Feature Selection for determining Diabetes Risk Factors (6 variations)

*School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg

Phindulo Ezekiel Makhado
Big Data Analytics Honors Student
University of the Witwatersrand
Johannesburg, South Africa
1832463@students.wits.ac.za

Supervised by: Dr Pravesh Ranchod
School of Computer Science and Applied Mathematics,
University of the Witwatersrand
Johannesburg, South Africa
Pravesh.Ranchod@wits.ac.za

***Abstract*—The undetectable impact that diabetes disease has on South Africa's financial system makes it imperative to make use of technological advancements to considerably aid in attempts to try to reduce the expenditures that diabetes has on South Africa's public health sector. Although there is no known cure for diabetes, there are a number of measures that can be used to lessen the damage that the disease causes. Focusing on diabetes' prognosis is the most popular strategy for reducing its negative effects because there is no known cure. The risk of undetected diabetes will be reduced by the use of risk prediction, which incorporates advanced machine learning algorithms with superior feature selection methodologies and implementations. We used a variety of Machine Learning models, such as Logistic Regression, Optimized Random Forest, and Decision tree classifier to predict Diabetes on train-test splits of 75–25%, following thorough data pre-processing and feature selection with a combinatorial approach and a Decision tree Importance approach. For that train-test split which then undergoes the proposed combinatorial feature selection approach, the Decision trees classifier model's accuracy was found to be 84.88% on the top 7 selected data set, the highest of all the models in all top selected data sets.**

***Index Terms*—Data Pre-processing, Machine learning techniques, risk prediction, Diabetes, feature Engineering, Feature Selection**

I. Introduction

Diabetes, One of the most serious diseases affects a large number of people. Diabetes may be brought on by hereditary factors, a lack of activity, advanced age, obesity, a poor diet, high blood pressure, a way of life, etc. High hazards for diabetics include renal disease, heart disease, nerve damage, visual problems, and stroke, among others. Given the current situation, diabetes has developed into a very serious illness in various nations, such as India [12]. Like those in other countries, an estimated 90–95 percent of South Africans who have diabetes typically have Type 2 Diabetes [1]. Diagnosed type 2 diabetics, comprising both diagnosed and undiagnosed patients, cost the public sector between R21.8 billion and R22.7 billion in 2018 [1]. Nearly all cases of type 2 diabetes are projected to cost the public sector R35.1 billion by 2030 [1]. The treatment of type 2 diabetes accounts for about 51% of total anticipated expenses for 2030, while complications account for 49%. The public health system in South Africa is under a lot of stress due to type 2 diabetes [1]. In 2018, treating all common ailments accounted for close to 12% of the total national medical budget [1]. The direct cost will steadily decline as the prevalence increases if the current care system is maintained and case detection is enhanced.

Recent studies have shown that properly investigated feature selection techniques and machine learning algorithms produce accurate results that are consistent with earlier findings. Machine learning is being used more and more in the healthcare industry. There is a greater likelihood that an illness can be prevented and successfully treated, the earlier it is detected. Thus utilizing several carefully implemented feature selection techniques and machine learning algorithms, we can achieve good experimental results in terms of early diagnosis of diabetes.

II. Literature Review

With the advancement of feature selection technique modifications, a number of new methods are being developed to aid machine learning algorithms in accurately diagnosing diabetes.

A. *Diabetes*

One of the illnesses linked to insulin hormones is diabetes. It might have a very detrimental effect when the body reacts to it. Diabetes sufferers can take measures to control their illness and lead healthy lives, but unfortunately, no medication has been shown to be beneficial. This subsection of the chapter discusses diabetes epidemiology. Diabetes is a significant direct and indirect stressor in South Africa. In South Africa, 4.5 million people have diabetes [1]. The costs of diabetes in South Africa's public health system were estimated to be R15 billion in 2021 [1]. In South Africa, diabetes affects an unacceptable number of people. It will take a lot of work to enhance healthcare such that the percentage of patients who achieve their blood pressure and glycemic control objectives and who are correctly monitored for consequences from later therapies can reach levels that have an impact on mortality and morbidity. Multi-level interventions, as well as fundamental

sickness prevention and improved management at the basic healthcare level, are needed [9].

This research focuses on making accurate predictions utilizing various feature selection techniques on popular and advanced machine learning algorithms in diagnostic approaches, with careful consideration of the number of attributes. People can discuss prophylactic treatment in advance with their doctor after receiving favorable forecasts [10]. Patients are in danger for a variety of life-threatening illnesses, from renal failure to heart stroke, if diabetes is not successfully handled [11]. Diabetes is a leading cause of cardiovascular illness, blindness, renal failure, and lower-extremity amputation in many nations [18]. Approximately 200 million people worldwide suffer from diabetes at this time, with women making up more than half of the total [19]. Research is being done to anticipate diabetes in order to diagnose and treat it early [12].

B. Chronic Illness Diagnosis Predictive Models

In this section of the report, studies based on various predictive machine-learning approaches that are used specifically for the diagnosis of chronic illnesses are given. One of the primary guidelines that this project will adhere to is the prediction models constructed in the studies presented in this portion of the report.

In [2], Kumar Dwivedi evaluated a range of machine-learning techniques for predicting diabetes. Methods like classification trees, logistic regression, support vector machines, artificial neural networks, and k-nearest neighbors were used. Specificity, accuracy, recall, precision, FPR, negative predictive value, rate of misclassification, F1 score, and ROC curve are a few of the measures used to gauge how well this system performs. Comparatively, logistic regression had a misclassification rate of 0.22 and the greatest accuracy of 78%. Using Naive Bayes and Logistic Regression, the greater precision of the negative predictive value was found to be 73% and 82%, respectively. And to segregate the data, 10-fold cross-validation is employed.

