**Diabetes Prediction from Health Indicators**

**Milestone: Final Project Report**

Group 24

Celia Harris

Anirudh Hegde

[harris.cel@northeastern.edu](mailto:harris.cel@northeastern.edu)

[hegde.anir@northeastern.edu](mailto:hegde.anir@northeastern.edu)

**Percentage of Effort Contributed by Student 1:** 50%

**Percentage of Effort Contributed by Student 2:** 50%

**Signature of Student 1:** Celia Harris

**Signature of Student 2:** Anirudh Hegde

**Submission Date:** 12th April 2024

Problem Setting

Diabetes is a prevalent chronic disease in the US, resulting from an inability to regulate glucose levels. Diabetes can lead to severe complications and life-long health issues. There is no cure for diabetes, however, if it is diagnosed at an early stage, treatments can be administered to prevent serious complications.

Problem Definition

Risk factors – such as obesity, poor diet, and lack of physical activity – can be precursors to diabetes. These factors can be used to predict whether a patient is at high risk of developing diabetes, so that the patient can take preventative action and avoid life-threatening complications resulting from untreated diabetes. Questions to be explored include whether a patient is likely to develop diabetes based on risk factors and which risk factors are the most useful and accurate in predicting diabetes risk.

Data Source

The data source is ‘Diabetes Health Indicators Dataset’ from the Behavioral Risk Factor Surveillance System (BRFSS), an annual survey that collects health-related data from US citizens: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data>

Data Description

This dataset includes a target variable column indicating whether a survey respondent has diabetes or no diabetes. Other columns include health indicators, such blood pressure, BMI, cholesterol, and stroke & heart disease history, and lifestyle habits, such as diet, alcohol consumption, smoking, education, and exercise. The Dataset contains 253,680 rows with 22 columns. All variables are in a numeric format and the data contains no missing values. Several columns, such as education, are in a numeric format, with numbers corresponding to set categories.

Data Exploration

The BRFSS database selected for this project is a subset of the survey responses collected in 2015, consisting of 22 indicators that are related to diabetes. The data includes 253,680 responses. The target variable is ‘Diabetes\_binary’ and it describes whether a patient has no diabetes (0) or has diabetes (1). There are 14 categorial variables, all of which are in binary format. There are seven numeric variables, many of which have been encoded, such that the numbers correspond to certain groups. ‘BMI’ is the only continuous variable, whereas all other numeric variables are discrete. All variables are described below in Table 1 and Table 2.

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| HighBP | Patient has been diagnosed with high blood pressure (1=true, 0=false). |
| HighChol | Patient has been diagnosed with high cholesterol (1=true, 0=false). |
| CholCheck | Patient has had their cholesterol level checked in the past five years (1=true, 0=false). |
| Smoker | Patient has smoked at least 100 cigarettes in their lifetime (1=true, 0=false). |
| Stroke | Patient has had a stroke (1=true, 0=false). |
| HeartDiseaseorAttack | Patient has had a myocardial infraction or coronary heart disease (1=true, 0=false). |
| PhysActivity | Patient has had physical activity or exercise in the past 30 days (1=true, 0=false). |
| Fruits | Patient consumes one or more fruits per day (1=true, 0=false). |
| Veggies | Patient consumes one or more vegetables per day (1=true, 0=false). |
| HvyAlcoholConsump | Men who have more than 14 drinks per week, or women who have more than 7 drinks per week (1=true, 0=false). |
| AnyHealthcare | Patient has any kind of healthcare coverage (1=true, 0=false). |
| NoDocbcCost | Patient needed to see a doctor but could not because of cost at least one time in the past 12 months (1=true, 0=false). |
| DiffWalk | Patient has difficulty walking or climbing stairs (1=true, 0=false). |
| Sex | ‘0’ is assigned to females, ‘1’ is assigned to males. |

Table 1: Descriptions of binary variables.

