Predicting the Presence of Diabetes Using Health Indicator Analysis

1st Jonathan Flanagan
School of Computing
National College of Ireland
Dublin, Ireland
x18143890@student.ncirl.ie

2nd Neil Fitzgerald
School of Computing
National College of Ireland
Dublin, Ireland
x18149693@student.ncirl.ie

***Abstract*—The purpose of this analysis is to predict the classification of diabetes in a person using a combination of 21 health and socioeconomic indicators from 70,692 surveys completed during the 2015 CDC Behavioral Risk Factor Surveillance System Survey. This is done by applying four data mining techniques: Logistic Regression, Random Forest, K-Nearest Neighbor and Naive-Bayes. Each models performance is measured and compared using their AUC-ROC score. A secondary objective to the analysis is to determine if there are any statistically significant indicators more prevalent in the diabetic diagnosis group than in the non-diabetic group based on features such as a persons sex, general health and physical activity. The study found that Logistic Regression performed the best of the four classification models with an AUC-ROC of 0.8202, Naive-Bayes scored lowest with an AUC-ROC of 0.7852, with Random Forest and K-Nearest Neighbors best models both achieving close to Linear Regression with an AUC-ROC of 0.8129. Along with these findings the study also found, using chi squared tests, that there is sufficient evidence to state there is an association between diabetes and high blood pressure, and high cholesterol. The study also indicates that there is no significant statistical difference between the number of males with diabetes compared to the number of females with diabetes in the surveyed group, but that there is a statistically significant difference between the number of people reporting low physical activity and suffering from diabetes versus those that do not suffer from diabetes.**

***Index Terms*—diabetes, data mining, health, predictive, R**

I. Introduction

During the normal digestion process, foods are broken down into sugars which are released into the bloodstream. This in-turn, signals the pancreas to release insulin. Insulin then aides in the breakdown of sugars in the bloodstream to be converted into energy the body can use.

Diabetes is generally characterized as a serious chronic disease and is determined by either the body not making enough insulin or not being able to use the insulin that it has made efficiently, resulting in an individual losing the ability to effectively regulate glucose levels in their blood.

A result of this can be a reduced quality of life and/or life expectancy. This condition is typically dissected into two groups, Type 1 or Type 2 diabetes. For the purposed of this study the Type has not been defined, only a diagnosis of Diabetes is used.

Chronic diabetes can lead to a range of additional health risks such as kidney disease, heart disease, vision loss and possible amputation of lower limbs. While there is no cure for diabetes, there are multiple lifestyle strategies the can be used to manage Diabetes like controlling weight, eating healthily, being active, and receiving medical treatments. These interventions can mitigate the harms of this disease in many patients.

Early diagnosis is an important factor in this and can lead to a more effective treatment. The aim of this analysis is to examine multiple data mining and classification algorithms in an attempt to classify whether a person is Diabetic or not.

Using 21 common health related questions as indicators, Logistic Regression, Random Forest, K-Nearest Neighbor (KNN) and Naive-Bayes algorithms will be used in the classification of diabetes being present in a person. By making predictive models for diabetes risk, it has the potential to help health officials in early diagnosis of diabetes and aide in the effective early treatment of the disease.

The analysis will use a readily available data set from Kaggle [1] that contains 70,692 observations and 22 questions from the 2015 Centers for Disease Control (CDC) Behavioral Risk Factor Surveillance System (BRFSS). The data set has an even split between Diabetic and non-Diabetic people.

The original survey had 253,680 respondents from across the United States and contained 330 questions. The sample taken from the original survey was not decided by this study and had been prepared previously. Responses to questions that contained answers such as "*don't know*" and "*refused to answer*" have already been removed from the data which helps in the preprocessing stages of the analysis.

As well as applying classification techniques, this study will also explore the health indications of the survey questions using statistical techniques such as chi squared tests, examining the relationship between Diabetes diagnosis and factors such as a persons sex, blood pressure, physical activity, presence of a pre diagnosed heart condition or if they have ever suffered from a stroke.

[1] https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

II. Literature Review

A. Knowledge Discovery and Data Mining: Towards a Unifying Framework [1]

This paper outlines process of Knowledge Discovery in Databases (KDD) and is a widely accepted methodology for use in data analytics, it is the reference for the methodology used throughout this analysis.

B. Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques [2]

This paper addresses many of the same prediction techniques that are to be applied in this study, Naive-Bayes, Logistic Regression and Random Forest. It found that Random Forest was the most successful model. The study contains 520 observations using a questionnaire from patients in Bangladesh. Although the data in this current study is based on different questionnaire questions and from people of a different cultural background the factor representation of the questions is helpful in understanding the problem domain addressed. It is the intention of this study to add to this work in attempt to use different health indicator questions to come to similar results.

