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*—With the unnoticeable effect that diabetes sickness has on South Africa's financial system, there is a huge need to utilize the advancement of technology to assist greatly in the efforts of trying to minimize the costs that diabetes has in South Africa's public health sector. As already known, There is no cure for diabetes, but there are several strategies that can be implemented to mitigate and reduce the harm that diabetes causes. With the cure unknown the most common way to mitigate the harm of diabetes is to focus on its prognosis. Advanced machine learning techniques with better feature selection approaches/implementations become an integral part of risk prediction that will be used to mitigate the risk of undiagnosed diabetes through various selected factors. After performing thorough data pre-processing and feature selection with a combinatorial approach and a Decision tree Importance approach, we used numerous Machine Learning models like Logistic Regression, Optimized Random Forest, and Decision tree classifier to predict Diabetes on train-test splits of 75-25. 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*—Machine learning techniques, risk prediction, Diabetes, Data Pre-processing, 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 [16]. Like other nations, an estimated 90-95% of South Africans who are affected by diabetes, are commonly affected by Type 2 Diabetes [1]. Diagnosed type 2 diabetics cost the public sector between R21.8 billion and R22.7 billion in 2018, including both diagnosed and undiagnosed patients [1]. In 2030, it is anticipated that nearly all instances of type 2 diabetes will cost the public sector R35.1 billion [1]. Approximately 51% of these projected 2030 costs are linked to type 2 diabetes treatment, whereas 49% are related to complications [1]. South Africa's public health system is heavily burdened by Type 2 Diabetes. In 2018, treating all common cases cost almost 12% of the overall national medical budget [1]. If the current care system is maintained and case detection is improved, the direct cost will see a steady decrease as the prevalence rises [1].

Identify applicable funding agency here. If none, delete this.

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*

Diabetes is one of the diseases connected to insulin hormones. When the human body responds to it, it can have a very negative impact. 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 a 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]. In many countries, diabetes is the main contributor to cardiovascular disease, blindness, renal failure, and lower-extremity amputation [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

The studies based on various predictive machine learning algorithms that are implemented specifically for the diagnosis of chronic illnesses are presented in this portion of the report. One of the primary guidelines that this project will adhere to is the prediction models constructed in the studies presented on this portion of the report.

For the purpose of predicting diabetes, Kumar Dwivedi in [2] assessed a variety of machine learning systems. The implementation of algorithms like classification trees, support vector machines, Artificial Neural Networks, logistic regression, and k-nearest neighbors was utilized. Specificity, Accuracy, Recall, Precision, FPR, Negative Prediction Value, Rate of Misclassification, F1 Score, and ROC Curve are some of the measures used to evaluate this system's performance. Comparatively, the best accuracy was achieved by logistic regression (78%), while the rate of mis-classification was 0.22. The negative predictive value's greater precision came out to be as 73% and 82% using Logistic Regression and Naive Bayes respectively. And 10-fold cross-validation is used to separate the data.

A fuzzy rule-based approach and a decision tree were developed by Pathak, Asim Kumar, and Arul Valan in [3] as a way to create a predictive model for heart disease diagnosis. This method uses a number of parameters, including BMI, minimum blood pressure, plasma glucose levels, and serum insulin levels. Specific insulin dosage is suggested based on a few factors; the probability of diagnosis presents the findings using five fuzzier figures and accurately predicts the likelihood of avoiding hypoglycemia (low blood sugar). Three fuzzier integers are used to represent the output results. In terms of area under the curve, specificity, accuracy, and sensitivity in predicting type 2 diabetes are utilized, when the fuzzy rule-based technique is compared to the predictive performance of the test data set as suggested in one of the many tested machine learning techniques. The predictive model created using the fuzzy rule-based approach with a decision tree offers insight into how a predictive model could be built.

Decision Trees, k-nearest neighbors, Gradient Boosting (GB), Logistic Regression, Random Forest, and support-vector machines are among the algorithms utilized in the study [4] named Diabetes Prediction using MLT. According to the results, random forest achieved the greatest classification accuracy, or 77%, in comparison. Support vector machine, random forest, and fully convolutional neural networks (CNN) are the approaches utilized for deep learning, according to a methodology proposed by A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe in [5]. Deep Learning (76.81%), support vector machine (65.38%), and random forest (83.67%) had the highest accuracy. The findings showed that Random Forest outperformed support vector machine and Deep Learning techniques for predicting diabetes.

An analysis of mammography feature selection methods for unified breast cancer detection is presented in [6]. Curvilinear features, textural features, Gabor features, and multi-resolution features were extracted from a 512x512 pixel patch containing either healthy tissue or breast cancer. An adaptive floating search and a GA were used to choose the features, and linear discriminant analysis was used to classify the cancerous and healthy regions.Az=0.90 is the overall ROC performance for SGA and LDA, Az=0.93 is the overall ROC performance for CHC and LDA, and Az=0.96 is the overall ROC performance for ASFFS and LDA. For each feature selection approach, fewer than 25% of the 86 features were picked, and at least one feature from each kind was chosen. Four different sorts of traits complement one another even if some of them may be linked. Performance is evaluated using the ROC curve's area under the curve (Az). In this work, feature selection and classification were accomplished using Linear Discriminant Analysis and Wrapper Based Selection, Adaptive Sequential Forward Floating Search Feature Selection, and GA for Feature Selection.‘

[7] presents a model for anticipating type 2 diabetes in nondiabetic patients with cardiovascular disease. In order to anticipate the likelihood that the disease will manifest throughout the follow-up period, the study reported a T2D prediction model. Korea University Guro Hospital provided the study's electronic health records (EHRs) (KUGH). 28 features in total were included, and 8454 subjects were followed up on for five years. The logistic regression (LR) model's AUC value was allegedly reached by the authors at 78.0%. The dataset used in this study only contained those who had cardiovascular risks.

C. Predictive Metrics For Predictive Modelling

A coincidence matrix is the main instrument used to evaluate performance in classification tasks. The formulae for the metrics that may be derived from the most popular coincidence matrix are also provided below (Fig. 1). From upper-left to lower-right, the numbers outside of this diagonal indicate the incorrect decisions that were made, whereas the numbers that run along it show the appropriate judgments. By dividing the total number of positives by the true positive count, one

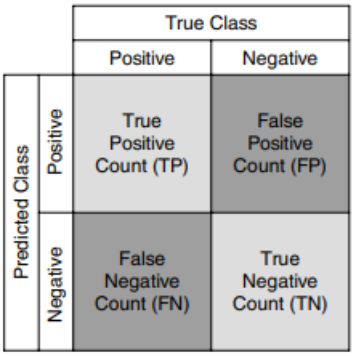

Fig. 1. A simple coincidence matrix

may get the true positive rate (also known as the hit rate or recall) of a classifier. By dividing the incorrectly categorized negatives (the false negative count) by the total negatives, one may determine the classifier's false positive rate, also known as the false alert rate. By dividing the total number of correctly identified positives and negatives by the total number of samples, a classifier's overall accuracy is calculated. A number of aggregated performance measures are created using recall (also known as sensitivity), specificity, and F-measure. When comparing the anticipated accuracy of two or more approaches, the bias caused by the random sampling of the training and holdout data samples may be reduced by using a technique known as k-fold cross-validation [20]. For rotation estimation, the entire data set is randomly split into k almost equal-sized, mutually exclusive portions. This procedure is known as k-fold cross-validation [20]. The categorization model is trained and tested K times. Every time, it is trained on all but one fold and tested on the others. A model's overall accuracy can be determined by simply summing the k different accuracy indices.

III. PROPOSED METHODOLOGY

Since a decade ago, the number of people with diabetes has dramatically increased. The key factor contributing to the rise of diabetes is current human behavior. 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 used in this research project, which was obtained from the Kaggle repository. Over 400,000 Americans participate in the survey each year, providing information on risky behaviors, chronic health issues, and the usage of preventative treatments. 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. 2. Detailed description of the features

In the above further description, "1" is to indicate the characteristic's presence, and "0" is to indicate the characteristic's absence. The above attributes are distributed by age among all surveyed patients as shown in Fig. 3.