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| BMI | Body mass index. |
| GenHlth | General health status on a scale from 1 (excellent) to 5 (poor). |
| MentHlth | Number of days out of the past 30 days, for which the patient had poor mental health. |
| PhysHlth | Number of days out of the past 30 days, for which the patient had poor physical health. |
| Age | Age, divided into categories, ranging from category 1 (ages 18-24) to category 13 (ages 80 and above). |
| Education | Highest level of schooling completed by patient, with responses ranging from 1 (never attended any school) to 6 (graduated college). |
| Income | Annual income, divided into 8 brackets, from 1 (< $10,000) to 8 (> $75,000). |

Table 2: Descriptions of numeric variables.

Figure 1 below provide insights into the central tendency and shape of the dataset. Figure 2 shows that the dataset contains no null values, thus, no corrective action is needed to account for missing values.

|  |  |
| --- | --- |
| **A table of numbers and symbols  Description automatically generated**  Figure 1: Details on central tendency and shape of the data. | **A screenshot of a test  Description automatically generated**  Figure 2: Count of null values present in each variable. |

A pie chart was generated to examine the target variable and is shown in Figure 3. Out of the total 253,680 responses, 13.9% are diabetic and 86.1%, the majority, are not diabetic. In the US, about 11.6% of the population has diabetes [1]. The prevalence of diabetes cases in the dataset and in the American population are very similar.

A blue and orange pie chart

Description automatically generated

Figure 3: Pie chart depicting the target variable and the prevalence of its categories: diabetic & healthy.

The binary categorial variables were examined using bar charts, displayed in Figure 4. Each variable was plotted on a separate bar chart, with the x-axis showing each possible category (0 or 1), and the y-axis showing the count of instances within the dataset. The graphs show that relatively few patients have had a stroke, have had heart disease, are heavy consumers of alcohol, have no healthcare coverage, have not been able to visit a doctor because of cost barriers, and have not had their cholesterol checked recently. The graphs also show that the majority of survey responders eat fruits and vegetables regularly, have no difficulty walking, are physically active, are not smokers, and have normal blood pressure and cholesterol levels. Additionally, the respondents consist of more females than males.

A group of blue and orange bars

Description automatically generatedA comparison of a bar graph

Description automatically generated with medium confidence

Figure 4: Count plots for each binary predictor variable.

Histograms were created for each of the numeric variables, in order to assess the variance, as shown in Figure 5. The CDC generally considers a healthy BMI to be between 18.5 and 25 [2]. When people have high BMI values, they may be considered overweight and are at higher risk of developing diabetes [2]. The histogram shows that the BMI data is centered around 25, and there are very few instances of patients being underweight (with a BMI less than 18.5), but a large portion of patients who can be considered overweight (with a BMI over 25). The histogram for GenHlth reveals that the majority of patients describe their health status as very good (2) or good (3), and relatively few who characterize their general health as poor (5). The histograms for MentHlth and PhysHlth appear to be very similar to each other. Both are highly right skewed, and over 70% of survey respondents have indicated that they only experienced poor mental and physical health for 0 to 5 days out of the past 30 days. Age seems to have a high variance, and is centered around category 9, which represents ages 60 to 64. Education and income appear to be left-skewed, with most patients having high levels of education and income.

