Diabetes Disease Progression using three Machine Learning Algorithms

Heba Dorazahi
department of Information and Technology
Technische Hochschule Ostwestfalen-Lippe
Lemgo, Germany
heba.dorazahi@stud.th-owl.de

***Abstract*—Diabetes is one of the most serious diseases in the world and affects a large population. It can be caused by various factors, such as lack of exercise, obesity, age, high blood pressure, lifestyle, etc. People with diabetes are at a higher risk of developing ailments such as strokes, heart disease, vision issues, kidney failure, and so on. Diabetes and its related health problems can be avoided with treatment and early detection. In industries, machine learning techniques are used to perform predictive analytics on big data. In this study, the performance and accuracy of three machine learning algorithms are evaluated and compared. The objective of the paper is to analyze three algorithms to reveal which algorithm can predict the disease progression with the highest accuracy.**

***Index Terms*—Machine learning algorithms, Diabetes prediction, Compare, accuracy**

I. MOTIVATION

The problem is a regression type problem in which the dataset of three popular machine learning algorithms linear regression, random forest regression, and decision tree analyzed to predict the progression of diabetes. These machine learning algorithms are analyzed and compared to reveal the model for diabetes progression with higher accuracy.

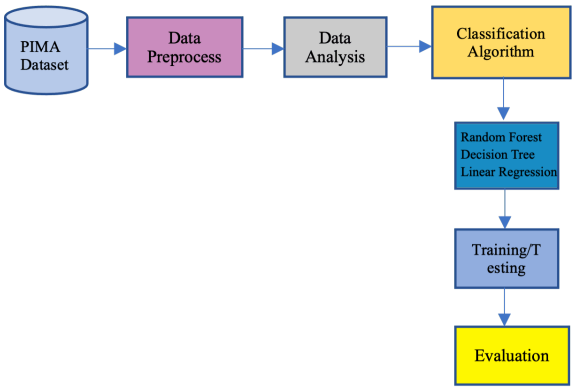

Fig. 1. Flow chart

Diabetes Mellitus, also known as diabetes, is currently one of the most rapidly spreading diseases. Diabetes is a metabolic condition defined by an abnormal rise in blood sugar levels caused by insulin insufficiency.

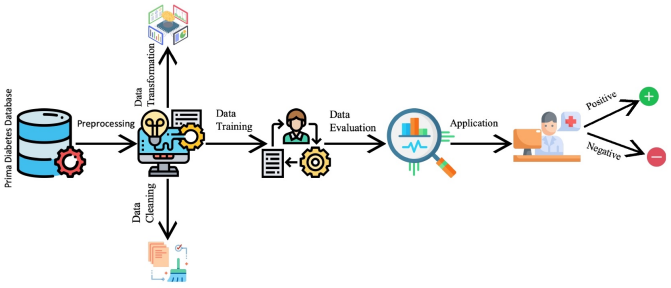

Fig. 2. Schematic Representation

In the current diagnosis procedures, diabetes-related medical errors are common, such as false-positive and false-negative diagnosis errors. In false-positive errors, the patient's test reports show diabetes, but in reality, they are not diabetic. In false-negative errors, the patient's test reports show that they are not diabetic, but in reality, they are diabetic.

Medical experts require a reliable framework for a more accurate diabetes analysis. Machine learning techniques are ideal for analyzing data because it is more beneficial to identify the disease and prevent it from occurring than to cure it. The disease prediction system has a higher accuracy for predicting diseases. It can provide valuable insights to physicians for disease diagnosis. In this study, machine learning algorithms are compared to analyze which methodology is best suited for predicting the diabetes progression. This approach will help physicians and health professionals in the prediction of diabetes disease progression.

Machine learning algorithms such as Logistic Regression, Naive Bayes, KNN, and Support Vector Machines have been analyzed by various researchers for predicting diabetes disease. This paper aims to discuss linear regression, random forest regression, and decision tree to determine if these

algorithms provide better accuracy for diabetes disease progression prediction. Data that contains outliers does not provide the desired results. This study helps to solve the problems associated with diabetes prediction by eliminating outliers and missing data to provide a more accurate result.