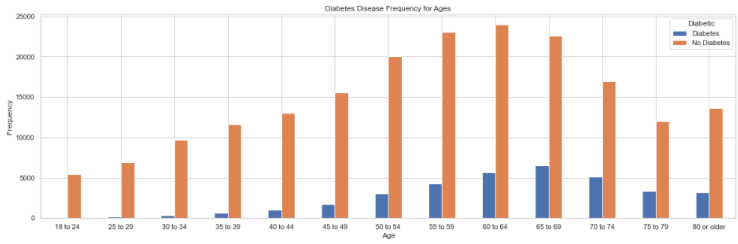

Fig. 3. Variation of Age for each target class

B. *Pre-processing of Data*

One of the important processes in machine learning is data pre-processing [21]. It enhances the performance of machine learning models and data quality [21]. The process entails preparing the raw data for training and evaluating prediction models by cleaning and converting them. Preparation, cleaning, feature selection, handling of missing values, and modification of data are all included in data pre-processing. Data preparation should produce a final dataset that is accurate and helpful for further data mining methods. 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. This assumption is broken in many real-world situations, though. The machine learning classifier has a tendency to be more

biased towards the majority class when working with datasets that have a class imbalance, which results in the incorrect categorization of 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]. Majority under-sampling and synthetic minority over-sampling (SMOTE) techniques were used to address the issue.

D. Feature Selection

The process of choosing a subset of the dataset's most pertinent characteristics to characterize the target variable is known as feature selection [22]. 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 approaches exclude features based on predetermined standards. Wrapper-based approaches evaluate and rank features using a modeling algorithm that is treated as a "black box." The embedded methods incorporate feature selection techniques including random forest (RF) feature selection and least absolute shrinkage and selection operator (LASSO) [23]. Exhaustive search, Pearson correlation, chi-squared, recursive feature removal, Lasso, and tree-based feature selection approaches are only a few examples of the several types of feature selection methods [24].

This section outlines a data-driven methodology for choosing features to forecast the occurrence of diabetes using statistical and machine-learning techniques. Both numerical variables from the diagnostic results and categorical entities from the questionnaire responses were included in the dataset created using the aforementioned approaches. Finding a collection of ideal traits that could effectively identify the two groups is the goal of the feature selection process.

1) ***Feature Importance - Logistic Regression****:* By counting the number of times each variable appears in a significant subset using logistic regression, the average impurity decrease is obtained from all subsets, and the most frequently appearing variable is labeled as the "most important." By standardizing the variables, repeating the regression, and ranking the coefficients from highest to lowest, one may determine the feature relevance from logistic regression. The winner's circle is reserved for those with the highest coefficients. Finally, we can assert that a feature is more significant the greater the score coefficient. In Fig. 4., the scores of various aspects are displayed.

2) ***Feature Importance - Optimized Random Forest****:* The average impurity decrease derived from all of the decision trees in the random forest is used to quantify the feature relevance. The random forest feature importance is estimated as a reduction in the node's impurity, which is weighted by

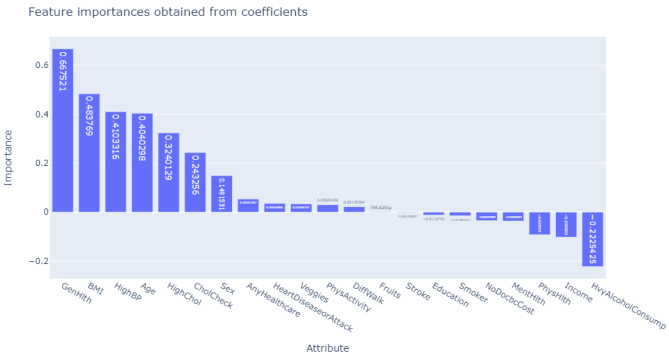

Fig. 4. Feature Importance on unselected dataset Obtained from coefficients

the chance of accessing that node. The count of samples that arrive at that specific node divided by the total number of samples yields the probability of it. Finally, we may state that the trait is more significant the higher the score. In Fig. 5., the scores of various aspects are displayed.

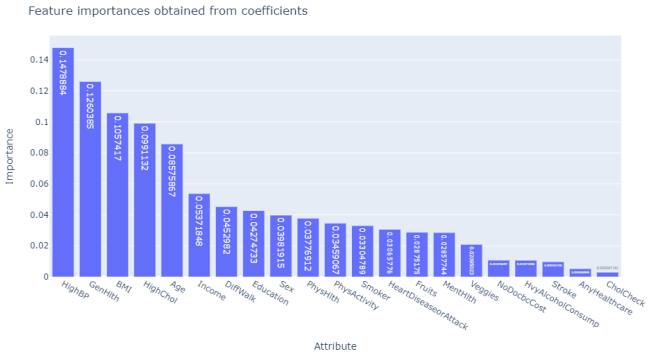

Fig. 5. Feature Importance Obtained from coefficients

The node impurity, which was weighted by the chance of reaching the node, was used to calculate the feature importance. The proportion of samples that reach the node to all samples was used to define the node probability [12]. The normalized value of the feature importance is shown on the x-axis in Figure 3. The feature is more significant the higher the value. The proposed data-driven feature selection strategy, which is consistent with multiple research [13] [14], identified the most significant and pertinent features to indicate the presence of diabetes.

3) ***Feature Importance - Decision Trees Classifier****:* A splitting rule connects each of the nodes that make up a decision tree. A feature and the value it should be split on are involved in the splitting rule. When a dataset observation is split, it indicates that it moves to the left of the node if the splitting rule is met. The observation moves to the right if the rule is not met. The likelihood that an observation will fall into a particular node is one of the metrics used to determine the importance of the features as displayed in Fig. 6. Each node in the decision tree has a probability that is computed

by simply dividing the number of samples in that node by the total number of observations in the dataset.

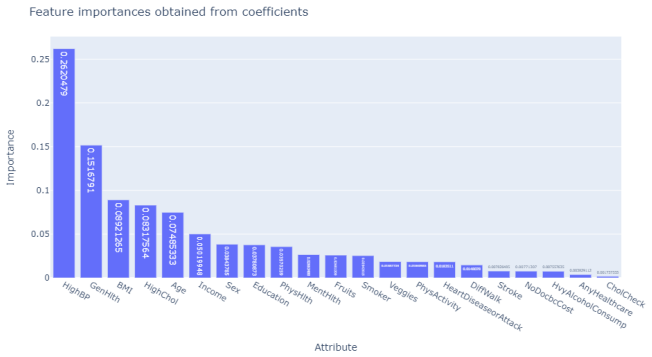

Fig. 6. Feature Importance Obtained from coefficients

E. *Feature Selection Techniques*

*1) **Sequential Forward Feature Selection**:* The family of greedy search algorithms, which includes sequential feature selection methods, is used to shrink an initial d-dimensional feature space into a k-dimensional feature subspace where k d. The goal is to choose the most pertinent subset of features for the task, which maximizes computing effectiveness and reduces generalization error by eliminating irrelevant information (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. The practical implementation is shown in Fig. 7.

- First and foremost, the best feature out of all the features is chosen (i.e., using some criterion function).
- The best pair of features is then chosen by combining this greatest feature with one of the remaining features.
- The best feature triplet is then constructed by combining the two best features with one of the remaining features.
- This process is repeated until K features, the predetermined number of features, are chosen.

```
from mlxtend.feature_selection import SequentialFeatureSelector as sfs
from sklearn.linear_model import LinearRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC

lr = LogisticRegression()

no_features = [3,5,7,10]
for i in range(len(no_features)):
    fs1 = sfs(lr, k_features=no_features[i], forward=True, verbose=1, scoring='accuracy', cv=5)
    startf = timer()
    fs1 = fs1.fit(X_train, y_train)
    endf = timer()
    t_timef = endf-startf
    trained_f = list(fs1.k_feature_names_)
    score = fs1.k_score_
    k_features = fs1.k_feature_idx_
    data.columns[1:][list(k_features)]
    X_train_sele = fs1.transform(X_train)
    X_test_sele = fs1.transform(X_)
    print(f'Top {no_features[i]} Features in Forward Selection')
    print(f"Training duration: {t_timef}")
    print("Selected features:", trained_f)
    print(f'prediction score for these {no_features[i]} features:',score)
    print("-----------------------------------------------------------------------------------------------")
```

Fig. 7. Sequential Forward Feature Selection Implementation

*2) **Sequential Backward Feature Selection**:* The sequential backward selection approach seeks to increase computational efficiency and decrease generalization error by decreasing the dimensionality of the initial feature subspace from N to K features with a minimum drop in model performance. To get to the list of K features, the main concept is to successfully delete features from the given features list, which consists of N features. The component that results in the least performance loss is removed at each level. 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. The practical implementation is shown in Fig. 8.