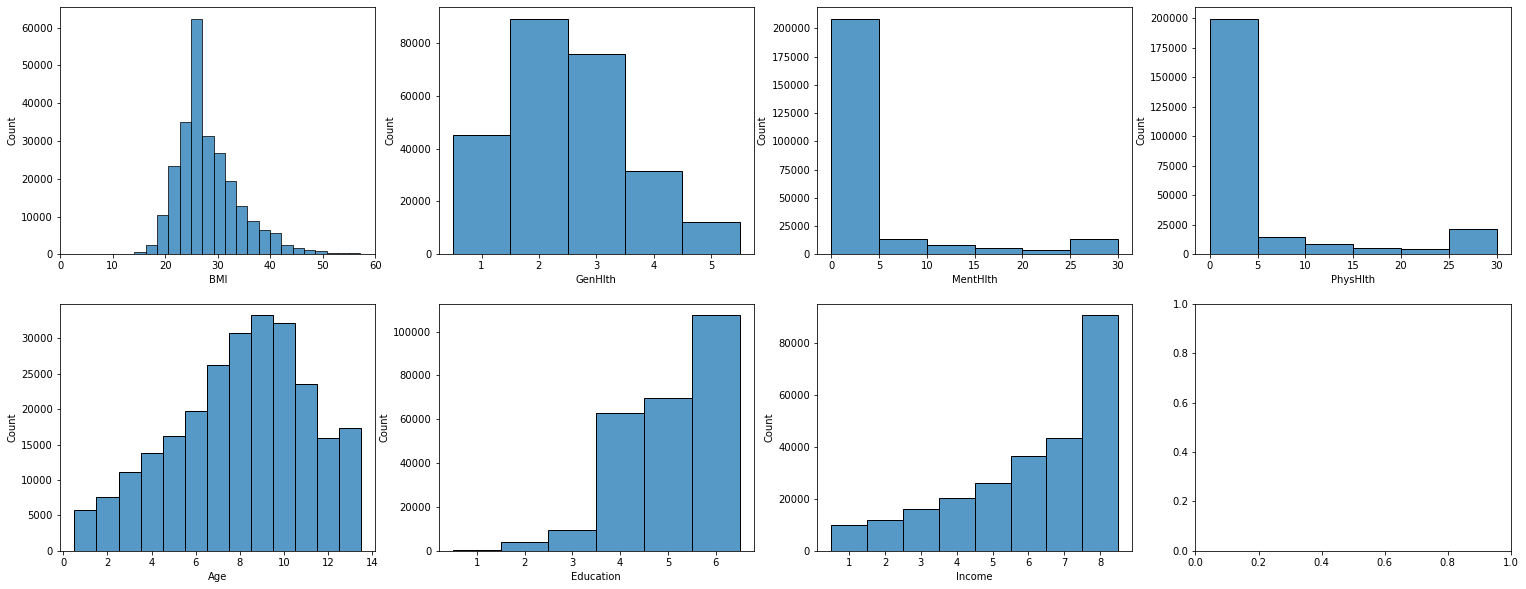
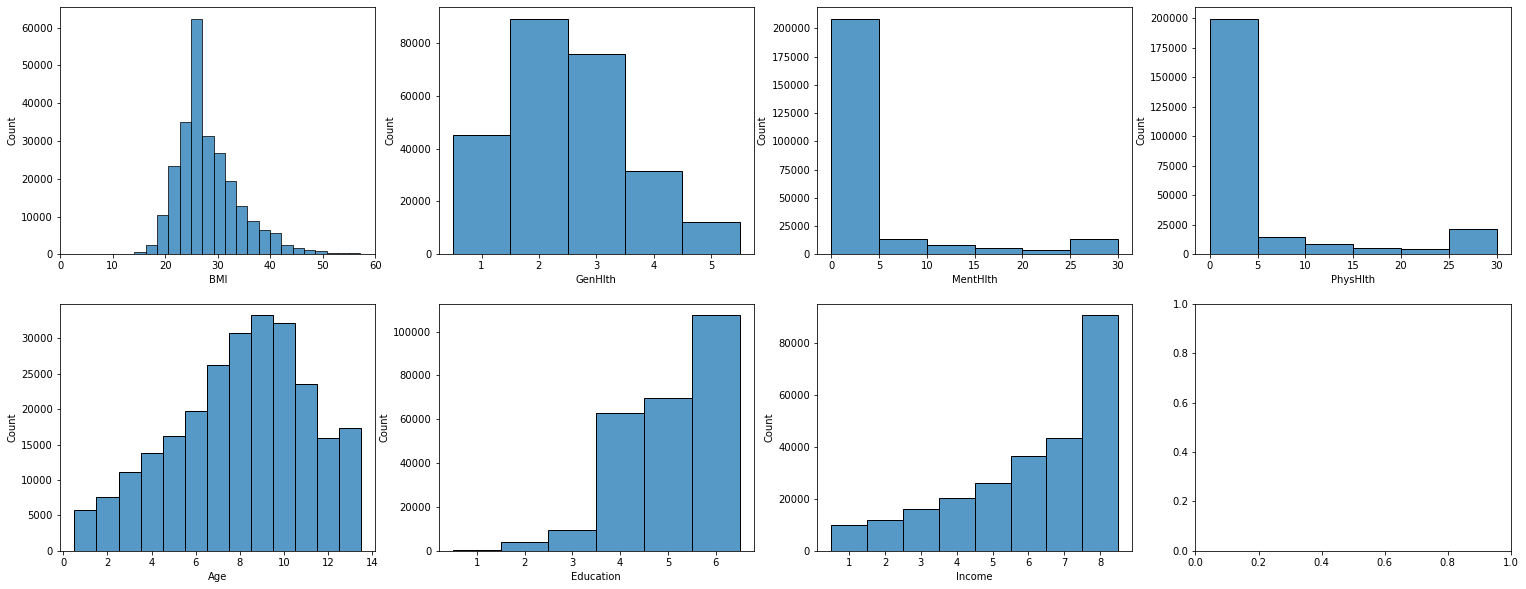