II. STATE OF THE ART

Machine learning algorithms have become increasingly significant in the medical field in recent years, particularly for diagnosing diseases using medical databases. Many companies are employing these techniques to improve medical diagnostics and early disease prediction [1].

Diabetes is a leading cause of death, disability, and economic loss worldwide [2]. According to the World Health Organization, diabetes mellitus affects roughly 346 million people globally. It is a disease in which blood glucose levels are not properly managed, increasing the risk of heart attack, kidney damage, and blindness [3].

Several researchers used machine learning techniques to predict diabetes using the PIMA Indian diabetes dataset.

Different machine learning algorithms are used and evaluated on the dataset, such as Adaboost, K Nearest Neighbor (KNN), Decision Tree (DT), Naive Bayes (NB) and Random Forest (RF) to predict the progression of diabetes in a patient [4].

There are several types of machine learning algorithms that are used to classify the dataset. As shown in Fig.3, there are supervised and unsupervised learning techniques. Regression and classification are two types of supervised learning, and clustering is an unsupervised learning type.

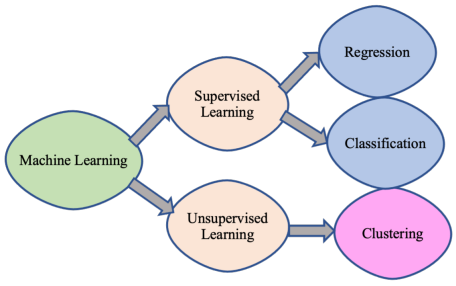

Fig. 3. Machine Learning Algorithms

The Decision Tree is one of the most popular machine learning algorithms for diabetes prediction. One study examined a predictive model to predict type II diabetes in order to help in the screening of diabetes. A total of 22,398 records were used for data preprocessing and preparation. The results showed that the model's precision and accuracy were 71.7 and 97.6 percent, respectively. Another classification algorithm that generates many decision trees is the Random Forest [5].

To determine the optimal size of the forest, several random forests are created with various numbers of trees so that an optimal random forest is obtained for classifying diabetic patients [6].

Random forest performance is evaluated by using its accuracy. The precision and accuracy of the model are 86.11 and 83.11 percent, respectively [7].

In one study, a random forest algorithm was used to predict the disease progression in amyotrophic lateral sclerosis (ALS) patients. In this study, the medical records of patients over a three-month period were used. The accuracy of the model was considered among the third-best models for predicting disease progression [8].

In machine learning, K-Nearest Neighbor is a common non-parametric classification algorithm because it is conceptually straightforward [2]. K-Nearest Neighbor (KNN) is a supervised machine learning technique that generates new data classifications based on the nearest neighbors to the K value.

This technique is based on the shortest the distance between the new data and the K nearest neighbors. A K-Nearest Neighbor algorithm has a 91 percent accuracy rate [9].

In a study, KNN algorithm was applied to predict the severity level of Parkinson's disease. The extracted features were used to build basic classifiers with the K-nearest neighbor algorithm. KNN achieved an accuracy of 95.02% for predicting the severity level of the disease [10].

Support Vector Machine was used for the classification of multiple sclerosis disease. The algorithm analyzed the disease data to classify patients based on the disease condition, whether the disease is worsening or not worsening. The model was 62 percent accurate [11] [12].

Naive Bayes is a machine learning classification algorithm that can handle missing values during classification. The fundamental idea behind Naive Bayes is that it works by applying the Bayes theorem to conditional probability. The data selection of naive bayes is based on numeric and category. The numerical data is based on the mean and variance, and categorical data is based on probability [13].The result of the algorithm indicated that the accuracy of the model was 79.687 percent [9].

III. SOLUTIONS

Long-term patient disease trajectories are a gold mine for a better understanding of disease progression and prediction. The ability of healthcare organizations to use predictions about the future health of individuals to make appropriate interventions is becoming increasingly crucial. A prognosis is a forecast of how an illness will progress after it has begun. It relates to the several possible outcomes of a disease as well as the likelihood of these outcomes occurring.