- Here, we have the sequential backward selection method in action. The original feature space must be reduced to a subset of features, which the class accepts as an instance of an estimator in the constructor. After feature scaling is completed to identify the subset of features, the training and test data sets can be passed.
- There is a fit and transform method offered by the class. The subset of feature indices and accuracy scores are determined using the fit approach.

```
lr1 = LinearRegression()
bfs2 = sfs(lr1, k_features=5, forward=False, verbose=1, scoring='accuracy', cv=5)

#startb = timer()
#bfs1 = bfs2.fit(independent, target)
#endb = timer()
no_features = [3,5,7,10]
for i in range(len(no_features)):
    bfs2 = sfs(lr1, k_features=no_features[i], forward=False, verbose=1, scoring='accuracy', cv=5)
    startb = timer()
    bfs2 = bfs2.fit(X_train, y_train)
    endb = timer()
    t_time = endb-startb
    trained_b = list(bfs2.k_feature_names_)
    score = bfs2.k_score_
    print(f'Top {no_features[i]} Features in Backward Elimination')
    print(f"Training duration: {t_time}")
    print("Selected features:", trained_b)
    print(f'prediction score for these {no_features[i]} features:',bfs2.k_score_)
    print("-----------------------------------------------------------------------------------------------")
```

Fig. 8. Sequential Backward Feature Selection Implementation

*3) **combinatorial feature selection approach(bi-directional elimination)**:* It is the combination of both Backward Elimination and Forward selection. The combinatorial feature selection approach is essentially a forward selection procedure but with the possibility of deleting a selected variable at each stage, as in the backward elimination, when there are correlations between variables. As visualized in Fig. 9., the Forward Selection and Backward Elimination methods are first used to the original feature set in this combination of a technique, which is also referred to as the floating search. The same number of features are selected from each subset in accordance with the majority of votes in the second step, which involves combining two subsets into a single pool. The feature with the higher pattern categorization performance rate when the other features are combined will win out if two features obtain an equal number of votes. In the third step, an exhaustive procedure will choose the final feature set based on the features that have garnered the most votes. By using voting, a good subset of all features was created, whose size is bigger than the desired subset size but significantly smaller than the original full set.

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

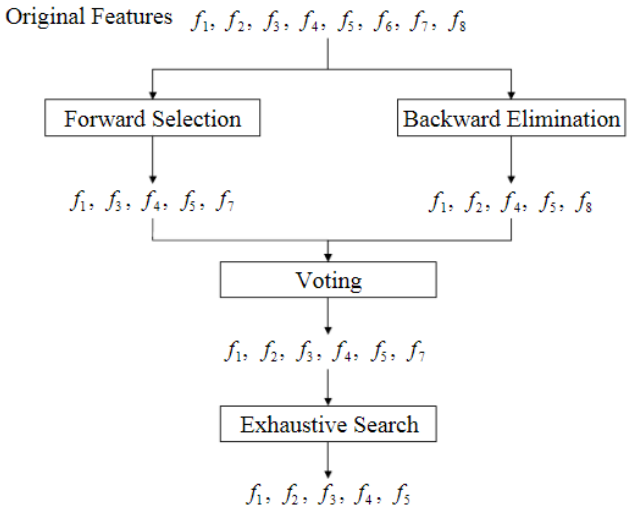

Fig. 9. Combined Feature Selection procedure

the regularized trees can choose high-quality feature subsets. Because category and numerical variables, missing values, disparities in scale across variables, interactions, nonlinearities, etc. may all be handled by tree models naturally. There are

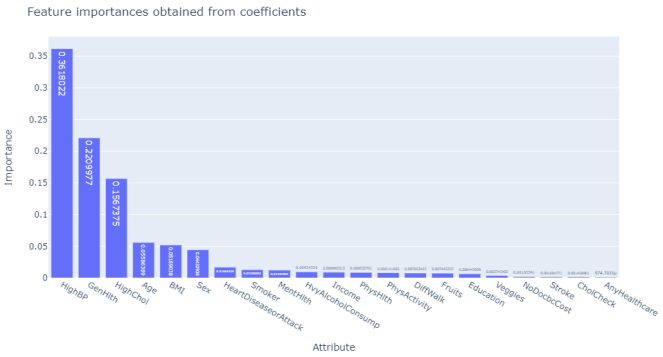

Fig. 10. Overall Regularized Trees with Overall Feature Importance

a few regularization parameters in a tree that we can use to control the size of the tree, like:

- max depth - the maximum length of the path from the root to the leaf
- min sample split - limit to stop the further splitting of nodes when the number of observations in the node is less than a given value.
- min sample leaf - minimum number of samples a leaf node must have. When a leaf node has too few observations further split will result in over-fitting.
- max feature size - maximum number of features evaluated before splitting.
- criterion - optional (default="gini") or Choose attribute selection measure: This parameter allows us to use the different-different attribute selection measure. Supported criteria are "gini" for the Gini index and "entropy" for the information gain.

In Fig. 10., we utilized the max depth = "15" and criterion =" entropy", to obtain that feature importance diagram.

F. Machine Learning Algorithms

*1) **Logistic Regression**:* Logistic Regression is a well-known supervised ML classification technique that is typically used to determine the likelihood of a binary response based on at least one prediction. They might be continuous or discrete. It is applied to the categorization of data. Additionally, Logistic Regression typically categorizes the data in binary form (0 and 1), which is utilized to determine if a patient has diabetes or not. This model's objective is to offer the greatest fit for assessing the correlation between predicted and desired values. The sigmoid function, which restricts the output to either 0 or 1, is used to anticipate the likelihood of the output.

- Split the data into a training set and testing set
- Fit a logistic regression model using sklearn
- Apply the model on the test data and make a prediction
- Evaluate the model accuracy using the confusion matrix
- Create the model and obtain the regression coefficients using stats model
- The most important step is to translate the regression coefficient into odds.

*2) **Random Forest**:* A well-known machine learning approach called Random Forest is applied to classification and regression problems. Bagging is an ensemble learning technique that combines a number of classifiers to enhance the performance of the model. It is used to reduce variance by producing extra data from training. The method acts during training and assigns the mode of classification of individual trees as output using a combination of multiple decision trees.

- Take a sample of random data from the training set first.
- Decision trees were then generated based on relationship with the chosen data.
- Give the decision tree your selection by number.
- Repeat step 1 and step 2 again.
- Locate each decision tree's predictions for the unknowable data, and then group them into the category with the most support.
- Lastly, determine the accuracy of the chosen category's results.

*3) **Decision Trees**:* A straightforward visual aid for classifying samples is a decision tree. This method of supervised machine learning constantly divides the data based on a given parameter. Both classification and regression issues can be resolved via decision tree analysis. A decision tree is incrementally built in conjunction with the decision tree algorithm, which divides a dataset into smaller subgroups. A decision tree is made up of leaf nodes, which are the terminal nodes that anticipate the conclusion, edges/branch, which correspond to the results of tests and connect to the next node or leaf, and nodes that test the value of various attributes.

- After loading the data, we understand the structure variables and determine the target feature variables (dependent independent variables respectively)

- Divide the data into training testing sets
- Performing The decision tree analysis using scikit learn
- But we should estimate how accurately the classifier predicts the outcome. The accuracy is computed by comparing actual test set values and predicted values.
- Now that we have created a decision tree, let's see what it looks like when we visualize it

*4) **Gaussian Naive Bayes classifier**:* This classifier is employed when the predictor values are continuous and are expected to follow a Gaussian distribution. Gaussian Naïve Bayes is the extension of naïve Bayes. While other functions are used to estimate data distribution, Gaussian or normal distribution is the simplest to implement as you will need to calculate the mean and standard deviation for the training data.