C. Cigarette Smoking and Diabetes [3]

The conclusions of this paper are that smoking increases the risk of diabetes. As one of the survey questions this study will be investigating is if a participant smokes these results will be examined to see if smoking is an impact-full measure in predicting diabetes. The paper draws conclusions from multiple sources that the increased risk of diabetes in both men and women increases by 50% with smoking, so it is expected in the study being performed that smokers will show as a statistically significant proportion of the diabetic cohort.

D. Physical inactivity and obesity: links with insulin resistance and type 2 diabetes mellitus [4]

In this paper, the links between physical inactivity and diabetes are explored with a conclusion that physical activity reduces the risk of diabetes. With an age adjusted risk decrease of 6%. Results where replicated across the citations showing a progressive reduction in diabetes risk with physical activity. This is relevant to the current study as physical activity is one of the factors addressed and as such it is expected to have an impact on predictions made. It also shows a reduction in the risk of hypertension which is another factor present in the BRFSS survey that is being analysed in this study.

E. Diabetes disease prediction using data mining [5]

This paper deals with predicting Diabetes through KNN and Naive-Bayes classification, although the paper doesn't go into detail on results or the complete architecture of the models used, there are a some useful insights in sample sizes and transformation of frequency tables, it is the intent of this study layer on top of these insights when creating the classification models.

F. Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm [6]

KNN algorithm to predict Diabetes is the main focus in this study, it uses K values of 3 and 5 with error rates of 30% for K3 and 35% for K5. The study to be completed and the study in the paper share similar attributes in the data set but the data set this study performed KNN on was of 100 observations and 11 features which is considerably less than to be used in the proposed study. Hopefully this will help in achieving similar or better error rates.

G. Predicting diabetes mellitus with machine learning techniques [7]

Random Forrest and Neural Networks were used as this papers prediction methods, although the proposed study does not intend to use Neural Networks the implementation of its random forest classification was examined, the best model prediction accuracy for this model was 80.84% (taken as an average of the five best models) using fourteen data features and 68,994 observations, this paper was of particular interest as the data sets being used are similar. 5 fold cross validation was used. Principal component analysis was used to reduce dimensions in this study, this is one advantage the study will have over the proposed study as principal component analysis will be done somewhat manually through p-value selection in the models and manually reducing dimensions with high numbers of same question answers by participants.

H. Prediction of Diabetes using Classification Algorithms [8]

The classification models used in this paper are Decision Trees, SVM and Naive-Bayes. The Naive-Bayes implementation is of particular interest from this study as it is one of the models chosen for the proposed study. This study found that Naive-Bayes performed the best of all three in terms of accuracy with an accuracy of 76.30%, results are verified using ROC curves with an AU-ROC of 0.819. Again the models in this study used a relatively small sample set of 768 observations with 8 data features, it is the hope that the larger data set to train the model in the proposed study will be beneficial in increasing that accuracy percentage.

I. Analysis and prediction of diabetes using machine learning [9]

This paper takes a similar approach to the proposed study by investigating the performance of multiple classification algorithms to see which performs best. KNN, Naive-Bayes and Decision Trees are implemented in the study. This study used bootstrapping in its methodology which may give it an advantage over the proposed study where bootstrapping will not be used. Decision tree performed best in the initial stages as well as after bootstrapping with accuracy rates or 78.43% before and 94.4%t after. This may highlight a limitation in the proposed study where bootstrapping will not be applied so a target of a similar pre-bootstrapping accuracy rate is aimed for.

J. Prediction of type 2 diabetes using machine learning classification methods [10]

The data set in this study is comparatively a lot less than in the proposed study with 952 observations recorded. There are similar questions in the study though, physical activity and high blood pressure being two of the common questions. With 12 questions total. The study uses similar methods from the proposed study mainly logistic regression, Naive-Bayes, Random Forest and KNN. Scores are also recorded using AUC similar to the proposed study. Logistic Regression scored 0.765, KNN 0.815, Naive-Bayes 0.760 and Random Forest 1.00. The Random Forest result is questionable due to the perfect scoring but in relation to the proposed study is the intent to use the same methods on a larger data set with slightly different survey questions to achieve similar or greater AUC results.

III. Methodology

A KDD approach is being used for this analysis and it will follow the five main steps of the process: selection, pre-processing, transformation, data mining and interpretation. KDD is a widely accepted methodology in data analysis and is a template used for the extraction of useful structured patterns from data. The approach will be applied iteratively in this analysis between different steps of the methodology especially in the transformation and data mining steps.

Along with the five main steps in the process there is arguably one additional step of the process and that is domain knowledge. Domain knowledge is needed in being able to understand and interpret the data and being analysed in the context of the specific problem domain addressed.

