National College of Ireland

BSc. Honours in Data Science

BSHDS4

Academic Year: 2022/2023

Benjamin Kelly

X19370681

X19370681@student.ncirl.ie

An analysis of the causes and prevention of diabetes

Data Analysis

4th year Project Report

Table of Contents

Executive Summary

The main aim of the project is to explore the factors that influence the onset of diabetes and examine how diabetes can be controlled by analysing and understanding risk factors associated with the disease. The project aims to identify the most suitable model or classifier that could be used to predict the likelihood of someone becoming diabetic based on various data points. The project datasets were collected in CSV format as this was most suitable for use on a variety of platforms. The project report outlines the approach to data cleansing, analysis in classification models and exploration of what models worked best.

Following data preparation, I carried out statistical analysis and exploratory analysis using various methods and platforms including R Studio. I then applied statistical methods such as Students T-Test and the Chi-Square test as they were appropriate for classification problems. These tests were able to explain important points about the data and features, and the initial results of these tests are discussed in the results section of this report.

A number of classification models were implemented in this project including Decision trees, Random Forest and Naïve Bayes in addition to Neural Network modelling. The results summarise how these performed in the context of the aims of the project. The results are further summarised in the conclusions section.

I identified previous research carried out in the area of diabetes data analysis in the past including academic sources and medical journals widely available on the web. They highlight the importance of this research to healthcare professionals and how it can help us understand the increasing rates of diabetes. When I analysed the results and models against the project objective (i.e. to predict the incidence of diabetes), the three top performing models were:

- Neural Networks Smote Hyper Model,
- Random Forest Down sampled model
- Down sampled Decision Tree model

In terms of overall model performance, the **Neural Network Smote model** performed the best with an accuracy > 90% beating other models' accuracy of between 70-75%. The project also illustrates that the use of AI over normal machine learning models was superior to normal machine learning models in the context of the project.

1.0 Introduction

1.1. Background

I have an interest in the healthcare industry, and wanted to do a study on a global disease that has been prevalent and increasing in recent years. In particular, diabetes was of interest because it is something that is prevalent in my extended family. It is important for healthcare providers to understand why diabetes is a growing problem and understand the factors which influence the disease causes so that society can find ways of coping with the increase. There have been many studies carried out in the past on diabetes, influencing factors and predicting trends based on data science and academic research. My project

objective was to be identify a classification model that would be best able to predict the likelihood of someone being diabetic.

1.2. Aims

This project aims included:

1. Exploring appropriate classification models to determine what model is best for predicting the output of someone being diabetic.
2. Exploring what are the factors that would most influence this outcome.
3. To use suitable statistical analysis to help us understand the importance of various lifestyle and other related data features and their correlations.

The research question for this project is: **"Given relevant data and the application of classifiers on that data, what is the likelihood of someone being diagnosed as a diabetic?".**

1.3. Technology

The following technology has been used in the project:

The coding platforms I have used in the project are: **Power BI, Jupyter Notebooks and RStudio.**

PowerBI was used initially to carry out some visualisation analysis of my chosen datasets. This helped me understand specific statistical details such as correlations among features and identify patterns amongst features.

Jupyter notebook was used for Python programming and the implementation of the Neural Network models.

RStudio was used to carry out statistical analysis of the datasets during the earlier parts of the project at the data exploration stage. It was used to implement four different types of classification models. I used this tool a lot during my internship in third year, and this allowed me develop a good understanding of model implementation in R. Neural Network models were not carried out in R, as python had more resources available.

There are other web based tools such as **Google Collab** and other software recommended by the computing department which I explored, in case performance issues arose running the neural network models. These options however were not required in the end.

1.4. Structure of this report

The introduction section covers the project background, the technologies used during the project, the aims of the project and reasons why I chose the project.

The **state-of-the-art section** details previous studies carried out in this area. I researched a variety of academic sources and have highlighted relevant aspects of related previous research and how they compared to my project.

The data section provides an overview and description of the datasets and the type of exploratory analysis to be applied to them.

Methodology and Implementation section outlines how the project is implemented using the CRISP-DM methodology. It summarises all the steps in the process as that are carried out in the project.

The analysis section details the activities carried out in Section 4 (methodology and implementation).

In **the results section** I have evaluated and compared the results of the various statistical tests and classification models. The results from the models are discussed and analysed based on effectiveness, accuracy and performance.

In the **conclusion section,** I have outlined the objectives achieved during the project, what I would have done differently if more time was allocated. I also comment on whether the study was effective and what could be learned from using risk factors to understand the causes of type 2 diabetes.