Decision Trees, k-nearest neighbors, Logistic Regression, Random Forest Trees, Gradient Boosting, and support-vector machines are among the algorithms utilized in the study [4] named Diabetes Prediction using MLT. The results show that random forest, with a classification accuracy of 77%, had the highest performance. According to a method given in [5], support vector machine, random forest, and fully convolutional neural networks are the methods used for deep learning. Deep Learning (76.81%), support vector machine (65.38%), and random forest (83.67%) had the highest accuracy. The results demonstrated that Random Forest performed better at predicting diabetes than support vector machine and Deep Learning methods.

A study of mammography feature selection algorithms for unified breast cancer diagnosis is presented in the article in [6]. A 512x512 pixel patch containing either healthy tissue or breast cancer was subjected to feature extraction, producing curvilinear features, textural features, Gabor features, and multi-resolution features. The characteristics were selected using an adaptive floating search and a GA, and the malignant and healthy regions were distinguished using linear discriminant analysis. SGA and LDA's overall ROC performance is 0.90, CHC and LDA's overall ROC performance is 0.93, and ASFFS and LDA's overall ROC performance is 0.96. Less than 25% of each approach's 86 features were chosen, and at least one feature from each category was selected. Even while some of the four various types of qualities may be connected, they all complement one another. The area under the ROC curve is used to assess performance (Az). Linear Discriminant Analysis and Wrapper Based Selection, Adaptive Sequential Forward Floating Search Feature Selection, and GA for Feature Selection was used in this work to choose features and classify them.

III. PROPOSED METHODOLOGY

The number of people with diabetes has significantly increased since a decade ago. Modern human activity is the main cause of the growth in diabetes. Finding effective feature selection techniques and machine learning models for diabetes prediction that are more accurate than current ones is the main goal of this paper. Different machine learning algorithms for diabetes prediction are used to achieve this. We will outline the procedure we followed below.

A. Description of the Dataset

The Behavioral Risk Factor Surveillance System (BRFSS), a health-related telephone survey that is collected annually by the CDC, provided the data set for this study, which was obtained through the Kaggle repository. Each year, more than 400,000 Americans take part in the survey, which collects data on dangerous behaviors, long-term health conditions, and the use of preventative medicines. The data collection is unbalanced and contains details on 253,680 patients and 21 characteristics. There are 2 classes for the target variable "Diabetic." 1 denotes either diabetes or pre-diabetes, whereas 0 indicates neither.

1. Diabetic : 0 = no diabetes 1 = pre-diabetes or diabetes
2. HighBP : 0 = no high BP 1 = high BP
3. HighChol : 0 = no high cholesterol 1 = high cholesterol
4. CholCheck : 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
5. BMI : Body Mass Index
6. Smoker : Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
7. Stroke : (Ever told) you had a stroke. 0 = no 1 = yes
8. HeartDiseaseorAttack : coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
9. PhysActivity : physical activity in past 30 days - not including job 0 = no 1 = yes
10. Fruits : Consume Fruit 1 or more times per day 0 = no 1 = yes
11. Veggies : Consume Vegetables 1 or more times per day 0 = no 1 = yes
12. HvyAlcoholConsump : (adult men >=14 drinks per week and adult women>=7 drinks per week) 0 = no 1 = yes
13. AnyHealthcare : Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes
14. NoDocbcCost : Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes
15. GenHlth : Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
16. MentHlth : days of poor mental health scale 1-30 days
17. PhysHlth : physical illness or injury days in past 30 days scale 1-30
18. DiffWalk : Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
19. Sex : 0 = female 1 = male
20. Age : 13-level age category 1 = 18-24 0 = 60-64 13 = 80 or older

Fig. 1. Detailed description of the features

In the above further description(Fig. 1.), "1" is to indicate the characteristic's presence, and "0" is to indicate the characteristic's absence. The distribution of the aforementioned characteristics by age across all patients surveyed is depicted in Fig. 2.

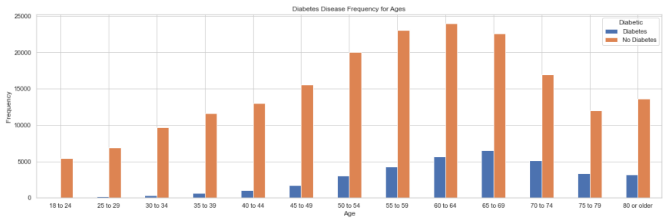

Fig. 2. Variation of Age for each target class

B. *Pre-processing of Data*

Data pre-processing is a crucial machine learning procedure [17]. It improves the effectiveness of machine learning models and the integrity of the data [17]. The procedure comprises cleaning and transforming the raw data in preparation for training and assessing prediction models. Data pre-processing entails the preparation, cleansing, feature selection, handling of missing values, and modification of the data. The resulting dataset from data preparation should be precise and useful for additional data mining techniques. The gathered electronic health records comprised a large dataset with many dimensions. Because the necessary measures were reliant on the participants, it is improbable that all the features were gathered during the medical examination. The missing values were eliminated to address the missing-values issue. Records containing null feature values were omitted from the dataset due to the size of the dataset and to provide quality analysis.

C. *Class Imbalance of Data*

The majority of machine learning algorithms presumptively consider similar prior probabilities for the target classes. But there are numerous instances in which this presumption is incorrect. When working with datasets that contain a class imbalance, the machine learning classifier has a tendency to be more biased towards the majority class, misclassifying the minority class. In such issues, the majority of the samples belong to one class while the remainder mostly belongs to the other class [8]. In our dataset, cases in the normal class made up 84.71% of the data, and then those with diabetes made up 15.29%. These two classes were distributed unevenly, which could have caused the prediction model to perform poorly when making predictions about the minority class [8]. The problem was addressed using the majority under-sampling and synthetic minority over-sampling (SMOTE) approaches.