The domain knowledge for this analysis has been acquired by the researchers studying the relevant literature on past analysis, the BRFSS study code book and relevant information on Diabetes from the world health organisation and available medical journals.

The purpose of using classification is to be able to possibly predict the presence of Diabetes in persons outside of the initial study and develop a model that can be used in parallel with other diagnosis methods currently used by health professionals. The success of such a model would be an invaluable addition to the already used methods of diagnosis by identifying people where further medical diagnosis should possibly take place.

To determine if there are any statistically significant correlations or frequency of health indicators in a specific group, Pearson correlation coefficient is to be used on numeric variables and chi squared tests of independence as well as chi squared tests of goodness of fit are to be implemented on factor variables using their p-value as the determining metric.

The null hypothesis in the case of using chi squared independence tests will be that there is no association between the tested variable and diabetes diagnosis. The null hypothesis in the case of using the chi squared goodness of fit tests will be that diabetic and non-diabetic people in a chosen group follows an approximate distribution of 50:50.

Data Selection

Preprocessing and Transformation

Preprocessing of the data will include checking for missing values and imputing where necessary or removing rows with null values altogether, this will be decided on by the type of data in the column missing data, if it can be imputed using mean or median values or if the the column has more then 25% of its data missing it could be removed [1]. The data in each feature may be returned to the original string value representation in the survey instead of the numeric representation in the Kaggle data set for human readability. For any sampling that is to be done a stratified approach will be taken to keep normal distribution of features within the data, this will be done using Kolmogorov-Smirnov tests on numeric values and for factors chi squared tests will be performed, making sure a p-value of above 0.05 and as close to 1 as possible is used for each feature, making sure that the distributions of the sample set matches the distributions of the full data set. For the classification algorithms the data set will need to be transformed in different ways, with logistic regression a model will be created using all features and then based on the significance of each column used a more refined model can be created. For KNN the original numeric representation can be used and the values scaled for use with the algorithm, then looping through a range of K variables will be used, and the best K selected for the final model. For Random Forest a random selection of features will be used and tested until all combinations are tested and the best combination of features will then be used.

Data Mining

Using the KDD methodology, this analysis will be applying classification data mining techniques, namely; Logistic Regression, Random Forest, KNN and Naive-Bayes, and assessing which model works best through the use of AUC-ROC scoring. These algorithms have been chosen as the problem domain of the study is classification and from the research literature these algorithms perform best in this field. [8] [2].

Logistic Regression

Logistic Regression is a classification algorithm and is used to predict the outcome of a binary variable, in this study it will be used to predict the binary option of Diabetes or no Diabetes. As it is primarily used as a binary classifier it would be well suited to the problem domain of this study. Each column of the data set will have a coefficient that is learned while training the model and used to predict the outcome. Best model selection will be done by first using all features in the data set and then using the alpha of 0.05 discussed previous to determine a features impact of on the model, reducing the number of features needed while maximising the performance of the model.

Random Forest

Random Forest is an extension of decision trees where the output of multiple decision trees are used to reach a single predicted output. For this analysis the predicted output again will be a classification output based on a certain probability. A base of 500 trees is to be used and refined by looping through random selections of columns (default will be 4), this is done to reduce correlation between features in the data set and to determine the best combination of columns and number of decision trees to use in maximising the accuracy of predictions. This classifier, from previous research done, has shown a high success rate in this problem domain.

Naive-Bayes

Naive-Bayes is slightly different from the other algorithms to be used in the analysis as it is a collection of algorithms based on Bayes Theorem and is used in data mining for classification problems which suit the problem domain of the proposed study. Naive-Bayes works under an assumption that the features in the data set are independent and share an equal contribution to the outcome. The independence assumption may not be true of features in the data set but still the performance of the model in previous studies shows that it is a positive model to use in this problem domain.

KNN

K-Nearest Neighbor is a type of supervised machine learning algorithm and is suitable for classification problems. It is a lazy learning algorithm which means it does not have a training phase but instead uses the complete data for training while classifying. KNN is also categorized as a non-parametric algorithm, meaning it makes no assumptions about the underlying data. Using Euclidean distance the classification is conducted by calculating the nearest K neighbors to the observation being classified. It has been used successfully in previous studies from the literary review and has displayed proficient accuracy scores.

These algorithms where chosen as they are established methods for classification and show high success rates in the problem domain of the analysis to be carried out.

Interpretation and Evaluation

Even though this analysis is dealing with a medical condition, the results will not be making any recommendations on treatments or medication and therefore when evaluating p-values, an industry standard alpha of 0.05 will be used. The researchers feel this is appropriate as the nature of the analysis is to find correlations and health indication factors that may assist in recommending further diagnosis but it is not diagnosis or treatment tools in itself. There are two main approaches that will require interpretation and evaluation throughout the proposed analysis.