- loading some basic libraries that will be used to import and view the dataset.
- Importing Dataset
- Preprocessing
- Visualizing Dataset
- To maximize the model's efficiency, it's always a good idea to normalize the data to a common scale.
- uses the train test split module from the sklearn package to divide the dataset into training and testing sections.
- Import and instantiate the Gaussian Naive Bayes module from SKlearn GaussianNB.
- Use the accuracy score to reflect how successfully our Sklearn Gaussian Naive Bayes model predicted our target column

*5) **Implementing Selected Machine Learning Algorithms**:*

- Logistic Regression

```
param_grid_lr = {
    'max_iter': [20, 50, 100, 200, 500, 1000],
    'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'],
    'class_weight': ['balanced']
}
#GridSearchCV is a technique to search through the best parameter values from the given set of the grid of parameters
logModel_grid = GridSearchCV(estimator=LogisticRegression(random_state=42), param_grid=param_grid_lr,
                             verbose=1, cv=10, n_jobs=-1)

start = timer()
logModel_grid.fit(X_train_scaled, y_train)
end = timer()
y_predicted = logModel_grid.predict(x_)
importances1 = pd.DataFrame(data={
    'Attribute': X_train.columns,
    'Importance': logModel_grid.best_estimator_.coef_[0]
})

importances1 = importances1.sort_values(by='Importance', ascending=False)

fig = px.bar(importances1, x="Attribute", y="Importance", text_auto=True)
fig.update_layout(barmode='relative', title_text="Feature importances obtained from coefficients")
fig.show()
```

Fig. 11. Execution of Logistic Regression

- Optimized Random Forest

```
model1 = RandomForestClassifier(min_samples_leaf = 1, n_estimators = 100, max_features = 'sqrt')
start1 = timer()
model1.fit(X_train_scaled, y_train)
end1 = timer()
y_predicted1 = model1.predict(x_)
importances2 = pd.DataFrame(data={
    'Attribute': X_train.columns,
    'Importance': model1.feature_importances_
})

importances2 = importances2.sort_values(by='Importance', ascending=False)

fig = px.bar(importances2, x="Attribute", y="Importance", text_auto=True)
fig.update_layout(barmode='relative', title_text="Feature importances obtained from coefficients")
fig.show()
```

Fig. 12. Execution of Optimized Random Forest

- Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
model2 = DecisionTreeClassifier()
start2 = timer()
model2.fit(X_train_scaled, y_train)
end2 = timer()
y_predicted2 = model2.predict(x_)
importances3 = pd.DataFrame(data={
    'Attribute': X_train.columns,
    'Importance': model2.feature_importances_
})

importances3 = importances3.sort_values(by='Importance', ascending=False)

fig = px.bar(importances3, x="Attribute", y="Importance", text_auto=True)
fig.update_layout(barmode='relative', title_text="Feature importances obtained from coefficients")
fig.show()
```

Fig. 13. Execution of Decision Tree

- Regularized Decision Tree

```
from sklearn.tree import DecisionTreeRegressor
from sklearn.tree import plot_tree
tree_regressor = DecisionTreeRegressor(max_depth=5)
start3 = timer()
tree_regressor.fit(X_train_scaled, y_train)
end3 = timer()
y_predicted3 = tree_regressor.predict(x_)
fig = plt.figure(figsize=(60,55))
_ = plot_tree(tree_regressor,
                    filled=True)
```

Fig. 14. Execution of Regularized Decision Tree

- Gaussian Naive Bayes classifier