D. *Feature Selection Techniques Implementation*

Choosing a subset of the dataset's most important traits to describe the target variable is the process of feature selection [18]. It makes machine learning problems more efficient in terms of computation time, generalization performance, and interpretational concerns. Filter-based, wrapper-based, and embedded types of feature selection strategies are the categories used to group them. Filter-based techniques eliminate features based on predetermined standards. Wrapper-based strategies analyze and rank characteristics using a modeling process that is thought of as a "black box." The embedded approaches employ feature selection methods based on both random forest (RF) and the least absolute shrinkage and selection operator (LASSO) [18]. This section describes a data-driven methodology for selecting variables to predict the likelihood of developing diabetes using statistical and machine-learning techniques. The dataset produced using the aforementioned techniques comprised both categorical entities from the questionnaire replies and numerical variables from the diagnostic results. Finding a collection of ideal traits that could effectively identify the two groups is the goal of the feature selection process.

*1) **Sequential Forward Feature Selection**:* The family of greedy search algorithms, which also includes sequential feature selection methods, first reduces a d-dimensional feature space into a k-dimensional feature subspace. By eliminating irrelevant information, the goal is to choose the subset of attributes that are most pertinent to the task at hand while increasing processing performance (that acts as noise). The sequential forward selection procedure involves carrying out the subsequent steps to find the N features that will fit in the K-features subset that are the most appropriate out of all of them.

*2) **Sequential Backward Feature Selection**:* The sequential backward selection strategy reduces the initial feature subspace's dimensionality from N to K features with the least possible impact on model performance in an effort to boost computing efficiency and reduce generalization error. The key idea is to successfully remove features from the N-feature given features list in order to reach the list of K features. At each level, the element that causes the smallest performance loss gets eliminated. The algorithm used to find features is a combinatorial search, where a subset of features is chosen from a combination and given a score before being compared to other subsets.

*3) **combinatorial feature selection approach(bi-directional elimination)**:* It is the combination of both Backward Elimination and Forward selection. When there are correlations between variables, the combinatorial feature selection method is essentially a backward elimination process with the option to delete a specified variable at each stage. As depicted in Fig. 3. In this combination of a method, also known as the floating search, the Forward Selection and Backward Elimination methods are initially applied to the initial feature set. In the second stage, two subsets are combined into a single pool, and the same number of features are chosen from each subset in accordance with the majority of votes. If two features receive an equal number of votes, the feature with the greater pattern categorization performance rate when the other characteristics are integrated will prevail. The final feature set will be chosen in the third step using a laborious process based on the features that have received the highest votes. The voting resulted in the creation of a good subset of all features, larger than the targeted subset size but considerably smaller than the original entire set.

*4) **Overall Regularized Trees with Overall Feature Importance**:* With reference to both strong and weak classifiers,

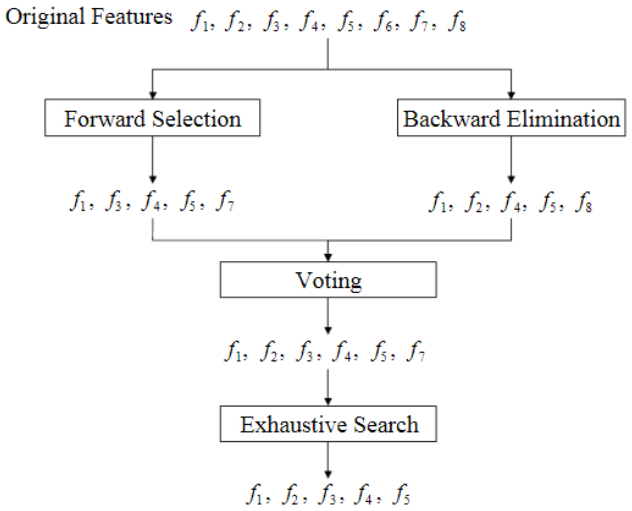

Fig. 3. Combined Feature Selection procedure

the regularized trees can choose high-quality feature subsets. Because tree models can naturally manage category and numerical variables, missing values, differences in scale across variables, interactions, nonlinearities, etc.

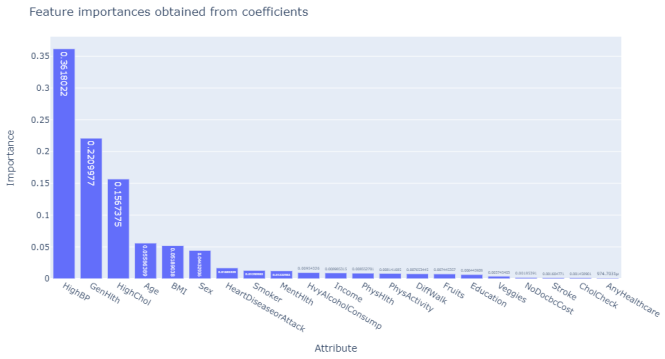

Fig. 4. Overall Regularized Trees with Overall Feature Importance

There are a few regularization parameters in a tree that we can use to control the size of the tree, like: max depth, Min sample split, min sample leaf, max feature size and optional criterion (default: "gini") or Select an attribute selection criterion. In Fig. 4., we utilized the max depth = "15" and criterion =" entropy", to obtain that feature importance diagram.