Research Work Section: I have included a comprehensive list of academic sources referenced in this paper which had a background of similar work.

2.0 State of the Art

The following two related studies were identified and researched as part of the project.

In the study, **Comparative Analysis of Machine Learning Models for Diabetes Prediction [5]** the authors carried out an analysis of diabetes data using **six different models** and cleansing techniques such as under-over sampling and feature standardization. The authors highlighted the use of models to detect diabetes early based on a range of data. They applied ensemble methods such as Extra Trees Classifier and Random Forest which I also applied in my study. Their project included pre-processing techniques which lead to 86% accuracy in diabetes prediction classification problems. Their objectives were similar to my project. In addition to using suitable variables in a model to predict diabetes, they also outlined an approach to using the results in redesigning healthcare systems. This underlines the relevance of analysis of this kind of data for the healthcare industry.

Checking model accuracy was important in the project [5] as this informed which models would be used in the application. The authors identified problems that can arise from applying deep learning and the need to ensure sufficient resources available. They highlighted another paper which applied Naïve Bayes as a deep learning approach.

The **dataset collected** for the analysis was from the National Institute of Diabetes and Digestive and Kidney Diseases, from the Unites States National Institutes of health. The study used medical data based on information from patients. This is similar to my project however they used only one dataset whereas I am applying two different datasets. In analying the dataset of the study they highlighted similar features to features in my datasets such as; glucose, insulin, blood pressure, BMI etc. They also had a binary variable which was presumably the target variable of the dataset and was supposed to tell whether someone has diabetes or not.

There was a similar use of **visualisation** and they provided graphical visualisation of a group of bar charts showing the ranges and distributions of the features.

Data preparation: The dataset in their study was cleaner than mine and they didn't have to deal with missing data but mentioned some common methods such as interpolating missing values and dropping rows. I however, had to implement several methods to clean my data for one of the datasets. Different over-sampling techniques were mentioned by the authors including Random Over-sampling and Synthetic Minority Over Sampling (SMOTE). I used SMOTE to create another subset based on the Binary dataset as an experiment. I also however, had utilised down sampling as a method as well which helps reduce the size of the majority class. This also helped with imbalanced data issues.

The paper also used feature standardization which was helpful with having the features the same range and are not dominated by one feature with larger values. Feature standardization has many techniques and two of the most common are standard scaler and Min Max scaler. Some of the algorithms implemented in the study were different to those which I had been familiar with. For example; Gaussian Naïve Bayes, Random Forest Classifier and XGB classifier. What was successful in this paper was using ensemble methods such as Random Forest which are proven to be high in accuracy and proving that machine learning can help with classification problems.

A second paper "**Performance Analysis of Classifier Models to Predict Diabetes Mellitus,"** from Procedia Computer Science. [6] also touches on the topic of how to predict diabetes at an early stage. This also focuses on comparing the performance of algorithms to predict diabetes using data mining techniques. The algorithms used in this study are; Decision Tree J48, KNN (K- Nearest Neighbour), Random forest and Support Vector Machines. The data used for the project was supplied form the UCI Machine Learning Repository.

In this study, there were two different subsets used during the experimentation, the original dataset (which contained noise) and the other dataset which was pre-processed. The results were compared in terms of accuracy, sensitivity and specificity. The paper outlined the complications associated with the disease including cardiovascular disease, blindness, kidney failure, and lower limb amputation. The J48 Decision Tree was explained using pseudocode which can be viewed in Figure 1.

The Pseudo code of J48

1) Check for base cases
2) For each attribute *a*
    a) It checks for normalized information gain on *a*.
3) Amount the best information gain of the attribute *a_best* it selects the attribute which has highest information gain.
4) It creates a decision node with that attribute.
5) This process is repeated with sub list of the nodes and added to its child node.

Figure 1: Pseudocode for J48 Decision tree

The KNN classifier was used to store all available cases and classify new cases based on a similarity measure where Euclidean distance formula was used. A case was classified by the majority vote of its neighbours, with the case being assigned to the class most common amongst its KNN measured by a distance function. The value of k was determined by inspecting the dataset. Cross validation is another way to retrospectively determine a good k value by using an independent dataset to validate the k. Random Forest is a statistical and machine learning algorithm that utilises multiple learning algorithms to obtain better predictive performance than others. In the study this was divided into two parts: tree bagging and random forest. I assume that Support Vector machines were used as a regression method maintaining the main features that characterise the algorithm. The primary use of the algorithm was to predict the class membership for categorical targets by constructing hyperplanes in a multidimensional space that separates cases of different class labels.