The patient's medical records can be used to better anticipate the patient's final result. Prognostic factors are traits that can be used to predict a patient's outcome. Prognostic factors do not have to cause outcomes, they merely have to be substantially related to them in order to

predict their progression [14].

A. Problem Description

A disease progression model examines characteristics of the data and trains a model to forecast future data and create meaningful output.

The objective of this study is to build a model for the prediction of disease progression with the use of machine learning models, such as linear regression, random forest regression, and decision tree regression. The dataset for this analysis is from Trevor Hastie's LARS software page [15].

B. Formalization

The diabetes dataset has ten baseline independent variables, age, sex, body mass index, average blood pressure, six blood serum measurements, and one dependent variable, namely disease progression, one year after baseline. A dataset containing data on the diabetes disease of 442 patients has been used in this study. The Python implementation of the three algorithms is discussed in this section.

The first step is importing essential Python libraries. After that, the dataset is loaded by using the read_csv() function from the pandas module. The target and feature variables are separated. To divide the data into a train set and a test set, use the train_test_split() module. After splitting the data, the machine learning regressor is fitted to the dataset. The R2 score is used to show the accuracy of the model.

In this study, three machine learning algorithms were used for regression analysis. Fig. 4 shows the general structure of the predictive model of machine learning.

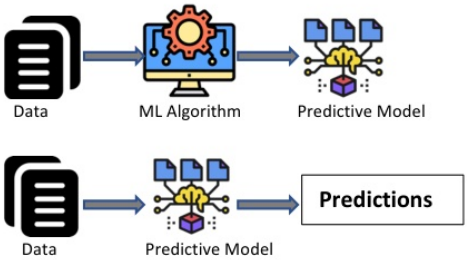

Fig. 4. Training and Testing phase of ML Predictive Model

C. Algorithm

The Decision Tree Regressor, Linear Regression, and Random Forest Regressor are the machine learning algorithms used in this regression analysis research work.

The decision tree is one of the most widely used algorithms for supervised learning. It can tackle both regression and classification problems. In the tree-structured classifier, there are three types of nodes. The root node is the first node and can be further divided into interior nodes and leaf nodes. This algorithm is useful for resolving decision-making problems [16].

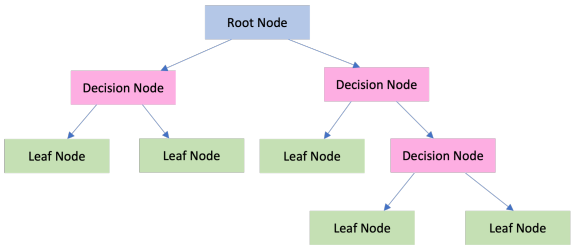

Fig. 5. Decision Tree Structure

Random Forest Regression is a supervised machine learning algorithm. It is a technique that combines several machine learning algorithms to get a more precise prediction as compared to a single machine learning model. A forest is comprised of several trees. Random forests generate decision trees, evaluate predictions from each tree, and choose the best possible solution [16].

Another machine learning algorithm is linear regression. In linear regression, the independent variables or features are linearly related to the dependent or target variable.

Y = bX + a

Algorithm Description:

1. Importing necessary libraries such as, Pandas, Numpy, Matplotlib, Sklearn, and Seaborn.
2. Analysing correlation between dataset with sns.heatmap()
3. Splitting the data between x and y
4. Applying feature scaling
5. Splitting data training and testing
6. Building the classification algorithm
7. Making Predictions by using regression algorithm
8. Making prediction on test data by using regression algorithm
9. Evaluating the model by using Mean squared error and R2_score.