E. Machine Learning Algorithms

*1) **Logistic Regression**:* It is common practice to assess the likelihood of a binary response based on at least one prediction using the well-known supervised machine learning (ML) classification technique known as logistic regression. It is possible to have both continuous and discrete ones. It is applied to the categorizing of data. Additionally, while determining whether a patient has diabetes or not, Logistic Regression frequently splits the data into binary groups (0 and 1). When analyzing the correlation between predicted and desired values, this model's objective is to offer the best fit. The sigmoid function, which restricts the output to either 0 or 1, is used to predict the likelihood of the output.

*2) **Random Forest**:* A well-known machine learning approach called Random Forest is applied to classification and regression problems. To improve the performance of the model, the bagging ensemble learning technique combines a number of classifiers. By generating more data through training, it is utilized to lower variation. The approach operates during training and provides the individual tree classification mode as an output. It uses a combination of many decision trees.

*3) **Decision Trees**:* A decision tree is a simple visual aid for identifying samples. The data is continuously split into different groups based on the parameter in this supervised machine-learning technique. Classification and regression issues can be solved by deterministic tree analysis. A decision tree is incrementally created together with the decision tree algorithm, which divides a dataset into more manageable divisions. A decision tree is made up of leaf nodes, which are the terminal nodes that foretell the outcome, edges/branches that correlate to the results of tests and connect to the next node or leaf, and nodes that assess the value of various attributes.

*4) **Gaussian Naive Bayes classifier**:* This classifier is employed when the predictor values are continuous and are assumed to follow a Gaussian distribution. The generalization of naive Bayes is called gaussian naive Bayes. The Gaussian or normal distribution is the most straightforward to implement among the several functions used to estimate data distribution, as you only need to figure out the mean and standard deviation for the training data.

F. Predictive Model Implementation

The crucial stage is model building. The aforementioned machine learning methods are all used early on to predict diabetes. The consolidated implementation of the predictive model is described below.

- Import both the Behavioral Risk Factor Surveillance System data set and the necessary libraries.
- Preprocessing the data will allow you to tidy it up and impute missing values.
- Eliminating the class imbalance in the data set
- Divide the data into 75%-25% for Training and Testing, accordingly.
- Implement the proposed feature selection techniques, namely: Sequential Forward Feature Selection, Sequential Backward Feature Selection, combinatorial feature selection approach(bi-directional elimination), and Overall Regularized Trees with Overall Feature Importance.
- Create different data sets for the top selected sets from the two proposed feature selection techniques.
- Train the Top selected and original dataset on the following machine learning algorithms: Logistic Regression, Random Forest, Decision Trees, and Gaussian Naive Bayes classifier.

- Create the classifier for the aforementioned machine learning method using training data.
- Test the Classifier for the aforementioned machine learning algorithm using the test set.
- Validate the trained models using k-Cross Validation, and then furthermore use the evaluation metrics to check the trained model's performances.
- Compare the effectiveness of the experimental findings for each classifier.
- After reviewing the data, decide which algorithm performs the best.

IV. RESULTS ANALYSIS

A machine with an AMD Ryzen 5 5600H processor running at 3.30 GHz, Radeon Graphics, and 16 GB of RAM was used for all testing. The research implementation code can be found on Github: https://github.com/Phindulo60/Research. This section displays the experimental results for the suggested models. The performance of the prediction models, which were produced by the Logistic Regression, Random Forest, and Decision Trees algorithms, was assessed using the metrics of accuracy, precision, recall, and F1-score. To attain higher precision and stability, numerous procedures were done in the suggested work, as demonstrated above. The suggested method employs various ensemble and classification machine-learning techniques, which are implemented in Python.

A. *Initial Risk Factors Relationship*

In the proposed work we intend to know how the independent variables and the target variable are related, this exploration broadens our understanding of the given information/data in order to solve the issue at hand optimally and best. A correlation test matrix and a correlation test bar graph, shown in the figures below (Figs. 5. and 6.), respectively, x and y are two distinct variables, and the correlation coefficients (r) between them serve as indicators of the strength of their linear relationship.

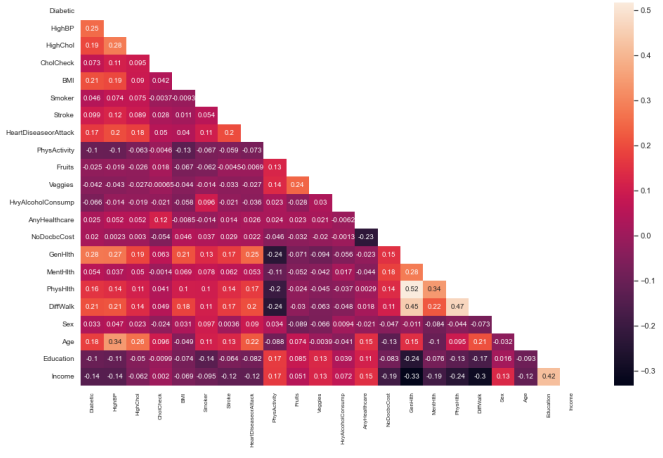

Fig. 5. Correlation test matrix

A positive relationship is shown by a linear correlation coefficient greater than zero. Indicating a negative association is a value less than zero. The two variables x and y do not have any relationship when their values are 0, to sum up. For further simplicity in understanding the variable's relationship and also to have a good correlation analysis, we implement a correlation bar graph in the below figure. Positive associations