Classification Models

Models will be assessed using AUC-ROC as the scoring criteria as it provides an aggregated measure of performance across all of the classification thresholds, is widely used in academia and provides a probability that the assessed model will rank a random positive diagnosis more highly than a random negative diagnosis. AUC-ROC is also classification threshold invariant, meaning the quality of the classifications is measured irrespective of the threshold chosen for classification.

A caveat to scoring with AUC-ROC is that calibrating probability outputs won't be possible for this reason specificity, sensitivity and miss-classification error rate will be used to tune models but the overall AUC-ROC score will be used when comparing models overall performance.

This evaluation method has been chosen as the researchers feel it is a favourable scenario to possibly falsely identify a negative case as positive rather than falsely identify a positive case as negative.

Chi Squared Tests

Two types of chi squared tests are to be be performed, testing for independence and testing goodness of fit to a predefined expect distribution. For these tests the alpha of 0.05 will be used by default when determining the statistical significance of a result.

As the chi squared tests are highly sensitive to the number of observations and it's results can vary dramatically as sample size grows, the tests are to be performed on the full data set first and the a sample sizes of 1000, 500 and 300 observations will be used from the original as per recommendations from a study on the affect of chi squared analysis on large data sets [12]. The samples are created using a stratified method and keeping the same probability distribution across all features.

For goodness of fit chi squared tests the theoretical distribution being used is 50:50, where the expected results will be that the variable being measured has a an equal representation in both the diabetic and non diabetic groups.

For the independence testing the variables will be evaluated under the null hypothesis that there is no statistical association between it and the presence of diabetes in a person.

The chi squared tests are to be used in this analysis as they provide a statistical test for the factor and logical type features in the data set and can provide extra insight into the prevalence of certain health indicators in people suffering from diabetes.

IV. Implementation

Analysis on the data is carried out using the programming language R and the R studio IDE. With some data manipulation carried out in Python using the Pandas package within the R code.

A. Data Selection and Preprocessing

The data sourced from Kaggle[2] is used in the analysis and prediction models. The data set contains 70,692 rows, evenly split between Diabetic and non-diabetic (35,346/35,346) with 21 features and 1 predictor variable (Diabetes). It is a subset of data taken from the BRFSS survey [11] carried out in 2015 where 253,680 United States residents responded to 330 questions. 21 questions have then been taken that provide health and socioeconomic indicators along with the Diabetes diagnosis making up the data set downloaded from Kaggle.

The data set is loaded into R studio and an initial view of the data shows that it all columns are read in as numeric values. These need to be altered to their correct data type before analysis can be begin. From domain knowledge gathered from the data set description online and the BRFSS survey handbook [11] using Python dictionary mapping the factors are returned to their original state making the data set more human readable while changing the data types.

Data types are a mixture of factors and numeric. Some of the binary factors are logical True/False answers and some are Yes/No answers. [Table I]

TABLE I
DATA TYPES

Column	Data Type
Diabetesbinary	BOOLEAN (TRUE/FALSE)
HighBP	BOOLEAN (TRUE/FALSE)
HighChol	BOOLEAN (TRUE/FALSE)
CholCheck	BOOLEAN (TRUE/FALSE)
BMI	Integer
Smoker	FACTOR (Yes/No)
Stroke	FACTOR (Yes/No)
HeartDiseaseorAttack	FACTOR (Yes/No)
PhysActivity	BOOLEAN (TRUE/FALSE)
Fruits	BOOLEAN (TRUE/FALSE)
Veggies	BOOLEAN (TRUE/FALSE)
HvyAlcoholConsump	BOOLEAN (TRUE/FALSE) k
AnyHealthcare	FACTOR (Yes/No)
NoDocbcCost	FACTOR (Yes/No)
GenHlth	FACTOR (1:Excellent, 2:Very Good, 3:Good, 4:Fair, 5:Poor)
MentHlth	INT (0-30)
PhysHlth	INT (0-30)
DiffWalk	FACTOR (Yes/No)
Sex	FACTOR (1:Male, 2:Female)
Age	FACTOR Thirteen-levels
Education	FACTOR Six Levels
Income	FACTOR Eight Levels

The data is then checked for any missing values and possible imputation needed. Using R data summary and the package Amelia, a data map is created showing there are no missing values and there is no need for any imputation.

B. Data Exploration

The numeric values in the data are represented by days for *MentHlth* or *PhysHlth* and an integer range for *BMI*. The first two may be turned into factors but the features are explored in their numeric representation first. As turning them into factors would lead to having two features with 30 levels of factors and may impact the models if all factors are not represented throughout training and test sets. Normality testing using Kolmogorov-Smirnov tests and QQ plots on all three numeric features show that the values are not normally distributed.