MSE = $\sum_{i=1}^{D}(x_i - y_i)^2$

$r2\_score = 1 - (RSS/TSS)$

RSS is sum of squares of residuals

TSS = total sum of squares

D. Implementation and Analysis

The data preprocessing technique is used prior to the application of machine learning algorithms. Data preprocessing is an approach to preparing datasets for predictive analysis with the use of machine learning algorithms. The preprocessing method involves data cleaning and removing outliers and null values from the data. It is also examined to see whether any data points are completely out of range. Outliers have

Algorithm 1 Regression Algorithms

1) Initialize the feature variables $x = AGE, SEX, BMI, \ldots, S5$
   $Initialize\ the\ target\ variables\ y = Disease\_progression$
   a) Fit the regressor to the training data using features variable
   b) Linear Equation

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon}_i \quad (1)$$

   c) Compute

$$\mathrm{M}SE = (1/N)(\sum_{i=1}^{D}(x_i - y_i)^2).$$

   d) Compute

$$\mathrm{r2\_}score = 1 - (RSS/TSS).$$

2) Output
   Linear Regression $\longrightarrow MSE = 3448.39$
   $R2\_score = 0.358$
   $RandomForest\ Regression \longrightarrow MSE = 3522.02$
   $R2\_score = 0.344$
   $DecisionTree\ Regression \longrightarrow MSE = 4131.41$
   $R2\_score = 0.21$

a tendency to degrade the model's efficiency. As a result, data is examined to check if it is normalized.

In Fig. 6, the histogram shows disease progression in a number of patients. It can be observed from the histogram that as the disease progresses, the number of patients decreases.

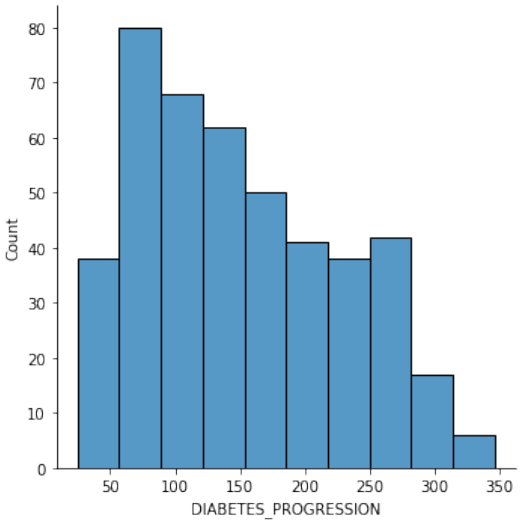

Fig. 6. Disease Progression of patients one year after baseline

In Fig. 7, the histogram shows that for more than 70 patients, the blood pressure value is between 85 and 95. A small number of patients have a blood pressure level of 130 or higher.

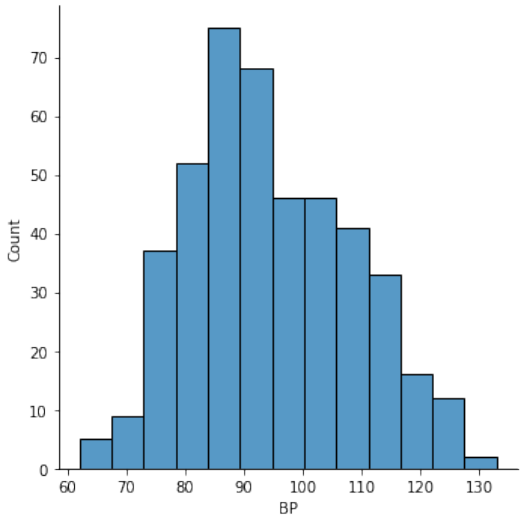

Fig. 7. Blood pressure versus number of patients

Fig. 8 shows the correlation between the variables. A value closer to one shows a positive correlation. From the correlation heatmap, it can be observed that BP, BMI, and glucose are highly correlated with disease progression.