```
gnb = GaussianNB()
start = timer()
gnb.fit(X_train_scaled, y_train)
end = timer()
y_predictedg = gnb.predict(x_)
```

Fig. 15. Execution of Gaussian Naive Bayes classifier

G. *Predictive Model*

The crucial stage is model building. The aforementioned machine learning methods are all used early on to predict diabetes.

- 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.
- Divide the data into 75%-25% for Training and Testing, accordingly.
- Select algorithms from 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.
- Compare the effectiveness of the experimental findings for each classifier.
- After reviewing the data, decide which algorithm performs the best.

IV. RESULTS ANALYSIS

All experiments were undertaken using a computer with an AMD Ryzen 5 5600H with Radeon Graphics 3.30 GHz processor and 16 GB of RAM. The experimental findings for the suggested models are shown in this section. The prediction models were constructed using the Logistic Regression, Random Forest, and Decision Trees algorithms, and their performance was assessed using the accuracy, precision, recall, and F1-score metrics. 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. The below figures(Fig. 16. and Fig. 17.) are a Correlation test matrix and a Correlation Test Bar Graph, which shows the correlation coefficients as indicators of the strength of the linear relationship between two different variables, x and y.

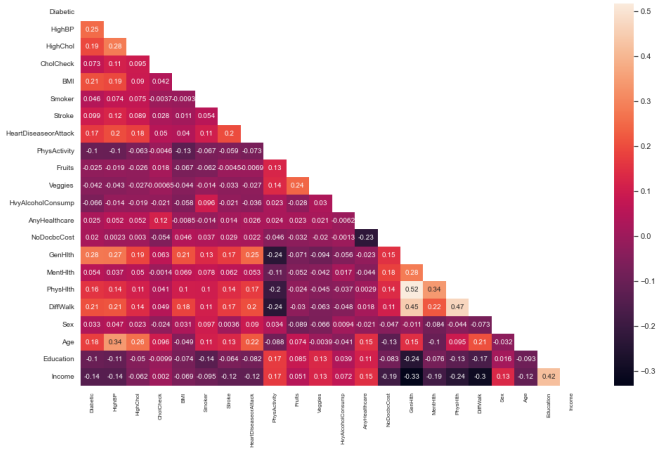

Fig. 16. Correlation test matrix

A linear correlation coefficient that is greater than zero indicates a positive relationship. A value that is less than zero signifies a negative relationship. Finally, a value of zero indicates no relationship between the two variables x and y. 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. A value less than zero denotes a negative association, while a value greater than zero denotes a positive relationship. Zero means there is no correlation between the two variables under comparison. A crucial idea in building diversified portfolios that may better resist portfolio volatility is the concept of a negative correlation, often known as an inverse correlation. Visibly from the above correlation testing bar graph, 15 independent variables have a positive relationship with the target variable and the rest have a negative association as denoted 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.

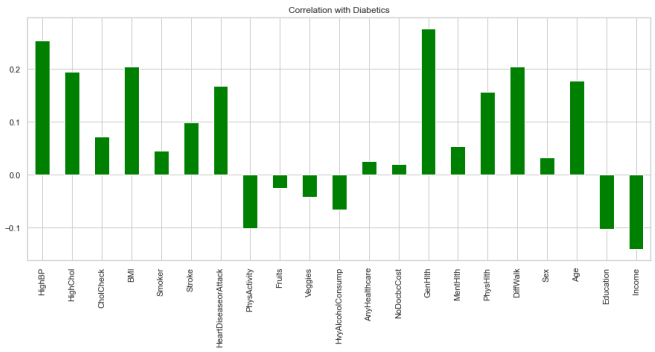

Fig. 17. Correlation Test Bar Graph

*1) **The relationship between diabetic patients and High Blood Pressure**:* Hypertension, another name for high blood pressure, is elevated blood pressure. Depending on one's activity, a person's blood pressure changes throughout the day. A diagnosis of high blood pressure may be made if blood pressure readings are frequently higher than normal (or hypertension).

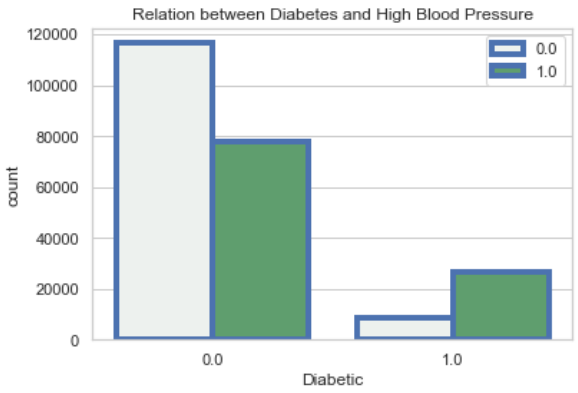

Fig. 18. The graphical representation of the relation between diabetic patients and High Blood Pressure factor

As it can be visibly seen in the above figure(Fig. 18.) double bar graph of the recorded data, when patients are non-diabetic or do not have diabetic disease, they tend to have less High Blood Pressure presence. But furthermore, when patients have been diagnosed with diabetes they tend to show a high presence of High Blood Pressure. The presence of High Blood Pressure in diabetic patients can often lead to many complications of diabetes, including diabetic eye disease and kidney disease, or make them worse. Most people with

diabetes will eventually have high blood pressure, along with other heart and circulation problems.

*2) **The relationship between diabetic patients and High Cholesterol**:* If one has type 2 diabetes, that person might have high cholesterol levels, too. With type 2 diabetes, one's body doesn't regulate or use glucose (sugar) the way it should [27]. That can lead to too-high levels of glucose in your blood. High glucose levels can contribute to other health conditions, including high cholesterol. But even people with type 2 diabetes who have well-controlled blood sugar may have cholesterol problems.

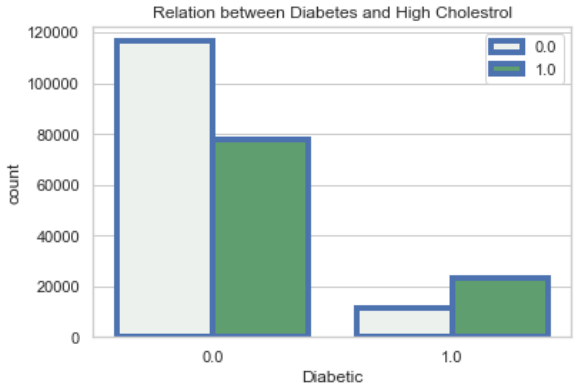

Fig. 19. The graphical representation of the relation between diabetic patients and High Cholesterol factor

Similar readings to that of the relationship between diabetic patients and High Blood Pressure appeared also on the above double bar graph in Fig. 19., patients who do not have diabetes or do not have the condition tend to have lower levels of high cholesterol but there's a significant amount of patients with high cholesterol levels. However, patients who have received a diabetes diagnosis also frequently exhibit elevated cholesterol levels. When a patient has diabetes and also has low levels of good cholesterol but high levels of bad cholesterol and high triglycerides, you have a condition called diabetic dyslipidemia. Up to 70% of people with type 2 diabetes have diabetic dyslipidemia. Additionally, type 1 diabetics with well-controlled blood sugar levels typically have normal cholesterol levels. However, individuals are more prone to have high cholesterol if they are obese or overweight.

*3) **The relationship between diabetic patients and Smoking**:* One of the causes of type 2 diabetes is smoking. 2 In actuality, compared to non-smokers, cigarette smokers have a 30%–40% increased risk of type 2 diabetes [25]. Smokers with diabetes are more prone than non-smokers to experience difficulties with insulin doses and maintaining their illness [25]. The likelihood of being positively diagnosed with type 2 diabetes increases with cigarette consumption [26].

Clearly visible seen from Fig. 20., there seems to be an almost balanced number of people who smoke and do not smoke when patients are diabetic. However, smoking makes diabetes harder to manage. If a patient has diabetes and the

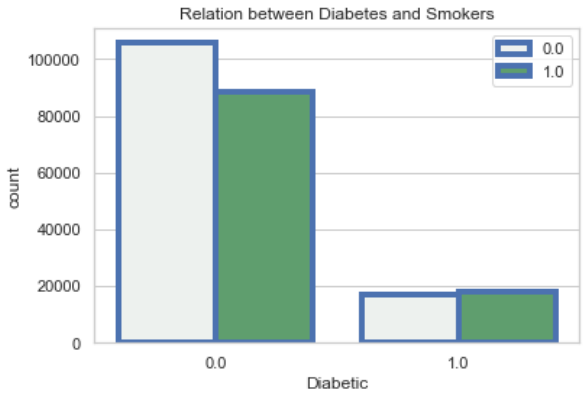

Fig. 20. The graphical representation of the relation between diabetic patients and Smoking factor

patient smokes, they are more likely to have serious health problems from diabetes.

*4) **The relationship between diabetic patients and Heavy Alcohol consumption**:* Chronic alcohol consumption is thought to increase the risk of type 2 diabetes mellitus (T2DM), which leads to insulin resistance and pancreatic beta-cell dysfunction, both of which are necessary conditions for the onset of diabetes [28]. The relationship between alcohol use and diabetes has been debatable, thus further research on how alcohol affects diabetes appears necessary.

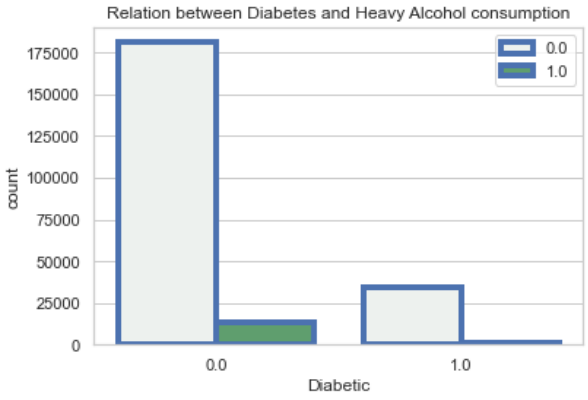

Fig. 21. The graphical representation of the relation between diabetic patients and Heavy Alcohol consumption factor