The frequency of each factor feature is checked along with some being tested using chi squared independence tests and chi squared goodness of fit tests (according to theorised 50:50 distribution).

For *HighBP* The data shows that from the survey there is a higher number of people with high blood pressure, with 56.35% of people reported having high blood pressure versus 43.65% of people not having high blood pressure. 75.27% of people from the survey who are diagnosed with Diabetes also have high blood pressure, whereas only 37.42% of the people that are diagnosed as not having Diabetes have high blood pressure.

For *HighChol* there is an approximate even split between people who have reported ever having high cholesterol versus people who have not with 52.57% True and 47.43% False. From the diabetic group there is 67.01% that have reported high cholesterol versus only 32.99% not having high cholesterol. This appears to be a significant difference and could be a heavy determining factor in diagnosing diabetes.

29.70% of the people in the survey report not having done any physical activity outside their normal job in the previous 30 days [*PhysActivity*]. From the Diabetes cohort as a whole 63.05% had completed physical activity while 36.95% hadn't but for the non-diabetic people 77.55% had and only 22.45% hadn't. Although this measure could be biased dependent on a person's perspective of physical activity it does indicate that Diabetic people have a higher amount of inactivity than non-diabetic people. From the group that have a diagnosis of diabetes 37.12% reported having difficulty walking or climbing stairs [*DiffWalk*] whereas only 13.42% of the non-diabetic people had difficulty. Again, there is no initial indication whether this is a result of the diagnosis or not but the presence of a higher percentage with diabetes that do have difficulty could result in an indicator.

54.30% of the people in the survey are female and 45.70% of the people are male [*Sex*] . From the group of Diabetic people 52.09% are female and 47.91% are male. The majority of people in this study fall between the ages of 55 and 74. (53.21%). Age groups with the higher percentage of diabetes prevalent are 70 to 74 (63.91%), 75 to 79 (63.09%), 65 to 69 (60.41%) and 60 to 64 (56.70%), this could be due to the number of people in each age category in the study.

C. Modeling

LOGISTIC REGRESSION

A training and test set of the data are created using an 80/20 split. For the initial model all the features are used. It is then refined in subsequent models using the features that have the highest impact on the model. VIF is checked for

[2]https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

multicollinearity and none of the features return a score of more than 1 so we can assume that multicollinearity is not an issue in our model.

Variable importance is checked with 9 of the features showing significant importance: *HighBP*, *HighChol*, *BMI*, *Smoker*, *Stroke*, *HeartDiseaseorAttack*, *GenHlth*, *Sex* and *Age*. [Table II]

A second model is then created using these features as the prediction variables and a third model is created using a further mixture of features in attempt to improve any performance of the model. The features for the third model are chosen from the a mixture of the VIF scores produced from the first model as well as features supported as being important to Diabetes diagnosis from literature reviews and the chi squared tests carried out in data exploration. The final combination of features tested in the third model are: *HighBP*, *HighChol*, *HeartDiseaseorAttack*, *Veggies*, *GenHlth*, *MentHlth*, *DiffWalk*, *Sex*, *Age*, *Income*, *BMI*, *Education*, *Income*

TABLE II
VIF SCORES

Feature	Score
HighBPTRUE	31.7810226
HighCholTRUE	27.0136022
BMI	41.2751598
SmokerYes	3.2872392
StrokeYes	3.9328665
HeartDiseaseorAttackYes	9.3917943
GenHlthFair	38.9111945
SexMale	12.9373274
Age70 to 74	14.2556660

RANDOM FOREST

The random forest is run in a default state with no parameters to begin with, then an attempt to tune the model is performed while testing for sensitivity, specificity, miss-classification error and AUC-ROC.

The same training and test sets created for the logistic regression model are used in this model.

The random forest model default used a classification type with 500 trees and 4 random variables at each point. The best models when graphed show that after approx 100 trees the error rate only shows minor decline but that the best number of trees for the model is the 500 default. [*fig. 1*]

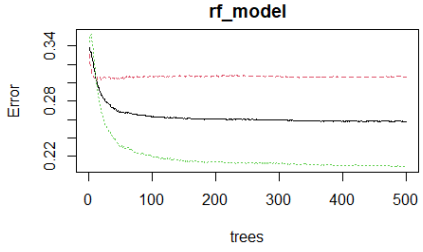

Fig. 1. Random Forest tree performance

Feature importance is tested and shows that BMI, AGE, GenHlth, Income and HighBP have the highest weighting on the model. [*fig. 2*]

Fig. 2. Random Forest Feature Importance

The model is then tuned using the five most important features with a tree size of 500, 4 random variables used at each point with a step factor of 1.5 and an improve rate of 0.01.