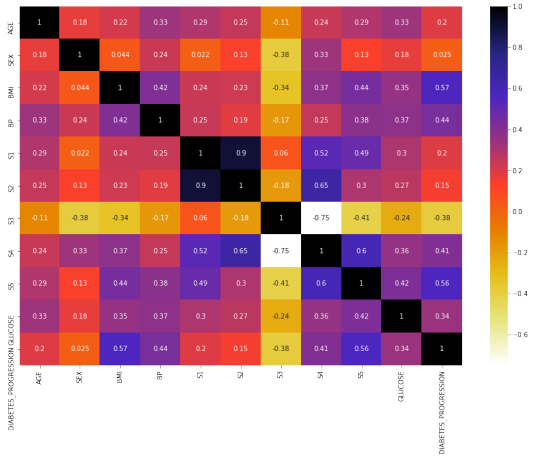

Fig. 8. Correlation Heat map among all the feature variables and target variable

In Fig. 9 scatter plot, 1 shows the male gender, and 2 represents the female gender. From the plot, it can be observed that the BP and glucose levels of males are higher than females. The BMI level of females is higher than males.

IV. EXPERIMENTS AND RESULTS

After preprocessing and regression analysis, the predictive performance of multiple models is compared and tested. In the linear regression scatter plot, which is shown in Fig. 10.,

Fig. 9. Scatter plot shows the relationship of each variable in different gender

the graph shows actual disease progression versus predicted disease progression. In this graph, some features are dropped to evaluate the model's accuracy level. Selected features are highly correlated with disease progression. Selected features are bmi, bp, S4, and S5. This variable's correlation level with disease progression is more than 40 percent. The accuracy score of linear regression for the selected feature is 35.1% . After including all features such as blood serum measurements, age, sex, and glucose level, the model accuracy score is 35.8% which is shown in Fig. 11. From the results, it is observed that non-correlated features do not have an effect on the accuracy level.

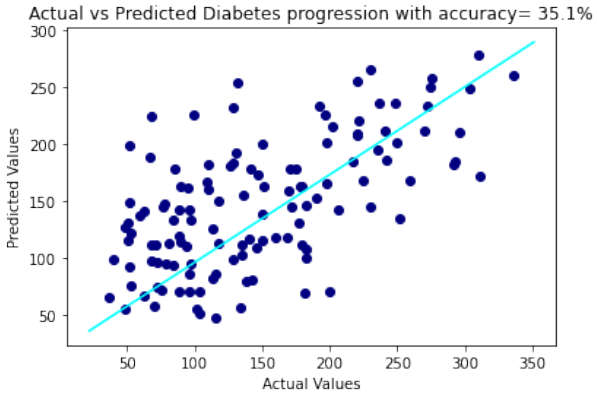

Fig. 10. Scatter plot of Actual versus Predicted regression model

In the random forest regression scatter plot, which is shown in Fig. 12., the graph shows actual disease progression versus predicted disease progression. By excluding some features, The accuracy score of the model was 31%. After including all features, the accuracy increased to 34%, which is shown in Fig.13. From the results, it is observed that reducing the

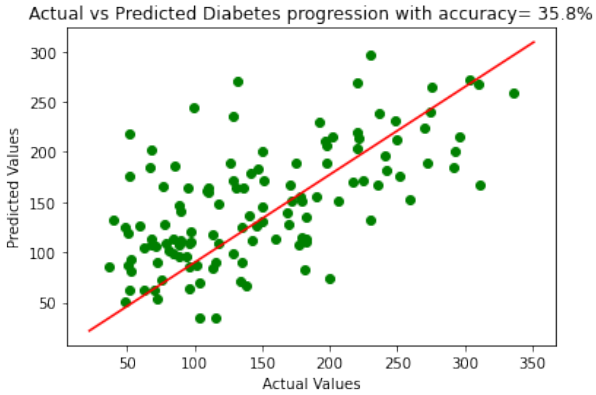

Fig. 11. Scatter plot of Actual versus Predicted regression model

features has a slight effect on model efficiency.