In the above double bar graph figure(Fig. 21.), we have a very interesting relation visualization, when patients were recorded to not be diabetic, a high number of them showed that they did not have Heavy Alcohol consumption, and a very small count of non-diabetic patients had Heavy Alcohol consumption. When patients are diabetic, there seemed to be a very low number of patients with Heavy Alcohol consumption. Diabetes, especially T2DM, causes dysregulation of various metabolic processes, which includes a defect in the insulin-mediated glucose function of adipocytes, and impaired insulin

action in the liver. In addition, neurobiological profiles of alcoholism are linked to the effects of a disruption of glucose homeostasis and of insulin resistance, which is affected by an altered appetite that regulates the peptides and neurotrophic factors. Since conditions, which precede the onset of diabetes that are associated with alcoholism are crucial public problems, research in efforts to prevent and treat diabetes with alcohol dependence, receives special clinical interest.

5) ***The relationship between diabetic patients and Heart Disease or Attack****:* Diabetes patients frequently develop heart problems. According to National Cardiac Association data from 2012, 65% of diabetics will pass away from a heart condition or a stroke [29]. People with diabetes are more than twice as likely to die from heart disease and stroke [29].

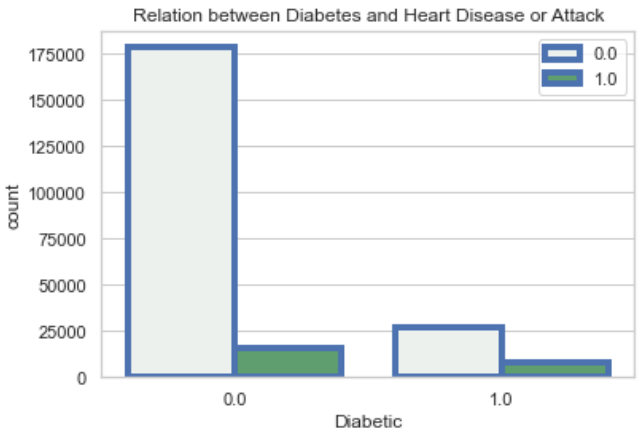

Fig. 22. The graphical representation of the relation between diabetic patients and Heart Disease or Attack factor

From the above double bar graph figure(Fig. 22.), regardless of whether the patient's recorded data showed that the patient had diabetes or not, there's a high number of people who were not affected by the heart disease or attack. While all people with diabetes have an increased chance of developing heart disease, the condition is more common in those with type 2 diabetes [30]. In fact, heart disease is the number one cause of death among people with type 2 diabetes [29].

6) ***The relationship between diabetic patients and Physical Activity****:* Physical activity improves glycemic control and reduces the risk of cardiovascular disease and mortality in patients with type 2 diabetes (T2D) [31]. Moderate to vigorous physical activity is recommended to manage T2D as shown in [31]; however, patients with T2D can be physically weak, making it difficult to engage in the recommended levels of physical activity [32]. With physical activity, the recorded data shows that when more patients do more physical activities they tend to be less prone to be diagnosed with diabetes, furthermore the data also suggests that people who take part in physical activities can be diagnosed with diabetes but can use physical activities to manage the condition. Daily physical activity includes various activities performed during both occupational and leisure time such as walking, gardening, and housework that diabetic patients should be able to perform without considerable physical burden [31]. This research focused on the association between daily physical activity and diabetes. Walking was the most common form of daily physical activity, with numerous studies demonstrating its beneficial effects on reducing the risk of T2D, cardiovascular disease, and mortality.

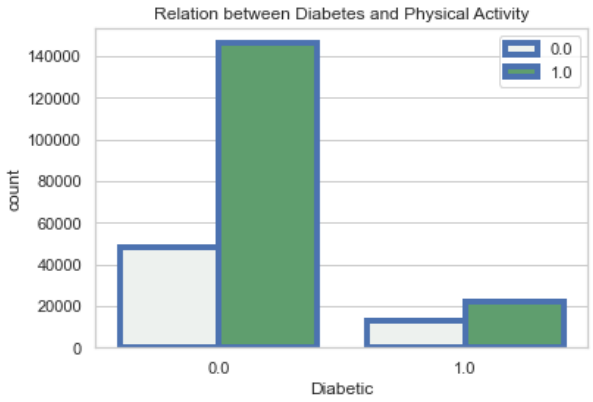

Fig. 23. The graphical representation of the relation between diabetic patients and Physical Activity factor

7) ***The relationship between diabetic patients and Vegetable intake****:* The increase in the incidence of diabetes has been attributed in part to high-fat, high-calorie diets, overweight and obesity – particularly excess abdominal fat, and lack of exercise [33]. These factors are associated with insulin resistance and metabolic syndrome- important risk factors for diabetes and cardiovascular disease. Estimates suggest that up to a 75% reduction in risk for diabetes could be achieved by preventing obesity [33]. Observations from population-based studies suggest that fruit and vegetable consumption may be associated with a reduced risk of diabetes or may be protective as shown in the above figure(Fig. 24.), although results have not always been consistent. A positive association between fruits and vegetables and reduced risk is supported by an inverse association between serum carotenoids, a marker for

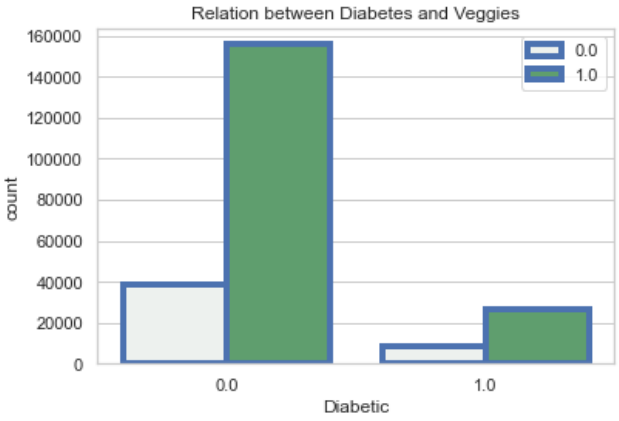

Fig. 24. The graphical representation of the relation between diabetic patients and Vegetables intake factor

fruit and vegetable intake, and diabetes and impaired glucose metabolism in adults [34]. Dietary patterns that include fruits and vegetables have been associated with a reduction in fasting blood glucose, improved results on glucose tolerance tests lower glycosylated hemoglobin values and enhanced insulin sensitivity [34]. However, energy intake may modify these associations. Taken together, these studies suggest that fruit and vegetable intake is potentially beneficial for the prevention and management of diabetes.

*8) **The relationship between diabetic patients and Age**:* Pre-diabetes and diabetes both had observed prevalences of 67% and 22% [35], respectively. 10% of girls and 6% of males who had never had a diabetes test before the study were found to have the disease [35], and 67% of both sexes were determined to be pre-diabetic [35]. As a result, a sizable fraction of South Africans are undiagnosed [35]. Significant interactions between various lifestyle, demographic, and anthropometric variables were also discovered, indicating that the influence each of these variables has on a person's chance of developing pre-diabetes or diabetes is complicated by additional variables [35]. We know that as age increases, the chances of diabetes

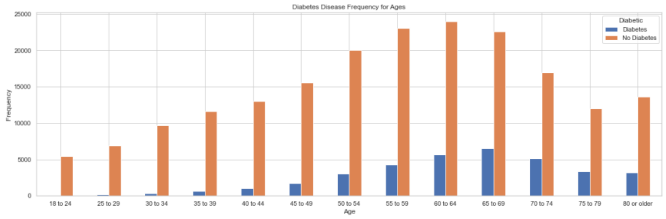

Fig. 25. The graphical representation of the relation between diabetic patients and Age factor

also commonly increase. From above figure(Fig. 25.) we can say, the median age of diabetic people is greater than that of non-diabetic people. When you reach your senior years, the prevalence of diabetes increases even more, as seen in the above double-bar graph figure(Fig. 25.). Diabetes sickness also is impacting ever more youths and even children. Potential harm from diabetes and age upon diagnosis are inversely correlated. The potential danger is greater the younger you are when it occurs.

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 accurately identify the condition, and in our instance, it is used to identify the individuals who have been given a diabetes diagnosis, or the number of persons who tested positive.
- **Specificity:** To determine who is healthy, that is, those who do not have diabetes or who tested negative