Figure 5: Histograms for the numeric variables.

A correlation heat map was generated for all variables and is shown in Figure 6. GenHlth and PhysHlth have the highest Pearson correlation coefficient of 0.52. This is likely because physical health has a great influence on how a patient rates their general heath level. Also, DiffWalk is highly correlated with PhysHlth (correlation of 0.46) and GenHlth (correlation of 0.48), because patients who have poor physical or general health are more likely to have difficulty walking. Education and Income are also highly correlated with a coefficient of 0.45, since in general, people with higher levels of education are able to attain higher-paying jobs.

A screenshot of a cell phone



Description automatically generated

Figure 6: Correlation heat map for all variables.

Data Mining Tasks

Principal component analysis was performed on the numeric variables, and chi-square test was performed on the categorial variables. The results of these analyses were used to perform dimension reduction on the dataset.

Principal Component Analysis

Principal component analysis was applied to the seven numeric variables to reduce the dataset’s dimensionality and identify the principal components that capture the most variance in the data. The dataset was standardized to ensure each feature contributes equally to the variance. The explained variance of each principal component was calculated, and it represented the amount of variance captured by each principal component. The proportion of variance explained by each principal component was calculated providing details of relative importance of each principal component. The cumulative proportion of variance was calculated to determine how many components are needed to capture around 95% of the total variance of the dataset.

A number with numbers on it

Description automatically generated with medium confidence

Figure 7: PCA analysis and scores.

Based on the PCA, no numerical variable was dropped. To retain 95% variance threshold, all 7 numerical must be retained. The line plot in Figure 8 helps us identify optimal number of principal components to retain for balanced dimensionality reduction.

**A graph with blue dots

Description automatically generated**

Figure 8: PCA cumulative variance by number of components.

Chi-Square Test

Chi-Square test was employed to check the association of each categorical variable against the target variable ‘Diabetes\_binary’. The variables with non-significant association i.e. p-value greater than the significance threshold (more than 0), were removed from the dataset.

A close-up of a text

Description automatically generated

Figure 9: Chi-square test results for each categorial variable.

After the dimension reduction that was performed based on the chi-square test results, the dataset contained 253,680 rows and 13 predictors, as shown below.

A screenshot of a computer

Description automatically generated

Figure 10: Column names and first five rows of dataset after dimension reduction.

Data Mining Models

k-Nearest Neighbors

The k-Nearest-Neighbors (k-NN) classifier uses numeric predictors to classify a categorical response variable. When there is a new record that needs to be classified, the model identifies k records from the training data that are similar to the new record. The new record is classified as the most prevalent class among the similar records that were identified. Euclidean distance is used to calculate distances between records and assess similarity. Selecting a higher value of k can reduce the risk of overfitting, but if the value is too high, the k-NN method will perform no better than a Naïve benchmark – classifying each new record as the most prevalent class in the dataset, rather than the most prevalent class of the k similar values. Accuracy can be used as an assessor to find the optimal k value in the validation dataset.