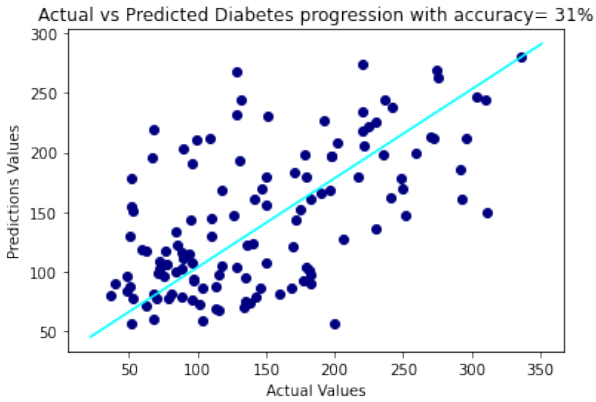

Fig. 12. Scatter plot of Actual versus Predicted regression model

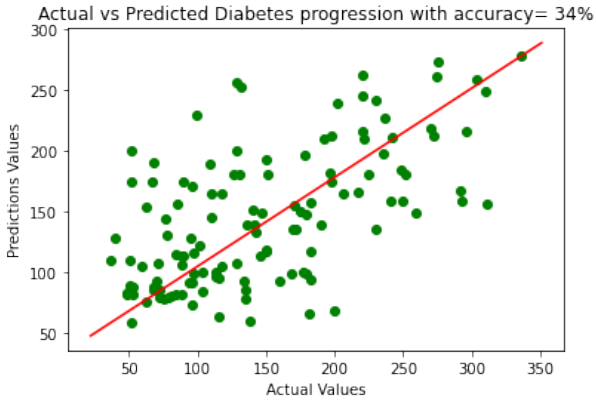

Fig. 13. Scatter plot of Actual versus Predicted regression model

In decision tree regression, the depth of the tree is calculated to get the maximum tree depth for more accurate results. The maximum depth of tree for all feature variables is measured as 2, and the accuracy of the model is 21% which is shown in Fig. 15.

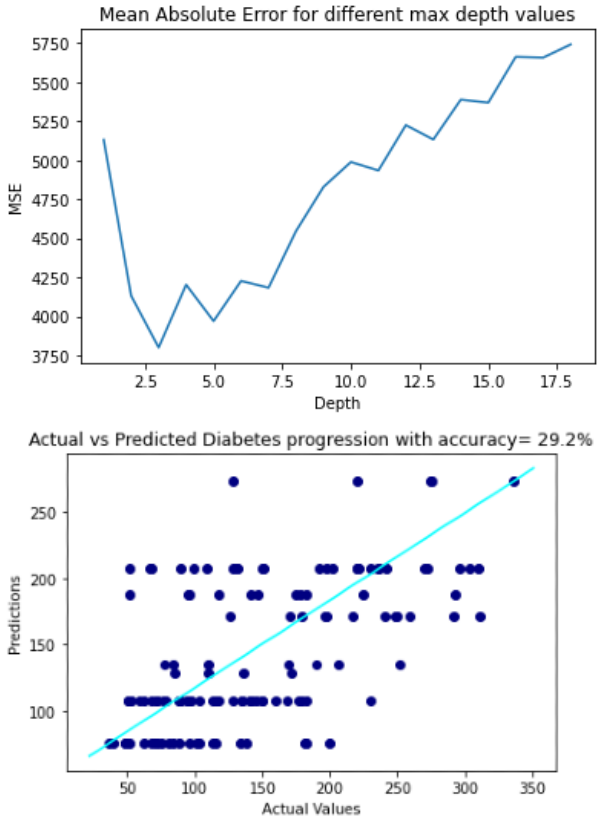

Fig. 14. Scatter plot of Actual versus Predicted regression and maximum tree depth