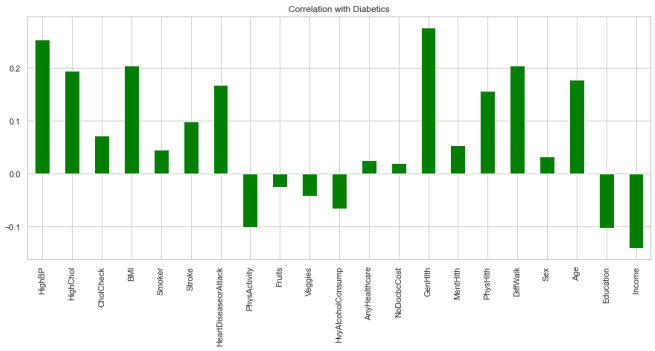

Fig. 6. Correlation Test Bar Graph

are shown by values greater than zero, whereas negative associations are indicated by values lower than zero. There can be no comparison if there is zero connection between the two variables. The idea of a negative correlation also referred to as an inverse correlation, is essential to the construction of diversified portfolios that may be better able to withstand portfolio volatility. The above correlation testing bar graph clearly shows that 15 independent variables are positively correlated with the target variable, while the remaining independent variables are negatively correlated, as indicated by the linear correlation coefficient. Our aim in this research work is to optimally select the best features ideally from the features which already have a positive relationship with the target variable. This will thus assist us in addressing the issue of having a high number of diabetic patients by giving an early diagnosis of diabetes.

B. *Evaluation Metrics*

The prediction accuracy, precision, and recall metrics were used to assess each model's performance. These measurements are based on a two-by-two matrix called the confusion matrix, which contrasts the anticipated class values of the model with the actual class values. The evaluation criteria that are used to base forecasts are as follows:

- **Sensitivity:** To correctly identify the condition, and in our case, it is used to count the number of people who tested positive or the number of people who have received a diabetes diagnosis.
- **Specificity:** Those who either tested negative for diabetes or did not have it were regarded as healthy.
- **F-Measure:** The Harmonic Mean between recall and precision is known as the Dice Similarity Coefficient, also known as the F-measure or F1 Score. The F1 Score is available between [0, 1]. It demonstrates how precise and dependable your classifier is.

- **10-fold Cross validation:** Cross-validation is a method for evaluating prediction models that divides the original sample into a training set and a test set.
- **Precision and Recall:** While the recall is the percentage of relevant instances that were retrieved, precision is the proportion of relevant examples among the recovered instances. Therefore, relevance serves as the foundation for precision and memory.
- **Accuracy:** How well our system identified patients with diabetes as having diabetes and those without diabetes as having non-diabetes.
- **True positives:** The first quadrant displays the percentage of diabetes patients who received a timely diagnosis.
- **False positive:** Patients without diabetes who were mistakenly identified as having the condition.
- **True negative:** People who are not diabetic appropriately classified themselves as such.
- **False negative:** People with diabetes were misclassified as non-diabetics.

C. *Feature Selection Results*

Depending on the criteria being assessed, feature selection selects a subset of pertinent traits from the input dataset. n subsets are generated from a set of features. The attributes are presented in ascending significance. A feature vector and the feature vectors around it could be redundant. Symmetric uncertainty is used to reduce overlap between two feature vectors. We can delete one of the duplicate features if there are two in the dataset because they almost always yield the same result. The patient's records contain a variety of characteristics that can be utilized to diagnose the patient's medical condition. Good qualities that are pertinent to the classification goal are chosen, but there shouldn't be any duplication. It is possible to choose the correlation between two attributes using the traditional linear correlation method or another way based on information theory.

The suggested bi-directional feature selection approach is made up of two phases joined in the last stage. The sequential forward feature selection filter approach was employed in the initial phase to rank features relative to the target class. By deleting irrelevant data, the goal of the first phase is to choose the subset of features that are most pertinent to the task at hand while optimizing computation efficiency and minimizing generalization error (that acts as noise). The most relevant features are selected in specific numbers to check for accuracy and performance, the details are as follows:

- **Top 3 features were selected in the first Sequential Forward Feature Selection iteration:** The selection score is 72.56%, in 39.00 seconds. The results of the Top 3 selected features are: ['HighBP', 'CholCheck', 'GenHlth']

- **Top 5 features were selected in the second Sequential Forward Feature Selection iteration:** The selection score is 72.48%, in 82.49 seconds. The results of the Top 5 selected features are: ['HighBP', 'CholCheck', 'Stroke', 'Veggies', 'GenHlth']

- **Top 7 features were selected in the third Sequential Forward Feature Selection iteration:** The selection score is 72.43%, in 135.96 seconds. The results of the Top 7 selected features are: ['HighBP', 'CholCheck', 'Stroke', 'HeartDiseaseorAttack', 'Fruits', 'Veggies', 'GenHlth']

- **Top 10 features were selected in the fourth Sequential Forward Feature Selection iteration:** The selection score is 72.48%, in 236.28 seconds. The results of the Top 10 selected features are: ['HighBP', 'CholCheck', 'Smoker', 'Stroke', 'HeartDiseaseorAttack', 'Fruits', 'Veggies', 'AnyHealthcare', 'NoDocbcCost', 'GenHlth']

Additionally, we can apply the hit-and-trial method for various values of k features to determine the ideal number of significant features, and then we can decide by plotting it against the performance of the model. A visual representation of the Sequential Forward Feature Selection process for choosing the precise significant number of features for identifying diabetes may be found in Figure. 7.