Advantages of the k-NN method are that this method is simple and does not rely on parametric assumptions. Also, this method performs well with large datasets. However, disadvantages are that considerable time may be needed to find nearest neighbors in a large dataset and, as the number of predictors increases, the number of records required in the training set to maintain dimensionality – and allow the k-NN method to function accurately – increases exponentially. Additionally, this method is not suitable for real-time predictions since the Euclidean distances need to be calculated between all training data and each new record to be classified.

Classification Tree

Classification trees is a classification method that is nonparametric, nonlinear, and uses categorical predictors. Trees operate by separating records into subgroups based on predictors and creating logical decision rules for classifying new records. A new record descends down the classification tree, through decision nodes, until it reaches a terminal node with no successive nodes. At this point, the training data that belongs to this terminal node is examined to classify the new record. Classification trees require large datasets for training. Also, classification tree sizes should be limited to prevent overfitting.

Advantages of classification tress are that they are relatively cheap-to-deploy methods that are highly automated, are not swayed by outliers, and can handle missing values. Disadvantages include that trees can fail to identify certain relationships between predictors, particularly linear ones. Also, despite their cheapness to deploy, it can be expensive to build classification trees due to the large training datasets required for tree construction. Furthermore, trees may favor predictors with many categories.

Naïve Bayes Classifier

The Naïve Bayes classifier is used for classification based on categorical predictors. It is a useful method when there is an asymmetric misclassification cost, and overclassifying records as a specific class of interest may be preferred rather than misclassifying records that actually belong to this class. The method involves identifying records from the dataset that have all the same predictor values as the new record, then classifying the new record based on the most prevalent class of those identified.

The Naïve Bayes classifier has many advantages. It is a simple and efficient method and often performs very well. On the other hand, a disadvantage is that a large dataset is needed for the method to perform well. Another disadvantage is that if a category for a predictor is rare and absent from the training data set, the new record is assigned a probability of zero. Also, the Naïve Bayes method produces highly biased results for propensity estimation, so should not be used for this application. But the method does perform well for classification based on propensities.

Logistic Regression

Logistic regression uses a logistic function, also known as a sigmoid function, to model the probability that a given observation belongs to a particular class. The logistic function outputs a value between 0 and 1, which can be interpreted as the probability of the observation belonging to the positive class. The model calculates these probabilities by applying a linear combination of the input features (weighted sum plus a bias) and then passing this through the logistic function. The coefficients or weights associated with the features are learned during the training process, which involves minimizing the difference between the predicted probabilities and the actual class labels (often using a method like gradient descent). For binary classification it uses the logistic function to model the probability that each entry falls into one of the two categories based on the input variables (health indicators and demographic information). The outcome is a value between 0 and 1, interpreted as the probability of belonging to the positive class.

Logistic Regression stands out for its simplicity and high interpretability, making it an ideal choice for projects, where understanding the impact of each variable on the outcome is important. This model is not only straightforward to implement but also efficient in computation, which is particularly beneficial for initial analyses or projects with limited computational resources. Moreover, logistic regression outputs probabilities, providing a nuanced view of risk levels that can be extremely valuable in medical and health-related decision-making processes.

Neural Network

Neural networks consist of input, hidden, and output layers. Each layer is made up of nodes (or neurons) that apply a weighted sum to their inputs, followed by a non-linear activation function. The weights of these sums are adjusted during training. Training involves forward propagation of input through the network to generate a prediction, calculation of the error (difference between predicted and actual values), and backward propagation of this error through the network to adjust the weights. This process uses techniques like gradient descent and backpropagation. Deep learning involves neural networks with multiple hidden layers, which can model very complex relationships but require substantial data and computational power.