- **F-Measure:** The Dice Similarity Coefficient (DSC), commonly referred to as the F-measure or F1 Score, is the Harmonic Mean between recall and precision. F1 Score has a range of [0, 1]. It reveals the precision and durability of your classifier (the proportion of correctly classified cases) (it does not miss a significant number of instances).
- **10-fold Cross validation:** Through the division of the original sample into a training set and a test set, cross-validation is a technique for assessing prediction models.
- **Precision and Recall:** The proportion of relevant examples among the retrieved instances is known as precision (also known as positive predictive value), whereas the proportion of relevant instances that were retrieved is known as recall (also known as sensitivity). Thus, relevance serves as the foundation for both precision and recall.
- **Accuracy:** how accurately our method has predicted diabetic patients as diabetic and nondiabetic patients as non-diabetic
- **True positives:** The number of patients with diabetes who were appropriately diagnosed, is shown in the first quadrant
- **False positive:** Patients without diabetes who were mistakenly identified as having the condition
- **True negative:** Non-diabetic people correctly identified as non-diabetic
- **False negative:** Diabetic people incorrectly identified as non-diabetic

Metric	Definition
Accuracy	$= \frac{TP+TN}{TP+FP+FN+TN}$
Precision	$= \frac{TP}{TP+FP}$
Recall	$= \frac{TP}{TP+FN}$
F1-score	$= \frac{2*(\text{recall}*\text{precision})}{\text{recall}+\text{precision}}$

TP = true positive, TN = true negative, FP = false positive, FN = false negative.

Fig. 26. Evaluation Metrics

C. *Feature Selection Results*

Depending on the criteria being assessed, feature selection extracts a subset of pertinent characteristics from the input dataset. There are n subsets created from a collection of features. The features are arranged in ascending relevance order. Between a feature vector and its neighboring feature vector, redundancy may exist. Symmetric uncertainty is used to reduce duplication between two feature vectors. If the dataset contains two duplicate features, we can eliminate one of them since they will virtually always produce the same outcome. 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.

Two phases combined make the proposed in the proposed bi-directional feature selection approach in the final phase. In

the first phase, the sequential forward feature selection filter method was used to rank features between the features and target class. In this first phase, the objective is to select the subset of features that are most relevant to the job at hand, maximizing computing efficiency and minimizing generalization error by removing unimportant data (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. Results are displayed in Fig. 27.

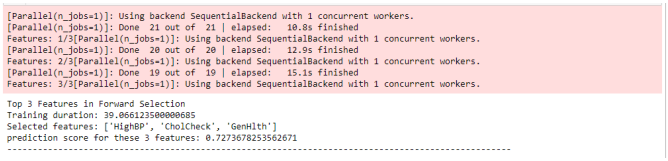

Fig. 27. Top 3 features in Sequential Forward Feature Selection

- **Top 5 features were selected in the second Sequential Forward Feature Selection iteration:** The selection score is 72.48%, in 82.49 seconds. Results are displayed in Fig. 28.

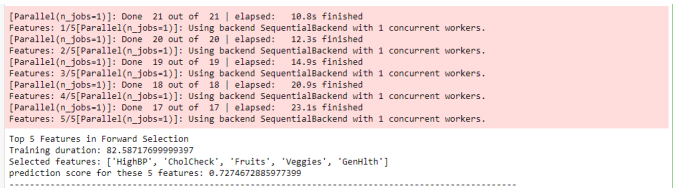

Fig. 28. Top 5 features in Sequential Forward Feature Selection

- **Top 7 features were selected in the third Sequential Forward Feature Selection iteration:** The selection score is 72.43%, in 135.96 seconds. Results are displayed in Fig. 29.

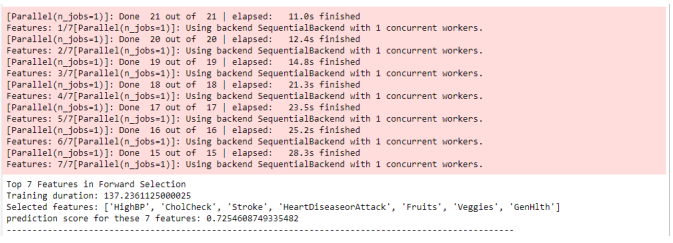

Fig. 29. Top 7 features in Sequential Forward Feature Selection

- **Top 10 features were selected in the fourth Sequential Forward Feature Selection iteration:** The selection score is 72.48%, in 236.28 seconds. Results are displayed in Fig. 30.

furthermore in order to find out the optimal number of significant features, we can use the hit and trial method for different values of k features and make the final decision by

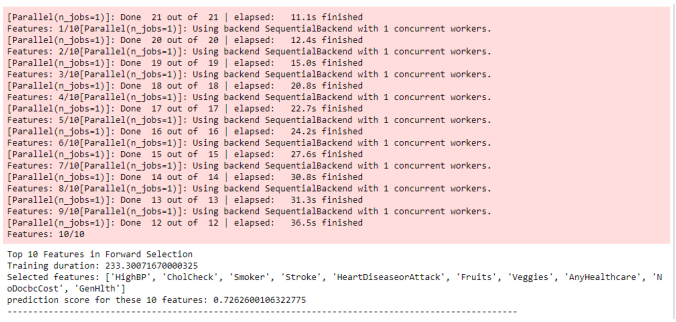

Fig. 30. Top 10 features in Sequential Forward Feature Selection

plotting it against the model performance. Fig. 31., is a visualization of the Sequential Forward Feature Selection selecting the specific significant number of features for determining diabetes.

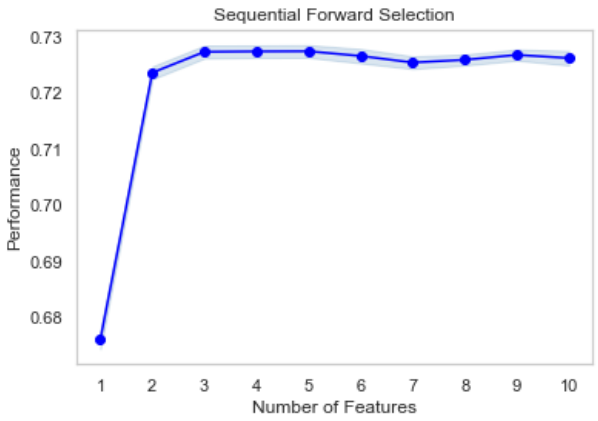

Fig. 31. Sequential Forward Feature Selection Performance

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. Results are displayed in Fig. 32.

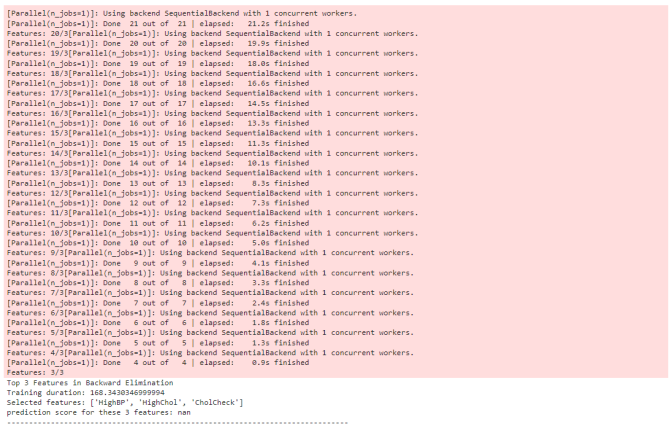

Fig. 32. Top 3 features in Sequential Backward Elimination Selection

- **Top 5 features were selected in the second sequential backward elimination selection iteration:** The selection duration is 119.29 seconds. Results are displayed in Fig. 33.

Fig. 33. Top 5 features in Sequential Backward Elimination Selection

- **Top 7 features were selected in the first sequential backward elimination selection iteration:** The selection duration is 116.52 seconds. Results are displayed in Fig. 34.

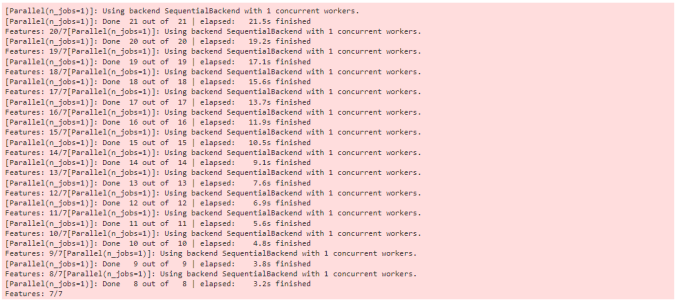

Fig. 34. Top 7 features in Sequential Backward Elimination Selection

- **Top 10 features were selected in the first sequential backward elimination selection iteration:** The selection duration is 107.43 seconds. Results are displayed in Fig. 35.

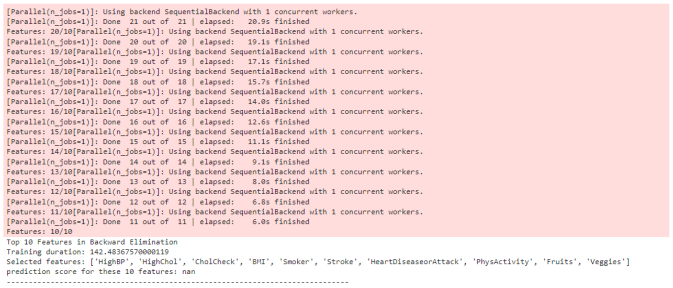

Fig. 35. Top 10 features in Sequential Backward Elimination Selection

In the final phase what we proposed comes through an exhaustive search approach of combining Sequential Forward and Backward Elimination Feature Selection, this makes up what we call the bi-directional feature selection approach which was proposed. In order to create at least two intermediate feature subsets, the Forward Selection and Backward Elimination methods are first used to the original feature set in this combination of a technique, which is also referred to as the floating search. The same number of features are selected from each subset in accordance with the majority of votes in the second step, which involves combining two subsets into a single pool. The feature with the higher pattern categorization performance rate when the other features are combined will win out if two features obtain an equal number of votes. In the third step, an exhaustive procedure will choose the final feature set based on the features that have garnered the most 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. Results are displayed in Fig. 36.

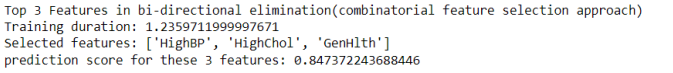

Fig. 36. Top 3 features in bi-directional feature selection approach

- **Top 5 features were selected in the second bi-directional feature selection approach iteration:** The selection score is 85.05%, in 10.32 seconds. Results are displayed in Fig. 37.