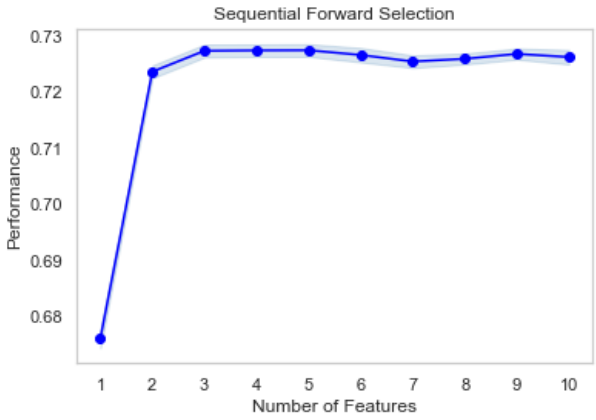

Fig. 7. Sequential Forward Feature Selection Performance

The second phase of this approach is to experiment with a sequential backward elimination selection approach. Based on the linear regression model, sequential backward elimination selection is a feature selection technique used to exclude unimportant features. The correct features for the model were anticipated using this approach. This approach has some benefits, such as lengthening the training period, reducing complexity, and enhancing accuracy and performance. The same "SequentialFeatureSelector()" function can be used to perform backward elimination by disabling the forward argument. The most relevant features are selected in specific numbers to check for accuracy and performance, the details are as follows:

- **Top 3 features were selected in the first sequential backward elimination selection iteration:** The selection

duration is 123.31 seconds. The results of the Top 3 selected features are: ['HighBP', 'HighChol', 'CholCheck']

- **Top 5 features were selected in the second sequential backward elimination selection iteration:** The selection duration is 119.29 seconds. The results of the Top 5 selected features are: ['HighBP', 'HighChol', 'CholCheck', 'BMI', 'Smoker']

- **Top 7 features were selected in the first sequential backward elimination selection iteration:** The selection duration is 116.52 seconds. The results of the Top 7 selected features are: ['HighBP', 'HighChol', 'CholCheck', 'BMI', 'Smoker', 'Stroke', 'HeartDiseaseorAttack']

- **Top 10 features were selected in the first sequential backward elimination selection iteration:** The selection duration is 107.43 seconds. The results of the Top 10 selected features are: ['HighBP', 'HighChol', 'CholCheck', 'BMI', 'Smoker', 'Stroke', 'HeartDiseaseorAttack', 'PhysActivity', 'Fruits', 'Veggies']

Sequential Forward and Backward Elimination Feature Selection is what we refer to as the bi-directional feature selection strategy that was proposed in the final phase of our exhaustive search approach. In this combination of a strategy, also known as the floating search, the Forward Selection and Backward Elimination procedures are first used to the original feature set in order to construct at least two intermediate feature subsets. In the second stage, two subsets are combined into a single pool, and the same number of features are chosen from each subset in accordance with the majority of votes. If two features receive an equal number of votes, the feature with the greater pattern categorization performance rate when the other characteristics are integrated will prevail. The final feature set will be chosen in the third step using a laborious process based on the features that have received the highest votes. The proposed approach, the combinatorial approach will deal with a significantly smaller set than the original full set in accordance with specific top-selected feature sets. The Below figures are the detailed results of the combinatorial feature selection approach which saw the selection top 10 relevant features with the highest prediction score of 85.11% but with the highest training duration.

- **Top 3 features were selected in the first bi-directional feature selection approach iteration:** The selection score is 84.73%, in 1.23 seconds. The results of the Top 3 selected features are: ['HighBP', 'HighChol', 'GenHlth']

- **Top 5 features were selected in the second bi-directional feature selection approach iteration:** The selection score is 85.05%, in 10.32 seconds. The results of the Top 5 selected features are: ['HighBP', 'HighChol', 'CholCheck', 'BMI', 'GenHlth']

- **Top 7 features were selected in the third bi-directional feature selection approach iteration:** The selection score is 85.08%, in 17.98 seconds. The results of the Top 7 selected features are: ['HighBP', 'HighChol', 'CholCheck', 'BMI', 'Stroke', 'HeartDiseaseorAttack', 'GenHlth']

- **Top 10 features were selected in the fourth bi-directional feature selection approach iteration:** The selection score is 85.10%, in 60.55 seconds. The results of the Top 10 selected features are: ['HighBP', 'HighChol', 'CholCheck', 'BMI', 'Smoker', 'Stroke', 'HeartDiseaseorAttack', 'PhysActivity', 'Fruits', 'NoDocbcCost', 'GenHlth']

The 'HighBP' feature is ranked the highest in all three feature sets and the 'CholCheck' follows, from the third feature in our feature sets there are different features which range from: 'Stroke', 'Fruits', 'Smoker','HeartDiseaseorAttack', 'PhysActivity', 'NoDocbcCost' to 'GenHlth'. This indicates that these characteristics are the most effective at predicting diabetes disease, at least in a statistical sense. Because of this knowledge, clinicians can diagnose patients by beginning with the most important characteristics and working their way down to the least important ones.

A very important and crucial part for healthcare practitioners is a medical diagnosis. The classification of diabetics in particular is extremely complicated. Early diabetes detection is crucial for managing the disease. A patient must undergo a number of tests, and after that, it is quite problematic for specialists to keep track of several elements during the diagnosis process, which can result in false results and make detection very difficult. This is the very same reason why we employed the use of another proposed approach of selecting the best features for predictive purposes, which is the Overall Regularized Trees with Overall Feature Importance. In the below figure(Fig. 8.) all features were ranked in accordance to their importance in predicting the target column('Diabetic'). There were also 4 different sets derived from this approach namely: **Top 3, Top 5, Top 7, and Top 10 feature sets**. These sets were utilized to predict diabetes and furthermore measure the performance of machine learning algorithms on them.