The team carried out a series of experiments for diabetes mellitus. The dataset collected contained 8 features and a predicted class. Weka was used as the main tool for the experiment. The results were also compared in it. From the confusion matrix the sensitivity, accuracy and specificity were calculated. The different sets of results were evaluated and the Decision tree J48 classifier achieved more accuracy than the other classifiers when the data was not pre-processed. The KNN and random forest achieved better when the data was pre-processed. It was proven that the accuracy was much higher when the data was pre-processed. This means when noise is removed it provides a better solution to the problem.

The usage of different data subsets was key in my project as I had used a regular splitting dataset, down sampled dataset, and smote dataset from the binary dataset (Also called the Diabetes Heath Indicators dataset). For the Non-Binary dataset, I had two different subsets used in the experiments which were a regular splitting and a stratified subset.

3.0 Data

In this project, I used **two different datasets** sourced from Kaggle. These are:

- The **'10 Years Diabetes Dataset'** which was sourced from Kaggle. It was extracted in CSV format. The size of the dataset was approx. 19.66 Megabytes. The dataset represents 10 Years (1999-2008) of clinical care sourced at 130 US hospitals and integrated delivery networks. It has over fifty features representing patient and hospital outcomes. The exploration analysis for this dataset was carried out in RStudio.
- The "**Diabetes Health Indicators"** dataset. This dataset primarily contains data about risk factors for diabetes also over a 10 year period. It however does not hold any appropriate binary information that I need for my project. This is where the **diabetes heath indicators dataset** comes into play. In this dataset, there is a column that is represented in binary format and gives information about whether someone is diabetic. In my data preparation work, I had intended to remove this column as mentioned above and add it to the diabetes dataset alongside other features that are relevant. In the end, I however, decided to carry out a separate analysis on both of the datasets.

3.1 10 Years dataset details

The information stored in the 10 Years dataset was drawn from the following encounters:

1. Inpatient encounter (a.k.a. hospital admission).
2. It is a diabetic encounter, that is one, during which any kind of diabetes was entered to the system as a diagnosis.
3. The length of stay was at least 1 days and at most 14 days.
4. Laboratory tests were performed during the encounter.
5. Medications were administered during the encounter.

The type of attributes contained in the 10 years dataset are:

- Patient number
- Race
- Gender
- Age
- Admission type
- Time in hospital
- Medical speciality of admitting physician
- Number of lab test performed
- HbA1c test result
- Diagnosis
- Number of medications
- Diabetic medications
- Number outpatient
- Inpatient
- Emergency visits in the year before the hospitalization.

3.2 Diabetes Health Indicators dataset details

The **'Diabetes Health Indicators'** dataset was sourced from Kaggle. It was extracted in CSV format as three separate files. The Behavioural Risk Factor Surveillance System (BRFSS) is a

health-related telephone survey that is collected annually by the CDC in the USA. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviours, chronic health conditions and the use of preventative services. It has been conducted every year since 1984. For this project, a CSV file of the dataset available on Kaggle for the year 2015 was used. This original dataset contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

Technical details:

- Clean dataset of 253,680 survey responses to the CDC's BRFSS2015.
- The target variable has 3 classes.
    - 0 = No Diabetes
    - 1 = Pre-Diabetes
    - 2 = Diabetes
- There is class imbalance in this dataset
- The dataset has a total of 21 features.

3.3 Exploratory analysis for '10 Years Diabetes Dataset':

I carried out a **correlation analysis** on the 10 Years diabetes dataset. This involved preparing the data that made it suitable to be used for such analysis. Steps taken here were; creating functions that would clean the data and factor if necessary. Instead of imputing the missing values in the dataset, I decided to swap the missing values with 'unknown'. All the categorical variables had to be encoded and the datatypes had to be changed also. For the correlation test, I created pairs of two features. I applied **Pearson's correlation** which is commonly used to find out about the relationships between pairs of features. I used pair panels to visualise the results of the correlation tests. However, these didn't say much, so I instead used the statistics to understand the test results. On one instance, I have used multiple correlation tests which meant having more than two features for comparison. I found this more effective to visualise because of the breadth of the visuals.

For the **statistics analysis**, I did statistics on two data frames for the dataset. One was prepared for the **correlation analysis** and the other was created for **PowerBI visualisation**. I mainly used the correlation dataset as everything had to be in numeric format for the statistics to work. Statistics were collected for both data frames which gave me some understanding of the datasets. However, I wanted to get them for individual columns, so I used the correlation dataset as it was most suitable for the task. Implementing **histogram analysis** of the features was beneficial in understanding the value distribution. **Frequency analysis** was difficult to interpret because all the values were in numbers so this impacted the headings of the tables. **Student t-tests** were carried out to see whether a particular feature has an impact on the dataset. **Non-parametric tests** such as **Pearson's Chi-Squared test** was used to understand whether two features are related to each other. The aim of using such a method was to identify whether a particular feature is independent of the other. In other words, is there a connection between the two.