In the project, we can employ neural networks to predict diabetes from a dataset filled with health indicators and demographic information. I'll start by preprocessing the data, converting categorical variables to numerical ones, and normalizing the features to ensure they contribute equally to the model's learning. The network's architecture will have an input layer matching the number of features, several hidden layers to capture complex relationships, and an output layer with a single neuron using a sigmoid function for binary classification. Throughout training, the network will adjust weights to minimize prediction errors, aiming to accurately predict diabetes presence. This approach allows me to leverage the power of neural networks to handle the intricacies of my dataset effectively.

Neural Networks, with their deep learning variants, are at the forefront of modeling complex and non-linear relationships in large datasets. Their ability to learn intricate patterns through multiple layers of processing makes them exceptionally powerful for analyzing health indicators data, where interactions between variables can be complex. Despite their need for substantial computational resources and data, neural networks' capacity to model almost any relationship makes them indispensable for cutting-edge health data analysis projects.

Ensembles

An ensemble model in machine learning is a technique that combines the predictions from multiple models to improve the overall performance, accuracy, and robustness of predictive analytics. The fundamental premise behind ensemble methods is that a group of weak learners (models that perform slightly better than random guessing) can come together to form a strong learner (a model with significantly improved accuracy). Ensembles work by exploiting the diversity among the models in terms of their assumptions, heuristics, or learned patterns, thereby reducing the risk of making poor predictions due to model bias or variance.

In approaching my dataset, which aims to predict diabetes from a mix of health indicators and demographic details, I'm considering the use of ensemble models. These models may significantly enhance the predictive accuracy of my project by leveraging the collective strength of multiple algorithms. This strategy can potentially address the complexities and nuances within my data more effectively than any single model could. By combining different models, might reduce the risk of overfitting, which is crucial given the diverse nature of my dataset. Ensemble methods, such as Random Forests, Boosting, or Stacking, could offer a more robust and stable prediction mechanism by averaging out biases and variances across various models. This approach will likely increase the reliability of my predictions across different segments of the population. In essence, the use of ensemble models may provide a more nuanced understanding and prediction capability for diabetes occurrence, which could be pivotal in the context of health data analytics.

Performance Evaluation

Prior to performing the model performance evaluations, the predictor data was normalized to eliminate any scale effects. Min-max normalization was used to transform all variables onto a scale from 0 to 1. The min-max normalization method was chosen for its convenience since many predictor variables are in binary format. Since the data is highly imbalanced with only 14% of cases being diabetic, synthetic minority oversampling technique (SMOTE) was used to oversample the minority class. The data was then split into training and validation sets, with 30% of the data assigned to training and 70% assigned to validation.

k-Nearest Neighbors Model

An iterative analysis was performed to determine the best k-value to use in the model. The accuracy was calculated and plotted for various k-values as shown in Figure 11. Values ranging from 1 to 20 were analyzed, as this is the typical k-value range. The plot revealed that the optimal k-value to achieve the highest accuracy of 85.5% is 1.

A blue line on a white background

Description automatically generated

Figure 11: Model accuracy at various k-values.

The k-NN classifier was fitted on the training data and evaluated on the validation data. The confusion matrix and other performance metrics were generated and are shown in Figure 12. An ROC curve was generated and is shown in Figure 13.

|  |  |
| --- | --- |
| A close-up of a number  Description automatically generated  A number of numbers on a white background  Description automatically generated  Figure 12: Confusion matrix and classification report for k-NN model. | A screen shot of a graph  Description automatically generated  Figure 13: ROC for k-NN model. |

**Classification Trees**

An iterative analysis was performed to determine the optimal maximum depth value to use in the model. Accuracy was calculated and plotted for various values of maximum depth, as shown below in Figure 3. The highest accuracy was 87% at a maximum depth of 19.

A graph with a line

Description automatically generated

Figure 14: Model accuracy at various maximum depth values.