D. Classification Methods Results

The pooled diabetes dataset was subjected to feature selection and extraction, as shown in the below table(Fig. 9.), and the outcomes varied significantly depending on the classification technique that was used to build the model. Even a few of the models created using the original dataset outperformed several models created using chosen feature selections. The model with the best accuracy and precision that we developed was a decision tree classifier, with an accuracy of 85.01% on the Top 5 feature set and it further produced accuracies of 84.88% and 84.26% respectively on the Top 7

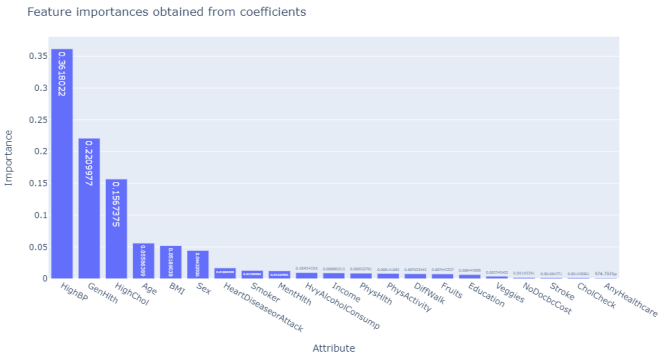

Fig. 8. Overall Regularized Trees with Overall Feature Importance

Combinatorial approach produced data set with unselected data set

Machine Learning Algorithms	Top 3 features	Top 5 features	Top 7 features	Top 10 features	Unselected set of data
Logistic Regression	0.6814	0.7092	0.7165	0.7158	0.7408
Optimized Random Forest	0.8464	0.8501	0.8483	0.8421	0.9077
Decision Tree Classifier	0.8464	0.8501	0.8488	0.8426	0.8606
Gaussian Naive Bayes classifier	0.8365	0.7801	0.7908	0.7179	0.7179

Fig. 9. Accuracy Metric measure on 4 Machine Learning Algorithms on Combinatorial approach produced dataset(Also on the unselected dataset)

and Top 10 feature sets. All feature sets were produced by the combinatorial feature selection approach.

Regularized Trees produced data set with unselected data set

Machine Learning Algorithms	Top 3 features	Top 5 features	Top 7 features	Top 10 features	Unselected set of data
Logistic Regression	0.6814	0.7127	0.7145	0.7162	0.7408
Optimized Random Forest	0.8464	0.8463	0.8400	0.8318	0.9077
Decision Tree Classifier	0.8464	0.8474	0.8405	0.8192	0.8606
Gaussian Naive Bayes classifier	0.8365	0.8289	0.8109	0.7967	0.7179

Fig. 10. Accuracy Metric measure on 4 Machine Learning Algorithms on regularized trees approach produced dataset(Also on the unselected dataset)

In addition, our suggested regularized decision tree feature selection method's implementation was made possible by changing the "max depth," "min sample split," "min sample leaf," and "max feature size," which specify the minimum number of samples a leaf must have in order to prevent further node splitting when the node's number of observations falls below a predetermined value and to prevent over-fitting when a leaf node has too few observations. According to the above table (Fig. 10), the decision tree had the most accuracy, at 84.05%, on the Top 7 feature set, while the optimized random forest had the highest accuracy on the Top 5 and Top 10 feature sets, at 84.63% and 83.18%, respectively, furthermore these have been validated. Due to the small differences across models across different machine learning methodologies, clinicians may choose to select a model that is straightforward to deploy without experiencing a major loss in accuracy. Although the Optimized Random Forest model, which has the highest accuracy in most selected datasets, is an effective model, it is crucial that a few patients who are actually susceptible to diabetes disease are misclassified in a medical context like this. Therefore, the model with the highest sensitivity, which is once again the Optimized Random Forest model, will most like be selected to be used in datasets produced by the combinatorial feature selection approach.

V. DISCUSSION

A machine learning approach was suggested in this work to forecast the occurrence of diabetes through the use of determining the crucial risk factors which have a high influence in predicting diabetes. This research addressed the potential transition between the two classifications of pre-diabetes and diabetes, whereas earlier efforts in [5] and [7] devised a strategy for anticipating the incidence of diabetes. Pre-diabetes prediction has received less attention because most research has concentrated on the prediction of undiagnosed diabetes. The prediction models in this work were developed utilizing a huge dataset and ensemble machine-learning approaches, in contrast to the studies described above. Additionally, by altering the amount of years utilized to train the models, the effect of the accumulated medical data on prediction accuracy was demonstrated. Predictors were developed using a data-driven feature selection methodology in order to successfully identify the different classes in the dataset. The resultant 10 features were 'HighBP', 'GenHlth', 'HighChol', 'BMI', 'NoDocbcCost', 'Fruits', 'PhysActivity' and 'Stroke'. 'HighBP' and 'HighChol' were the most important predictors selected by the combinatorial feature selection approach. When compared to the five most common predictors of diabetes ('HighBP', 'GenHlth', 'BMI', 'age', and 'sex'), the suggested models incorporating the selected variables showed enhanced prediction ability. The greatest accuracy of the machine learning method was 90.77% for the unselected feature set and 85.01% for the Top 5 chosen feature sets when more than 200,000 entries of data were used for training. We can draw the conclusion that using a lot of cleaned data helped to increase prediction accuracy.

The prediction models were created using Logistic Regression, Optimized Random Forest, Decision Tree, and Gaussian Naive Bayes classifier. The created prediction models outperformed the existing models, the standard statistical analysis technique, according to experimental data. On the test data, the performance disparity amongst the methods was minimal. Class overlap in the feature space can explain this.

The Optimized Random Forest findings revealed a sizable performance disparity between the prediction models. In general, the study's findings showed that the developed prediction models outperformed the current clinical screening approach. The study's conclusions and the produced prediction models

are put to use to the advantage of patients and practitioners. The models can be a useful tool for practitioners to use while making clinical decisions and providing patient counseling. Additionally, early diagnosis of the condition permits persons with diabetes and those at risk of developing it to take precautions that may slow the disease's course and its potentially fatal complications.

