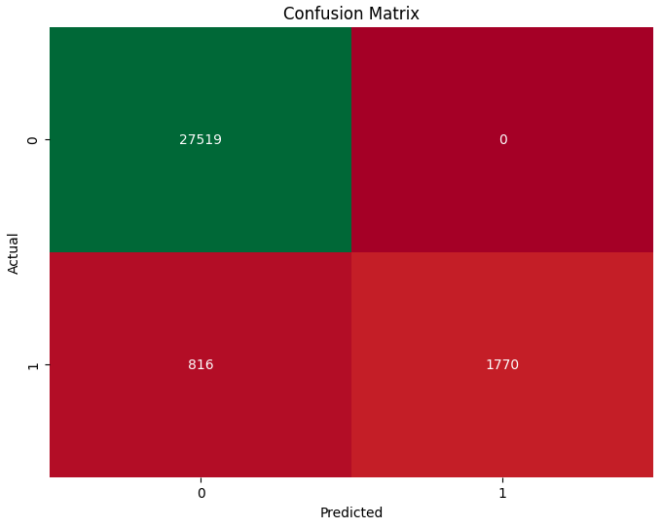


Fig 8: Confusion Matrix(Random Forest)

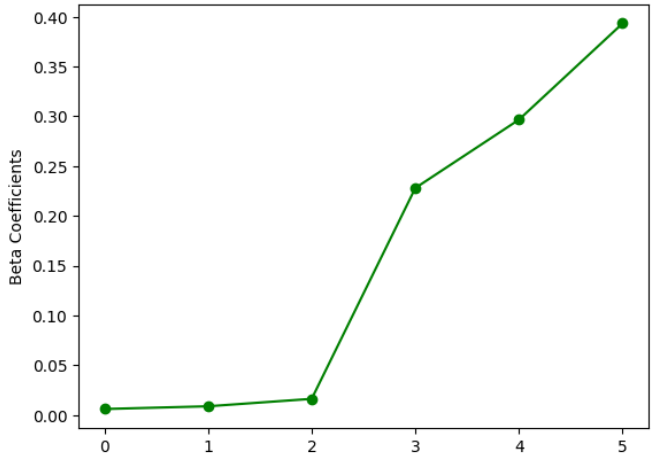


Fig 9: Beta Coeeficients(Logistic Regression)

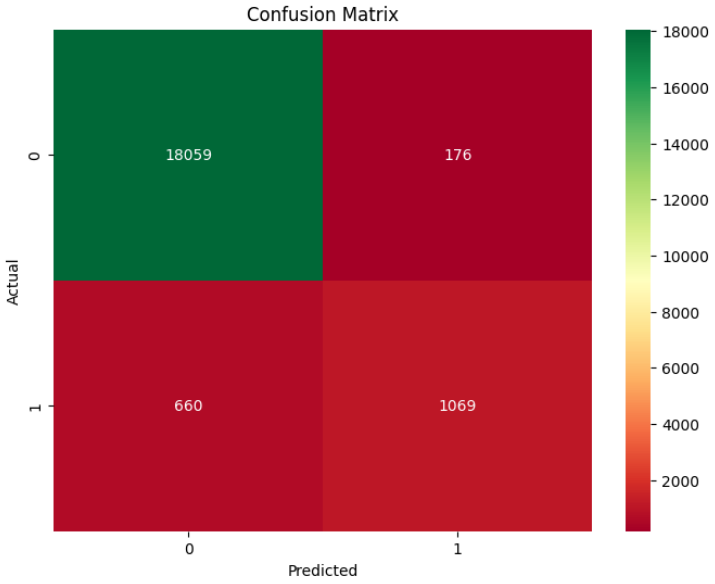


Fig 10: Confusion Matrix(Logistic Regression)

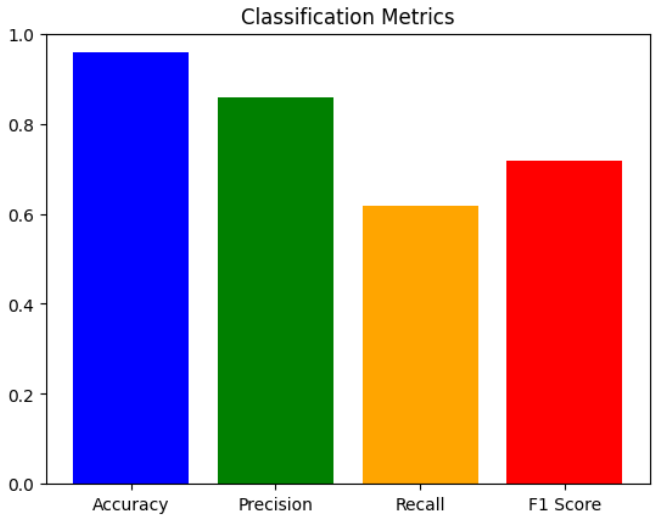


Fig 11: Metrics(Logistic Regression)