The confusion matrix, additional performance metrics, and an ROC were generated to evaluate this model’s performance, as shown below.

|  |  |
| --- | --- |
| A close-up of numbers  Description automatically generated  A number of numbers on a white background  Description automatically generated  Figure 15: Confusion matrix and classification report for classification tree. | Figure 16: ROC curve for classification tree. |

**Naïve Bayes Classifier**

The Naïve Bayes classifier was run on the data and the resulting confusion matrix and performance metrics are shown below. The performance metrics reveal that the Naïve Bayes classifier resulted in significantly lower accuracy (70%) and recall (62%) compared to the other models discussed.

|  |  |
| --- | --- |
| A close-up of numbers  Description automatically generated  A number of numbers on a white background  Description automatically generated  Figure 17: Confusion matrix and classification report for Naive Bayes classifier. | **A graph of a positive rate  Description automatically generated with medium confidence**  Figure 18: ROC curve for Naive Bayes classifier. |

**Logistic Regression**

Logistic regression was applied on the dataset, and the resulting confusion matrix and performance metrics are shown below. The logistic regression resulted in accuracy of 75% which is not at high as many of the other models that were examined.

|  |  |
| --- | --- |
| A close-up of numbers  Description automatically generated  A number of numbers on a white background  Description automatically generated  Figure 19: Confusion matrix and classification report for logistic regression. | **A graph of a positive rate  Description automatically generated with medium confidence**  Figure 20: ROC curve for logistic regression. |

**Neural Network**

A neural network model was employed to analyze the dataset and the outcomes of the classification model are displayed in the confusion matrix and performance metrics below. As evidenced by the displayed figures, the neural network achieved an accuracy of 76%, which is similar to the logistic regression model accuracy. However, the neural network model had a much higher AUC of 84% compared to the logistic regression AUC of 75%.

|  |  |
| --- | --- |
| A black text on a white background  Description automatically generated  A number of numbers on a white background  Description automatically generated  Figure 21: Confusion matrix and classification report for neural network. | A graph of a function  Description automatically generated  Figure 22: ROC curve for neural network. |

Random Forest

The random forest classifier was applied to the dataset and the resulting confusion matrix and accompanying performance metrics are presented below. The random forest achieved an accuracy of 89% which is the highest of all the models. Additionally, this model achieved higher precision, recall, f1-score, and AUC compared to all other models examined.

|  |  |
| --- | --- |
| A black and white text with black numbers  Description automatically generated with medium confidence    Figure 23: Confusion matrix and classification report for random forest. | A graph of a function  Description automatically generated  Figure 24: ROC curve for random forest. |

Project Results

Performance metrics used to evaluate the model performances are summarized in Figure 25 below, and include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). The results revealed that the random forest model offers the best performance parameters among all the models. This model correctly classified 89% of all observations, and thus has an error rate of 11%. Additionally, the model had high sensitivity, recall, and F1-score values of about 89%. Overall, these performance metrics suggest that the random forest model can identify diabetes cases with a high accuracy, while maintaining a balance between sensitivity and specificity. However, it is important to note that based on the error rate, about 11% of patients could be misdiagnosed by this model.

A close-up of numbers

Description automatically generated

Figure 25: Summary of performance metrics for each model.