```
Top 5 Features in bi-directional elimination(combinatorial feature selection approach)
Training duration: 10.327816600001825
Selected features: ['HighBP', 'HighChol', 'CholCheck', 'BMI', 'GenHlth']
prediction score for these 5 features: 0.850544725603556
--------------------------------------------------------------------------------
```

Fig. 37. Top 5 features in bi-directional feature selection approach

- **Top 7 features were selected in the third bi-directional feature selection approach iteration:** The selection score is 85.08%, in 17.98 seconds. Results are displayed in Fig. 38.

```
Top 7 Features in bi-directional elimination(combinatorial feature selection approach)
Training duration: 17.989302800000587
Selected features: ['HighBP', 'HighChol', 'CholCheck', 'BMI', 'Stroke', 'HeartDiseaseorAttack', 'GenHlth']
prediction score for these 7 features: 0.8508061938932628
--------------------------------------------------------------------------------
```

Fig. 38. Top 7 features in bi-directional feature selection approach

- **Top 10 features were selected in the fourth bi-directional feature selection approach iteration:** The selection score is 85.10%, in 60.55 seconds. Results are displayed in Fig. 39.

```
Top 10 Features in bi-directional elimination(combinatorial feature selection approach)
Training duration: 60.5573570000201
Selected features: ['HighBP', 'HighChol', 'CholCheck', 'BMI', 'Smoker', 'Stroke', 'HeartDiseaseorAttack', 'PhysActivity', 'Fruits', 'NoDocbcCost', 'GenHlth']
prediction score for these 10 features: 0.8510676621829697
--------------------------------------------------------------------------------
```

Fig. 39. Top 10 features in bi-directional feature selection approach

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 means these features, at least in a statistical sense, are the most influential for predicting diabetes disease. This information can be very useful to clinicians because when diagnosing a patient they can start by testing for the most influential features before the least influential ones.

Furthermore, medical diagnosis is a very essential and critical aspect for healthcare professionals. In particular, the classification of diabetics is very complex. Early identification of diabetes is much important in controlling diabetes. A patient has to go through several tests and later it is very difficult for the professionals to keep track of multiple factors at the time of the diagnosis process which can lead to inaccurate results which makes the detection very challenging. 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. 40.) 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.

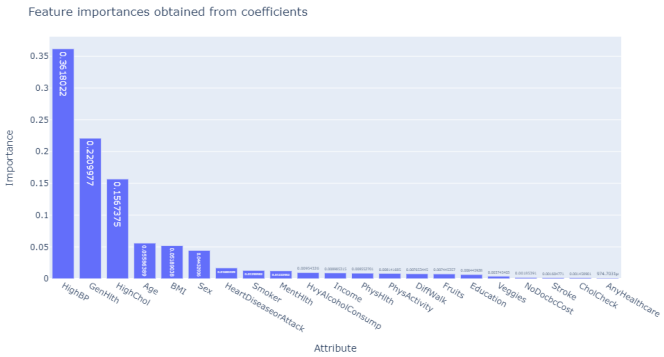

Fig. 40. Overall Regularized Trees with Overall Feature Importance

D. *Classification Methods Results*

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

The pooled diabetes dataset was subjected to feature selection and extraction, as shown in the above table(Fig. 41.), 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 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. 42. Accuracy Metric measure on 4 Machine Learning Algorithms on regularized trees approach produced dataset(Also on the unselected dataset)

Moreover, our proposed regularized decision tree feature selection approach, its implementation was made possible by altering the 'max depth', 'min sample split', 'min sample leaf', and 'max feature size', in order to stop the further splitting of nodes when a number of observation in the node is less than given value and further also set the minimum number of sample a leaf must-have when a leaf node has too few observations further split which might result in over-fitting. As shown in the above table(Fig. 42.), the optimized random forest was seen to have had the highest accuracy on the Top 5 and Top 10 feature sets of 84.63% and 83.18% respectively, whilst the decision tree had the highest accuracy of 84.05% on the Top 7 feature set. Clinicians may opt to select a model that is simple to implement without suffering a significant loss in accuracy due to the little variances across models across machine learning approaches. 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 little attention because the majority of studies have concentrated on the prediction of undiagnosed diabetes. Compared to the works stated above, this study used a big dataset and ensemble machine-learning approaches to create the prediction models. Additionally, by altering the number of years utilized to train the models, the effect of the accumulated medical data on prediction accuracy was also demonstrated. To develop predictors that were effective in identifying the different classes in the dataset, a data-driven feature selection method was used. 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. Compared to using the traditional five predictors of diabetes ('HighBP', 'GenHlth', 'BMI', 'age', and 'sex'), the proposed models employing the selected features showed a superior prediction performance. When more than 200,000 entries of data were utilized in training, the maximum machine learning algorithm accuracy was 90.77% for the unselected feature set and 85.01% for the Top 5 selected feature sets. It can be concluded that the use of a large amount of cleaned data contributed to improved accuracy of prediction.

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 study recommends two additional studies that are worthwhile conducting. To reduce the complexity of identifying diabetes, which results from the overlap between normal and diabetes classes, the first step would be to include a variety of datasets. The second would be to make the prediction models more accessible and enhance the usability of web and mobile applications.

VI. CONCLUSIONS

The goal of the study was to develop and use a viable ML model to predict diabetes at an early stage. And after examining many Machine Learning models, we came to the conclusion that, with the aid of appropriate hyperparameter adjustment, the Optimized Random Forest model had the greatest testing accuracy of 91.0% on the unselected feature set and 85.01% for the Top 5 selected feature set. Additionally, our accuracy was impacted by the features we selected, the scaling mechanism we used, the hyperparameters, and the class imbalance we eliminated. A different outcome could result from altering the aforementioned factors.

Our findings demonstrate that by applying feature selection and extraction to the diabetes disease dataset, it is possible to produce a model for diabetes prediction that is more accurate. To obtain the best model possible, it is required to evaluate a wide range of combinations of feature selection approaches with machine learning algorithms. The improvements over utilizing the original dataset rely largely on the machine learning algorithm utilized. The ranking of features by our feature selection methods(both combinatorial and regularized trees) shows us that 'HighBP' is universally the most influential feature for predicting heart disease followed by 'GenHlth', 'HighChol' and 'BMI' features; however, these features are ranked differently across feature selection methods. If patient data for the relevant factors are available, clinicians and other healthcare professionals can utilize the models created in this work to detect diabetes in new patients. Knowing which specific factors were used during pre-processing is

also beneficial because it tells which are more statistically significant for predicting diabetic illness.

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: There was no external support for this study.

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] Choi, B.G.; Rha, S.-W.; Kim, S.W.; Kang, J.H.; Park, J.Y.; Noh, Y.-K. Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in nondiabetic patients with cardiovascular risks. Yonsei Med. J. 2019, 60, 191–199.

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

[9] 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

[10] 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

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

[12] Rahimloo, P. and Jafarian, A. (2016). Prediction of diabetes by using artificial neural network, logistic regression statistical model and combination of them. Bulletin de la Soci´et´e Royale des Sciences de Li'ege, 85:1148–1164

[13] Ronaghan, S. The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark. Available online: https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3 (accessed on 9 March 2021).

[14] Inoue, K.; Matsumoto, M.; Kobayashi, Y. The combination of fasting plasma glucose and glycosylated hemoglobin predicts type 2 diabetes in Japanese workers. Diabetes Res. Clin. Pract. 2007, 77, 451–458.

[15] 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.

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

[17] 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.

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

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] Cateni, Silvia, et al. "Variable selection and feature extraction through artificial intelligence techniques." Multivariate analysis in management, engineering and the Science 6 (2012): 103-118.

[25] Yeh, Hsin-Chieh, et al. "Smoking, smoking cessation, and risk for type 2 diabetes mellitus: a cohort study." Annals of internal medicine 152.1 (2010): 10-17.

[26] Meisinger, Christa, et al. "Association of cigarette smoking and tar and nicotine intake with development of type 2 diabetes mellitus in men and women from the general population: the MONICA/KORA Augsburg Cohort Study." Diabetologia 49.8 (2006): 1770-1776.

[27] Can, I. "What is diabetes?." (2008).

[28] Kim, Soo-Jeong, and Dai-Jin Kim. "Alcoholism and diabetes mellitus." Diabetes metabolism journal 36.2 (2012): 108-115.

[29] Writing Group Members, et al. "Heart disease and stroke statistics—2012 update: a report from the American Heart Association." Circulation 125.1 (2012): e2-e220.

[30] Glovaci, Diana, Wenjun Fan, and Nathan D. Wong. "Epidemiology of diabetes mellitus and cardiovascular disease." Current cardiology reports 21.4 (2019): 1-8.

[31] Liu, Gang, et al. "Influence of lifestyle on incident cardiovascular disease and mortality in patients with diabetes mellitus." Journal of the American College of Cardiology 71.25 (2018): 2867-2876.

[32] Hu, Jinbo, et al. "Weight change, lifestyle, and mortality in patients with type 2 diabetes." The Journal of Clinical Endocrinology Metabolism 107.3 (2022): 627-637.

[33] Ghanim, Husam, et al. "Liraglutide treatment in overweight and obese patients with type 1 diabetes: a 26-week randomized controlled trial; mechanisms of weight loss." Diabetes, Obesity and Metabolism 22.10 (2020): 1742-1752.

[34] Ford, Earl S., and Ali H. Mokdad. "Fruit and vegetable consumption and diabetes mellitus incidence among US adults." Preventive medicine 32.1 (2001): 33-39.

[35] Khan, Radia Marium Modhumi, et al. "From pre-diabetes to diabetes: diagnosis, treatments and translational research." Medicina 55.9 (2019): 546.