Out of the Bag (OOB) error rate is then plotted and shows that the best OOB error rate achieved is when using 2 random features which is being used. [*fig. 3*]

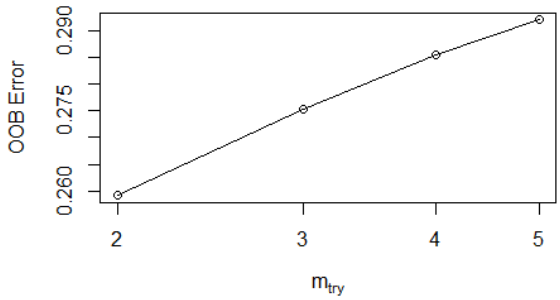

Fig. 3. Random Forest best OOB

NAIVE-BAYES

Naive Bayes model is used as the third model to be test for predicting diabetes using the health indicators from the data set. A stratified sample of 5000 observations from the data set is used due to processing restrictions and is again split into a 80/20 train and test sets.

A summary of the test set is then checked to make sure all factors are accounted for in the sample. An X and Y variable are created by splitting the data set on the dependent variable against the other independent variables.

Train control parameters are set for the model using 10 fold cross validation. Search grid parameters are set using kernel set to (True, False), and fl of 0:5 and a sequence adjustment range from 0 to 5 in 1 step increments.

Once training is complete the top 5 models produced are selected and plotted. [*fig.4*]

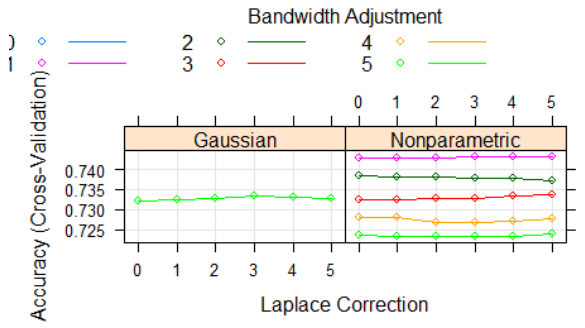

Fig. 4. Naive-Bayes Laplace Graph

KNN

For K-Nearest Neighbor the original numeric dataset and the same features as the previous models are used in the implementation.

The original datasets numeric values are needed for the algorithm and values are scaled before the model is applied.

Even though KNN uses the complete number of observations a training and test are created from a stratified sample. The sample size chosen is 20,000. this figure is chosen by the researchers again due to processing restrictions.

Training control parameters are set with 3 repeats, 10 fold cross validation and class probability set to True. The scale process variables are set the *center* and *scale* with the measure metric set to ROC as this will be the overall metric used to evaluate the performance of the models.

The best number for K is tested in the training and a graph plotted with results. [*fig. 5*]

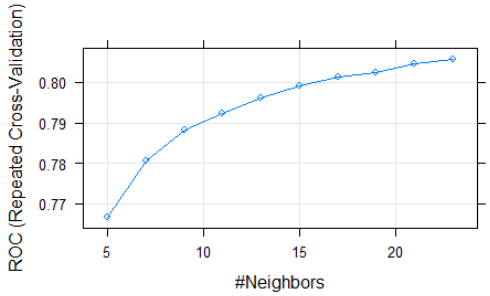

Fig. 5. KNN Best N

V. Results and Findings

Logistic Regression

Using McFaddens method, the initial models pseudo R^2 value is 0.2619746 with sensitivity of 0.767148, specificity of 0.718471, a total miss-classification error rate of 0.255200 with an AUC-ROC score of 0.8202 [*fig. 6*].

The second model using the features with highest impact returns a pseudo R^2 value of 0.2578206, sensitivity of 0.769252, specificity of 0.715499, a total miss-classification error rate of 0.256800 with an AUC-ROC score of 0.8184.

The third model created using the further mixture of features produces a pseudo R^2 value of 0.2615656, sensitivity of 0.768832, specificity of 0.717339, a total miss-classification error rate of 0.255600 with an AUC-ROC score of 0.8184. It has a slightly higher sensitivity score than then second model, and slightly lower specificity and error rate scores. [Table III]

TABLE III
Model Results

Model Number	Spec	Sens	Err	AUC
1	0.718471	0.767148	0.255200	0.8202
2	0.715499	0.769252	0.256800	0.8184
3	0.717339	0.768832	0.255600	0.8184