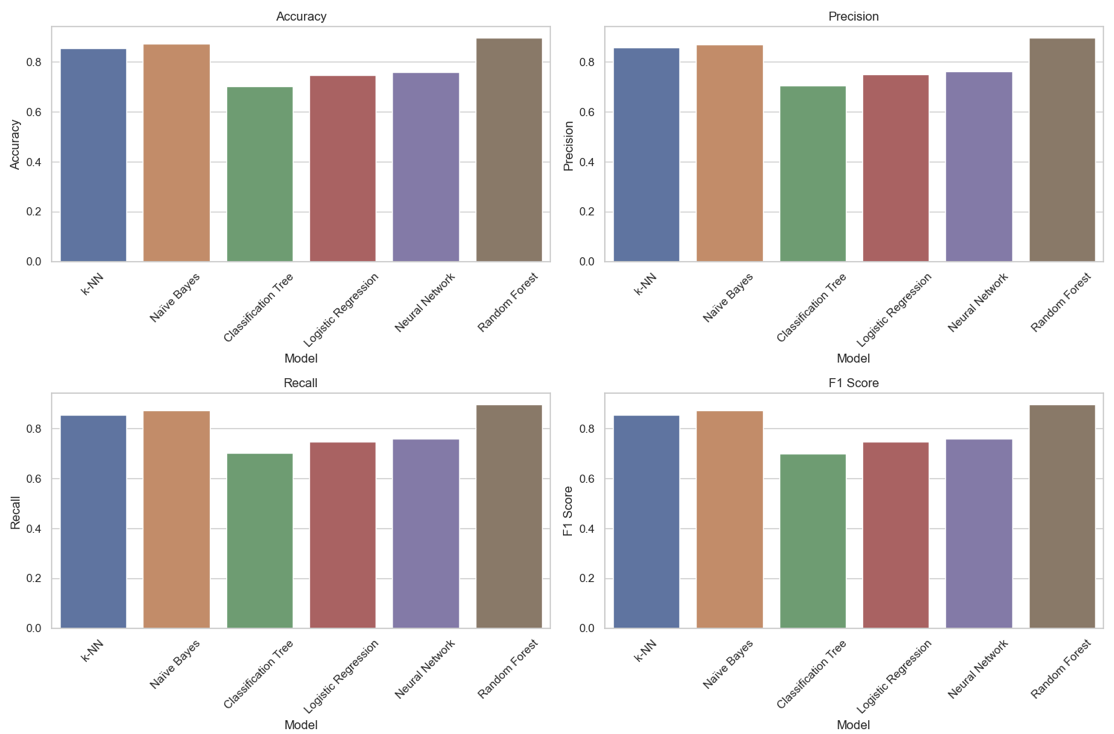


Figure 26: Bar chart comparison of performance metrics for each model.

Figure 26 above shows a comparison of the performance metrics for each model. The random forest model performed the best in terms of all performance metrics, while the classification tree performance the worst. The logistic regression and neural network models performed fairly. The k-NN and naïve Bayes models performed very well.

All of the models have an AUC (area under the ROC curve) value between 0.7 and 0.9. These values are higher than the baseline of 0.5, suggesting that all the models are performing better than a random classifier. The random forest model results in the highest AUC value of 0.89.

Impact of Project Outcomes

The applications of the random forest model described in this report are extensive. Having a model that can predict whether a patient is likely to have diabetes or not based on several simple predictors that can be answered in a survey is extremely useful. It provides a simple and cost-effective method for patients to assess their own risk level and determine their necessity for medical visits and diagnosis tests. It also encourages early-stage diagnosis and therefore prevention of severe complications that can developed with untreated diabetes.

Overall, the model developed for this project is a highly beneficial tool in assessing diabetes risk. However, the model does have an error rate of 11%, so should not be relied upon completely for diagnoses. Future work is required to further optimize this diabetes prediction model in order to reduce the error rate and increase reliability. This could be done by investigating additional models not covered in this project, or finding additional predictors that may be better indicators of diabetes risk.

References

|  |  |
| --- | --- |
| [1] | "Statistics About Diabetes," American Diabetes Association, [Online]. Available: https://diabetes.org/about-diabetes/statistics/about-diabetes#:~:text=Overall%20numbers,of%20the%20population%2C%20had%20diabetes.&text=Diagnosed%20and%20undiagnosed%3A%20Of%20the,and%208.7%20million%20were%20undiagnosed.. [Accessed 21 March 2024]. |
| [2] | "Healthy Weight, Nutrition, and Physical Activity," Centers for Disease Control and Prevention, 3 June 2022. [Online]. Available: https://www.cdc.gov/healthyweight/assessing/index.html#:~:text=If%20your%20BMI%20is%20less,falls%20within%20the%20obese%20range. [Accessed 21 March 2024]. |