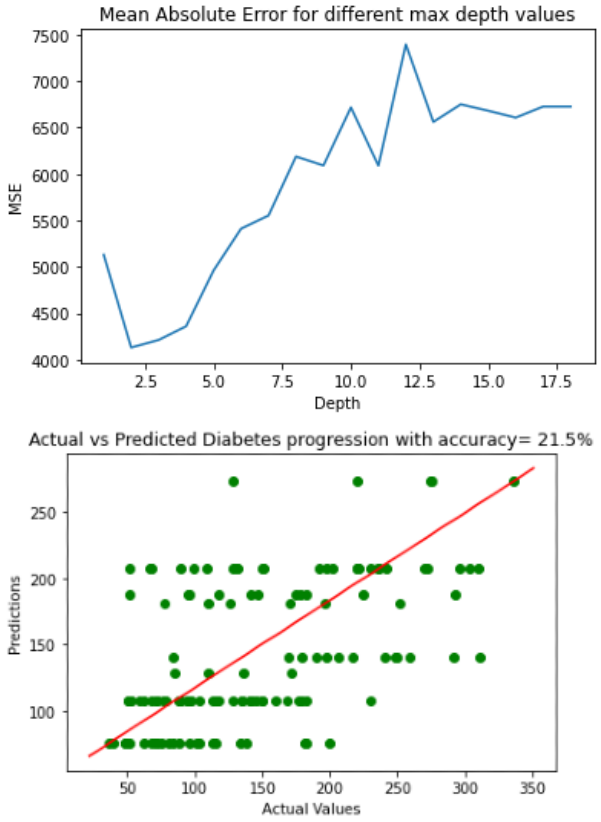

Fig. 15. Scatter plot of Actual versus Predicted regression and maximum tree depth

In Fig. 14, only those features that are highly correlated with disease progression are selected. The maximum tree depth for selected features in 3 and the model accuracy increased to 29.2%. Results show that the model accuracy is increased after taking only those features which are highly correlated with the disease progression.

After conducting the analysis, the performance parameters of the three algorithms are tabulated in Table I. The linear regression model with selected features produced an RMSE value of 59.00 units, which is one unit higher compared to all the selected features. The R-squared values of the selected feature model and all feature models are 0.351 and 0.358, respectively, which are approximately the same values.

The random forest model with selected features produced an RMSE value of 60.00, which is one unit higher compared to all the selected features. The R-squared values of the selected feature model and all feature models are 0.31 and 0.344, respectively. The R-squared value in all features shows better accuracy compared to selected features.

The decision tree model with selected features produced an RMSE value of 61.65, which is 3 units lower compared to all selected features. The R-squared values of the selected features model and all features models are 0.292 and 0.21, respectively. The R-squared value in selected features shows better accuracy compared to all features.

TABLE I
PERFORMANCE ANALYSIS OF ALGORITHMS

b)

Methods Used		MSE	R-Squared	RMSE
Reduced Features	Linear Regression	3481.14	0.351	59.00
	Random Forest	3690.35	0.31	60.74
	Decision Tree	3800.73	0.292	61.65
All Features	Linear Regression	3448.39	0.358	58.72
	Random Forest	3522.02	0.344	59.34
	Decision Tree	4131.41	0.21	64.27

When comparing the three algorithms, it is observed that linear regression shows better accuracy in both methods (selected features and all features) as compared to the other two algorithms. A random forest shows better accuracy than a decision tree.

V. SUMMARY AND OUTLOOK

Three regression algorithms were used for the predictive analysis of diabetes progression one year after the baseline. A dataset containing data on the diabetes disease of 442 patients with respect to 10 input factors provides insight into the disease progression. The disease progression method helps the doctors offer precise medical advice. It is an effective way of

understanding the lifestyles of individuals. This model ignores any changes in the lifestyle of a person or any other blood sample readings.

In the results, it was observed that the Linear Regression algorithm provides slightly higher accuracy as compared to Random Forest and Decision Tree. Random Forest provides better accuracy than the Decision Tree.

Three algorithms do not seem effective in the prediction of disease progression. The dataset that has been used for this analysis is real data, and it has some non-linear properties, so it does not provide high efficiency for Linear Regression. Random Forests and decision trees are more effective at predicting whether or not a person is diabetic. They do not provide high efficiency in the disease progression.