I carried out some **imputation methods** in the hope that this would clear the missing data from the dataset. Some of these tasks involved replacing missing values with NA's and analysing the frequency distribution of missing values within each column. As there were NA's in the dataset, I tried to use "**Mice"** as an imputation method but due to data requirements this was discontinued. I had also tried to replace range values with numbers, but missing data resulted from this, so I had to stop using this method. I did checks of NA's per column as well, just in case I made a decision to drop the column due to its' quantity of missing values.

In addition to correlation and statistics analysis of the 10 Years diabetes dataset, I prepared the dataset again. In this instance, the binary columns remained unchanged and the same encoding and data changing process was carried out.

3.4 Exploratory analysis for Diabetes health Indicators dataset:

I carried out some data cleaning for this dataset by removing all the missing values with 'unknown', as I had done for the other dataset.

I carried out some data preparation for the dataset by setting the binary values in the binary column as categorical values; 'Yes' and 'No'. I had originally planned to use this format in PowerBI for visualisation work but instead I decided to move to the machine learning phase of the project.

For statistical analysis, I carried out similar steps on the other dataset and experimented with using heatmaps to investigate the correlation in the dataset. However, due to the large number of features in the dataset, this was difficult to manage.

I then decided to use the **Cohens Kappa score** to test the accuracy of the classification algorithms. It is a popular tool for classification problems. To do this, I split the data into two separate variables and applied one hot encoding to get the data into a particular format. This was done by using Dummy variables and the Kappa score could then be computed. However when I attempted to do this with my dataset, this approach became problematic due to the excessive size of the dataset.

4. Methodology and Implementation

I used the CRISP-DM methodology for my project as it best mapped onto the project structure and objectives.

4.1 Business Understanding:

The challenge of predicting diabetes trends is an important task as health care services need to understand the underlying causes of the disease and how these relate to lifestyle data. The selected datasets were identified to support analysis of the data associated with the above problem. By identifying risk factors that cause the disease to occur, it should be possible to spot patterns from the data and draw conclusions about lifestyle recommendations which could be of use to healthcare providers.

4.2 Data Understanding:

After compiling the data from the relevant sources, I sought to understand it better by using exploratory and statistical analysis. I carried out this work using RStudio and PowerBI. In PowerBI, I carried out some visualisation tests on a number of features to see what patterns could be understood and to test the importance of the features.

Correlation Analysis was a key area in my project, and this involved carrying out tests and understanding how different variables were related and understanding the use of multiple variables. This type of test is known as a **parametric test**. The key value taken away from the test was the p-value which helped me understand the significance level of a feature.

I carried out statistical analysis on the data frames to understand the **mean distribution** of values in every column and check the **standard deviation**. This is important for other statistical methods like the **chi-square test**. **Histogram analysis** was used also to understand the value distribution within each feature - this method works for large datasets.

The **student t-test** was used to check the means of two features and see whether they were similar. This particularly useful for data that has an even distribution.

Pearson's chi-square test was also utilised to investigate whether distributions of categorical variables differed from one another. When carrying out the Chi-Square test the example would appear as; 'Test whether a person's gender is independent of their insulin level at .05 significance level.' If the p-value was greater than that significance level then the feature is independent of the other feature and this would show a very weak correlation between the two variables.

4.3 Data Preparation:

The main preparation techniques carried out involved getting the data ready for statistical and exploration analysis. The techniques which I explored included **Imputation methods, Replacing values, creating data subsets, merging data frames, One-Hot encoding and Encoding values**.

Imputation methods were used for dealing with the NA's problem at the beginning of the data cleaning process. The packages which I experimented with in this area included MICE, DPLYR, TIDYR and a number of others. These methods however required data to be presented in a particular format and as a result I chose not to continue with this method.

Creating data subsets was proposed to help understand correlation amongst variables better using a lower volume of features. As a next step, I had intended to create a new data frame which would merge information from the two datasets together to create one whole data frame.

I looked at implemented the One-Hot encoding method to transform my data for the Cohens Kappa calculation. I found however that, as my data was in numeric format, implementing OHE was unnecessary. This was because the data exceeded limits for the kappa function to work.