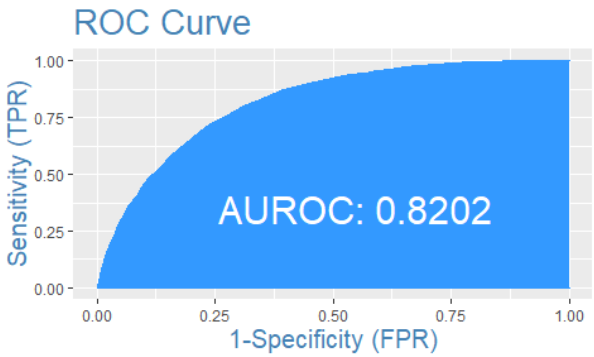

Fig. 6. Best Logistic Regression AUC-ROC

Random Forest

After tuning the Random Forest model inline with the OOB plot using 2 random features and the 500 trees, the best model returned scores of: sensitivity of 0.792538, specificity of 0.685775, and a total miss-classification error rate of 0.259100 with an AUC-ROC score of 0.8129.[Table IV] [*fig.7*]

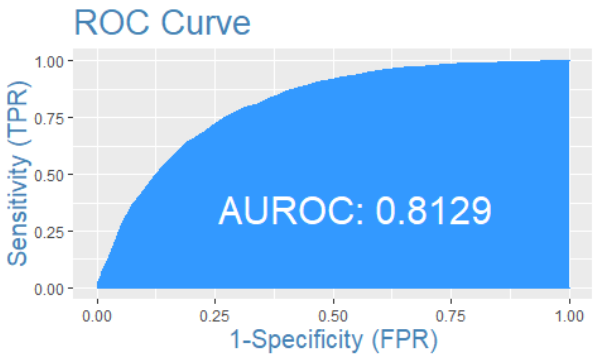

Fig. 7. Random Forest AUC-ROC

TABLE IV
MODEL RESULTS

Model	Spec	Sens	Err	AUC
RF	0.685775	0.792538	0.259100	0.8129

NAIVE-BAYES

The best training model, using 10 fold cross validation, returns an average accuracy of 74.34%. The test set returns a balanced accuracy of 72.27%, sensitivity of 0.698413, specificity of 0.746964, and a total miss-classification error rate of 0.277600 with an AUC-ROC score of 0.7852.[Table V] [*fig. 8*]

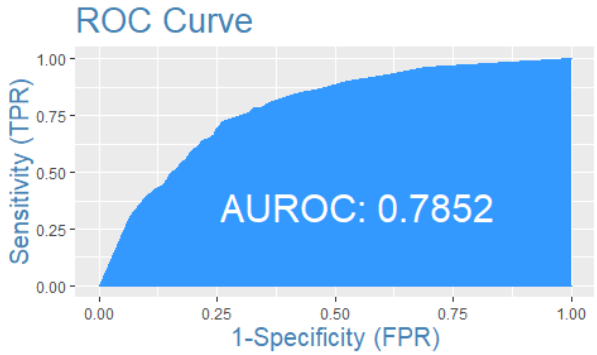

Fig. 8. Naive-Bayes AUC-ROC

TABLE V
MODEL RESULTS

Model	Spec	Sens	Err	AUC
Naive-Bayes	0.746964	0.698413	0.277600	0.7852

KNN

Using the best N of 23 for the KNN algorithm on the unseen test set, the model is applied and returns scores of: sensitivity of 0.781951, specificity of 0.698799, and a total miss-classification error rate of 0.259500 with an AUC-ROC score of 0.8129.[Table VI] [*fig. 9*]

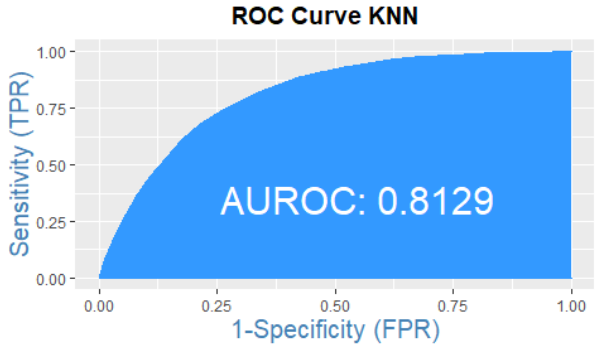

Fig. 9. KNN AUC-ROC

TABLE VI
MODEL RESULTS

Model	Spec	Sens	Err	AUC
KNN	0.698799	0.781951	0.259500	0.8129

TABLE VII
OVERALL MODEL RESULTS

Model	Spec	Sens	Err	AUC
Logistic Regression	0.718471	0.767148	0.255200	0.8202
Random Forest	0.685775	0.792538	0.259100	0.8129
KNN	0.698799	0.781951	0.259500	0.8129
Naive-Bayes	0.746964	0.698413	0.277600	0.7852

Chi squared tests of independence on high blood pressure returns a p-value of less than 2.708e-09 therefore the null hypothesis is rejected meaning there is sufficient evidence to say that there is an association between high blood pressure and Diabetes.