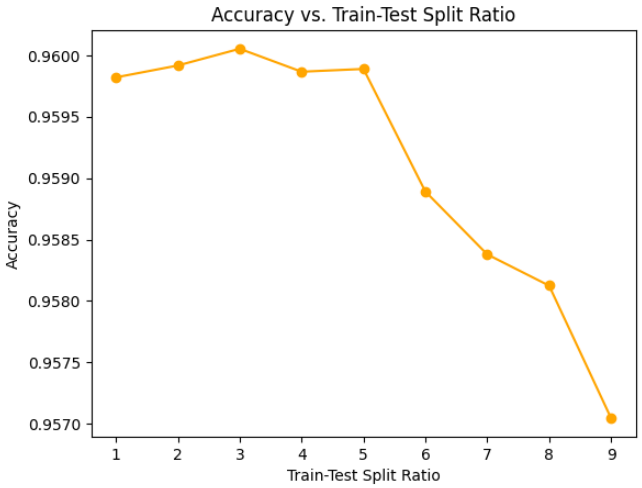
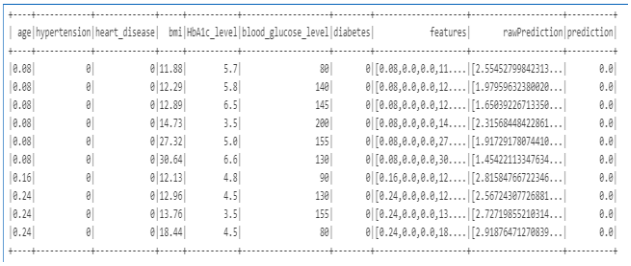


Fig 12: Different Train-Test Ratio



age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes	features	rawPrediction	prediction
0.08	0	0	11.88	5.7	80	0	[0.08,0.0,0.0,11....	[2.55452799842313...	0.0
0.08	0	0	12.29	5.8	140	0	[0.08,0.0,0.0,12....	[1.97959632380020...	0.0
0.08	0	0	12.89	6.5	145	0	[0.08,0.0,0.0,12....	[1.65039226713350...	0.0
0.08	0	0	14.73	3.5	200	0	[0.08,0.0,0.0,14....	[2.31568448422861...	0.0
0.08	0	0	27.32	5.0	155	0	[0.08,0.0,0.0,27....	[1.91729178074410...	0.0
0.08	0	0	30.64	6.6	130	0	[0.08,0.0,0.0,30....	[1.45422113347634...	0.0
0.16	0	0	12.13	4.8	90	0	[0.16,0.0,0.0,12....	[2.81584766722346...	0.0
0.24	0	0	12.96	4.5	130	0	[0.24,0.0,0.0,12....	[2.56724387726881...	0.0
0.24	0	0	13.76	3.5	155	0	[0.24,0.0,0.0,13....	[2.72719855210314...	0.0
0.24	0	0	18.44	4.5	80	0	[0.24,0.0,0.0,18....	[2.91876471270839...	0.0

Fig 13: Comparing Models

Model	**Accuracy**
Logistic Regression	95.18
SVM	94.52
Decision Tree	97.05
Random Forest	97.28
Naïve Bayes	90.21

V. CONCLUSION

This study looked at using the PySpark framework to predict diabetes using machine learning methods, particularly Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Naive Bayes.. To identify people at risk of diabetes, an analysis was performed on a heterogeneous dataset that included pertinent health and lifestyle factors.Our comparative analysis's findings identified each algorithm's distinct advantages and traits. As a baseline model, logistic regression demonstrated interpretability and effectiveness. Decision trees were versatile and could capture complicated relationships; however, they needed to be carefully pruned to prevent overfitting. An ensemble of decision trees called Random Forest demonstrated better generalization and robustness. SVM, which is renowned for its ability to handle high-dimensional data, proved useful in identifying non-linear decision boundaries . Despite its naive assumptions, Naive Bayes turned out to be effective and performed well, especially when dealing with high-dimensional datasets. This study highlights the significance of utilizing distributed computing frameworks to handle the rising volume of healthcare data and contributes to the expanding body of knowledge in healthcare analytics. This study highlights the significance of utilizing distributed computing frameworks to handle the rising volume of healthcare data and contributes to the expanding body of knowledge in healthcare analytics.

VI. REFERENCES

[1] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang“Type 2 diabetes mellitus prediction model based on data mining”

[2] Mayo Clinic Q and A: Childhood diabetes, March 31, 2021,06:00 p.m. CDT

[3] Science Saturday: Could regenerative medicine provide a newapproach to diabetes care?, Nov. 28, 2020, 12:00 p.m. CDT

[4] Iancu, I., Mota, M., and Iancu, E. (2008). “Method for the analysing of blood glucose dynamics in diabetes mellitus patients,” in Proceedings of the 2008 IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca. doi: 10.1109/AQTR.2008.4588883

[5] Cox, M. E., and Edelman, D. (2009). Tests for screening and diagnosis of type 2 diabetes. Clin. Diabetes 27, 132–138. doi: 10.2337/diaclin.27.4.132

[6] American Diabetes Association (2012). Diagnosis and classification of diabetes mellitus. Diabetes Care 35(Suppl. 1), S64–S71. doi: 10.2337/dc12-s064

[7] Lee, B. J., and Kim, J. Y. (2016). Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE J. Biomed. Health Inform. 20, 39–46. doi:10.1109/JBHI.2015.2396520

[8] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C.,Ehrman, J., and Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the henry ford exercise testing (FIT) project. PLoS One 12:e0179805. doi: 10.1371/journal.pone.0179805

[9] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Comput. Struct. Biotechnol. J.