Our research suggests that two additional studies need to be conducted. The first step would be to combine several datasets in order to lessen the difficulty of diagnosing diabetes, which is caused by the overlap between normal and diabetes classes. The second step would be to improve web and mobile application usability and increase accessibility to the prediction models.

VI. CONCLUSIONS

The goal of the project was to develop and use a practical machine-learning model for early diabetes prediction. And after looking at a variety of machine learning models, we came to the conclusion that the Optimized Random Forest model had the highest testing accuracy, scoring 91.0% on the unselected feature set and 85.01% on the Top 5 selected feature set with the help of the proper hyperparameter adjustment. The characteristics we chose, the scaling mechanism we employed, the hyperparameters, and the class imbalance we removed all had an effect on our accuracy. The aforementioned factors could be changed to produce a different result.

Our findings demonstrate that by applying feature selection and extraction to the diabetic disease dataset, a more accurate model for diabetes prediction may be produced. To create the best model, a variety of feature selection tactics and machine learning algorithm combinations must be evaluated. The improvements over using the original dataset are significantly influenced by the machine learning technique that was utilized. The ranking of the features by our feature selection procedures shows that "HighBP" is consistently the most critical feature for predicting heart disease, followed by "GenHlth," "HighChol," and "BMI" values; however, these characteristics are scored differently by our feature selection strategies. Clinicians and other healthcare professionals can use the models developed in this work to identify diabetes in new patients if patient data for the pertinent criteria is available. Knowing which specific parameters were employed during pre-processing is also beneficial because it shows which are more statistically significant for forecasting diabetic disease.

Conflict of Interest: Regarding the research, writing, and/or publication of this work, the author declared that there were no potential conflicts of interest.

Funding: The study received no outside funding.

Institutional Review Board Statement: The study was exempt from ethical review and approval because it made use of pre-existing data.

ACKNOWLEDGMENT

Thank you to my supervisor Dr Pravesh Ranchod for providing helpful, insightful and constructive feedback

REFERENCES

[1] Agnes Erzse, Nicholas Stacey, Lumbwe Chola, Aviva Tugendhaft, Melvyn Freeman, and Karen Hofman. The direct medical cost of type 2 diabetes mellitus in south Africa: a cost of illness study. Global health action, 12(1):1636611, 2019.

[2] A. Kumar Dwivedi, "Analysis of computational intelligence techniques for diabetes mellitus prediction," Neural Comput. Appl., vol. 13, no. 3, pp. 1–9,2017.

[3] Pathak, A. K. and Arul Valan, J. (2020). A predictive model for heart disease diagnosis using fuzzy logic and decision tree. In Smart computing paradigms: new progress and challenges, pages 131–140. Springer

[4] Mitushi Soni, Dr. Sunita Varma, 2020, Diabetes Prediction using MachineLearning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) Volume 09, Issue 09 (September 2020)

[5] A. Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe, "A Decision support system for Diabetes Prediction Using Machine Learning and Deep learning techniques," 2019 1st International Informatics and Software EngineeringConference (UBMYK), 2019,pp. 1-4, doi:10.1109/UBMYK48245.2019.8965556.

[6] Sun, Y., Babbs, C., and Delp, E. (2005). A comparison of feature selection methods for the detection of breast cancers in mammograms: Adaptive sequential floating search vs. genetic algorithm. In 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, pages 6532–6535

[7] Rahman, M.M.; Davis, D.N. Addressing the class imbalance problem in medical datasets. Int. J. Mach. Learn. Comput. 2013, 3, 224–228.

[8] Bradshaw, D., Norman, R., Pieterse, D., and Levitt, N. S. (2007). Estimating the burden of disease attributable to diabetes south africa in 2000. South African Medical Journal, 97(8):700–706

[9] Gotfredsen, C., Buschard, K., and Frandsen, E. (1985). Reduction of diabetes incidence of bb wistar rats by early prophylactic insulin treatment of diabetes-prone animals. Diabetologia, 28(12):933–935

[10] Despres, J. (2006). Intra-abdominal obesity: an untreated risk factor for type 2 diabetes and cardiovascular disease. Journal of endocrinological investigation, 29(3):77

[11] Norberg, M.; Eriksson, J.W.; Lindahl, B.; Andersson, C.; Rolandsson, O.; Stenlund, H.; Weinehall, L. A combination of HbA1c, fasting glucose and BMI is effective in screening for individuals at risk of future type 2 diabetes: OGTT is not needed. J. Intern. Med. 2006, 260, 263–271.

[12] Ramachandran, Ambady, and Chamukuttan Snehalatha. "Current scenario of diabetes in India." Journal of diabetes 1.1 (2009): 18-28.

[13] Grundlingh, Nina, et al. "Assessment of prevalence and risk factors of diabetes and pre-diabetes in South Africa." Journal of Health, Population and Nutrition 41.1 (2022): 1-12.

[14] Deshpande, Anjali D., Marcie Harris-Hayes, and Mario Schootman. "Epidemiology of diabetes and diabetes-related complications." Physical therapy 88.11 (2008): 1254-1264.

[15] Saeedi, Pouya, et al. "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas." Diabetes research and clinical practice 157 (2019): 107843.

[16] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." Ijcai. Vol. 14. No. 2. 1995.

[17] Kang, Myeongsu, and Jing Tian. "Machine Learning: Data Preprocessing." Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things (2018): 111-130.

[18] Khalid, Samina, Tehmina Khalil, and Shamila Nasreen. "A survey of feature selection and feature extraction techniques in machine learning." 2014 science and information conference. IEEE, 2014.

[19] Panda, Debjani, et al. "Predictive systems: Role of feature selection in prediction of heart disease." Journal of Physics: Conference Series. Vol. 1372. No. 1. IOP Publishing, 2019.