Similar results are found when testing high cholesterol (p-value less than 6.371e-08), Stroke (p=0.04521) and Heart Disease (p less than 3.487e-05) is calculated, rejecting the null hypothesis in each case, giving sufficient evidence to say that there is an association between these factors and Diabetes. Confirming findings in previous studies. [13] [14]. For Physical activity and difficulty walking p values of 0.0009368 and 0.0002079 are calculated, again rejecting the null hypothesis and confirming results from a previous study. [15]. Contrary to this study [16]. For males and females the study found there was no increased risk across the genders and that from all the participants in the study, for the number of Males, Diabetes does approx. follow a hypothesized distribution of 50:50 (p=0.6284) and for Females the diagnosis follows a hypothesized approx distribution of 50:50 (p=0.6833).

VI. CONCLUSIONS & FUTURE WORK

The analysis carried out, in the opinion of the researches, shows that there is potential in using a predictive model to assess the possible presence of Diabetes in a person. Although the results did not score higher than 0.83 on the AUC-ROC metric, with more tuning of the models they could be improved. The Logistic Regression model performed the best of the models tested with KNN and Random forest performing closely behind. The analysis also found that there is a statistical association between diabetes and high blood pressure, high cholesterol and low physical activity, and also indicates that there is no significant statistical difference between the number of males with diabetes compared to the number of females with diabetes.

Future work in this field could be improved by covering a wider initial question set, possibly made up of a mixture of the health and socioeconomic questions in this analysis's survey as well as some of the numeric and body metric questions present in the surveys in the literature review carried out. With more time and an expanded data set this analysis could be improved with further tuning of the Random Forest and Naive-Bayes models.

REFERENCES

[1] Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P., 1996, August. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In KDD (Vol. 96, pp. 82-88).

[2] Islam, M.M., Ferdousi, R., Rahman, S. and Bushra, H.Y., 2020. Likelihood prediction of diabetes at early stage using data mining techniques. In Computer Vision and Machine Intelligence in Medical Image Analysis (pp. 113-125). Springer, Singapore.

[3] Eliasson, B., 2003. Cigarette smoking and diabetes. Progress in cardiovascular diseases, 45(5), pp.405-413.

[4] Venables, M.C. and Jeukendrup, A.E., 2009. Physical inactivity and obesity: links with insulin resistance and type 2 diabetes mellitus. Diabetes/metabolism research and reviews, 25(S1), pp.S18-S23.

[5] Shetty, D., Rit, K., Shaikh, S. and Patil, N., 2017, March. Diabetes disease prediction using data mining. In 2017 international conference on innovations in information, embedded and communication systems (ICIIECS) (pp. 1-5). IEEE.

[6] Saxena, K., Khan, Z. and Singh, S., 2014. Diagnosis of diabetes mellitus using k nearest neighbor algorithm. International Journal of Computer Science Trends and Technology (IJCST), 2(4), pp.36-43.

[7] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y. and Tang, H., 2018. Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, p.515.

[8] Sisodia, D. and Sisodia, D.S., 2018. Prediction of diabetes using classification algorithms. Procedia computer science, 132, pp.1578-1585.

[9] Saru, S. and Subashree, S., 2019. Analysis and prediction of diabetes using machine learning. International Journal of Emerging Technology and Innovative Engineering, 5(4).

[10] Tigga, N.P. and Garg, S., 2020. Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 167, pp.706-716.

[11] cdc.gov. 2015. Behavioral Risk Factor Surveillance System. https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

[12] Shi, D., DiStefano, C., McDaniel, H.L. and Jiang, Z., 2018. Examining chi-square test statistics under conditions of large model size and ordinal data. Structural Equation Modeling: A Multidisciplinary Journal, 25(6), pp.924-945.

[13] Barrett-Connor, E. and Khaw, K.T., 1988. Diabetes mellitus: an independent risk factor for stroke?. American journal of epidemiology, 128(1), pp.116-123.

[14] Wilson, P.W., 1998. Diabetes mellitus and coronary heart disease. American Journal of Kidney Diseases, 32(5), pp.S89-S100.

[15] Venables, M.C. and Jeukendrup, A.E., 2009. Physical inactivity and obesity: links with insulin resistance and type 2 diabetes mellitus. Diabetes/metabolism research and reviews, 25(S1), pp.S18-S23.

[16] Gucciardi, E., Wang, S.C.T., DeMelo, M., Amaral, L. and Stewart, D.E., 2008. Characteristics of men and women with diabetes: observations during patients' initial visit to a diabetes education centre. Canadian Family Physician, 54(2), pp.219-227.