VI. Conclusion

Three machine learning models were used to compare the accuracy of disease progression. Linear Regression and Random Forest provided almost the same accuracy level, with Decision Tree providing slightly lower accuracy as compared to the other two algorithms. Decision Tree and Random Forest provide more accurate results by using numeric data. Many studies have shown that using numeric and categorical values for analyzing the data provides more precise results, such as predicting whether a person has diabetes or not. In this study, I have analyzed the algorithms by using continuous data. As it was observed, the accuracy level of Linear Regression is higher compared to Random Forest and Decision Tree. These algorithms were analyzed by reducing the features of the data and using only those features that were more correlated with the disease progression. By analyzing reduced feature methods, it was observed that reducing the feature variables was only effective in decision tree algorithm. The decision tree accuracy increased from 21 percent to 29 percent by dropping some features.

References

[1] I. Ibrahim and A. Abdulazeez, "The role of machine learning algorithms for diagnosing diseases," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 10–19, 2021.

[2] M. Panwar, A. Acharyya, R. A. Shafik, and D. Biswas, "K-nearest neighbor based methodology for accurate diagnosis of diabetes mellitus," in *2016 sixth international symposium on embedded computing and system design (ISED)*. IEEE, 2016, pp. 132–136.

[3] K. Saxena, Z. Khan, and S. Singh, "Diagnosis of diabetes mellitus using k nearest neighbor algorithm," *vol*, vol. 2, pp. 36–43, 2014.

[4] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, 2021.

[5] N. K. Kumar, D. Vigneswari, M. V. Krishna, and G. P. Reddy, "An optimized random forest classifier for diabetes mellitus," in *Emerging Technologies in Data Mining and Information Security*. Springer, 2019, pp. 765–773.

[6] S. Benbelkacem and B. Atmani, "Random forests for diabetes diagnosis," in *2019 International Conference on Computer and Information Sciences (ICCIS)*. IEEE, 2019, pp. 1–4.

[7] A. Choudhury and D. Gupta, "A survey on medical diagnosis of diabetes using machine learning techniques," in *Recent developments in machine learning and data analytics*. Springer, 2019, pp. 67–78.

[8] T. Hothorn and H. H. Jung, "Randomforest4life: a random forest for predicting als disease progression," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 15, no. 5-6, pp. 444–452, 2014.

[9] A. Nurhayati, "Implementation of naive bayes and k-nearest neighbor algorithm for diagnosis of diabetes mellitus," in *Proceedings of the 13th International Conference on Applied Computer and Applied Computation Science*, 2014.

[10] H. Zhao, R. Wang, Y. Lei, W.-H. Liao, H. Cao, and J. Cao, "Severity level diagnosis of parkinson's disease by ensemble k-nearest neighbor under imbalanced data," *Expert Systems with Applications*, vol. 189, p. 116113, 2022.

[11] A. Tousignant, P. Lemaître, D. Precup, D. L. Arnold, and T. Arbel, "Prediction of disease progression in multiple sclerosis patients using deep learning analysis of mri data," in *International Conference on Medical Imaging with Deep Learning*. PMLR, 2019, pp. 483–492.

[12] C. Cavaliere, E. Vilades, M. Alonso-Rodríguez, M. J. Rodrigo, L. E. Pablo, J. M. Miguel, E. López-Guillén, E. M. Morla, L. Boquete, and E. Garcia-Martin, "Computer-aided diagnosis of multiple sclerosis using a support vector machine and optical coherence tomography features," *Sensors*, vol. 19, no. 23, p. 5323, 2019.

[13] V. V. Vijayan and C. Anjali, "Prediction and diagnosis of diabetes mellitus—a machine learning approach," in *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE, 2015, pp. 122–127.

[14] J. Futoma, M. Sendak, B. Cameron, and K. Heller, "Predicting disease progression with a model for multivariate longitudinal clinical data," in *Machine Learning for Healthcare Conference*. PMLR, 2016, pp. 42–54.

[15]

[16] M. M. K. Sabariah, S. A. Hanifa, and M. S. Sa'adah, "Early detection of type ii diabetes mellitus with random forest and classification and regression tree (cart)," in *2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*. IEEE, 2014, pp. 238–242.