I also used label encoding to create individual factor ranges for one of my variables that had range like data. However, this resulted in errors with NA values automatically being put into the data frame.

4.4 Modelling:

After the datasets were cleaned off and all missing values were removed, I moved on to selecting classification models. The models chosen for this project were **Random Forest, Decision trees and Naïve Bayes**. I also planned some artificial intelligence analysis by implementing **Neural Networks**.

The **Diabetes Health Indicators dataset** had three different subsets created; Regular splitting, Down Sample and Smote. The reason for this was due to modelling issues and uneven proportions in the dataset. I decided to experiment and explore which of the three subsets performed best in the classification analysis across all four classification models. Down sample has the benefit of downsizing the majority class to the same number as the minority class. Smote however, does the opposite by creating more copies of the minority class to match the majority class in the training data before fitting to the model.

The **"10 Years Diabetes Dataset"** did not have the same binary classification class imbalance issue that the other dataset had. Different R scripts were created during the modelling phase to separate out the models. I used regular splitting in my approach and also initially looked at using stratified sampling. Stratified sampling involves creating a subset so that both binary classes of a variable are even. I however dropped the stratified approach because the proportions for the train and test datasets were fairly balanced in terms of gender attributes (this was the target variable in that dataset).

Platforms used for modelling: As mentioned earlier, most of the work is carried out in RStudio with the models. Random Forest, Decision trees and Naïve Bayes were carried out on both datasets. The Neural Networks were implemented in Jupyter Notebook using Python as this had better library resources.

Smote: Smote (Synthetic Minority Oversampling Techniques) is a common method for dealing with imbalanced datasets in classification problems. This technique was useful in overcoming the majority class outweighing the minority class. This was for the target variable in the Binary dataset. This would produce more copies of the existing data of the minority class to balance out the majority class.

Down sampling involves reducing the number of samples in the majority class and making it more balanced with the minority class. It is another common technique in overcoming imbalanced data issues in classification problems.

The scripts differed in terms of modelling structure because the two datasets have different subsets.

4.5 10 years Diabetes Dataset - modelling structure

With this dataset, the implementation had a different procedures for Random Forest, Naïve Bayes and Decision Trees models.

For Random Forest, the process was as follows. Split the data into train and test, set a seed, fit train data to the model with hyperparameters defined, carry out predictions, access confusion matrix and check for variable importance.

For **Decision trees, the process was as follows:** Split the data into train and test set a seed, fit train data to the model with hyperparameters defined, carry out predictions, access confusion matrix and get the accuracy of the model.

For the Naïve Bayes implementation, I implemented scripts for this dataset using regular splitting. I created a subset for some variables instead of using the whole dataset and regular splitting was performed on this. Train data would then be fit to the model and the predictions would be carried out. Typically, I would use the train data to fit the model. Following this, I would compute the confusion matrix and accuracy for the train and test sets from the regular splitting.

4.6 Diabetes Health Indicators Dataset - modelling structure

For this dataset, the process depended on the subset in question. These subsets were; Regular Splitting, Down sample or Smote.

The Random Forest modelling had the following steps followed for each subset. For regular splitting, the process was; splitting the data, fitting to the model, getting predictions, accessing confusion matrix, plotting the model and finding variable importance.

- For **down sample**, the process implemented was; collecting a down sampled version of the data, split the data and analyse the proportions, for the data to model and define the target variable, calculating predictions, accessing the confusion matrix, plotting the model and checking variable importance.
- For **Smote,** the process implemented involved collecting a Smote version of the dataset, checking distributions, splitting the data, checking proportions, fitting data and defining the target variable for the model, collecting the predictions, computing the confusion matrix, plotting the model and checking for variable importance.

The **Decision trees modelling** had the following steps implemented. For **regular splitting,** the data was split, fitting to the model with hyperparameters defined, collecting predictions, and then getting the overall accuracy.

- In the case of **down sampling**, a down sampled version of the data was collected, the data was split, data was fitted to the model, predictions are calculated, and accuracy was computed.
- For **Smote**, the smote version of the data was collected, data was split, fitting the data to the model was followed, predictions calculated, and accuracy computed.

The **Naive Bayes modelling** had the following steps implemented**;**

- For **regular splitting** the process was; creating a subset of the original data, splitting the data, factoring the features before fitting to the model, data fitting to the model, calculating the predictions, getting the confusion matrix and accuracy for the train and test sets.
- For **down sampling**, the process was, collecting a down sampled version of the data, splitting the data, checking proportions, factoring variables, fitting the data to the model, collecting predictions, computing the confusion matrix and accuracy for the train and test data.
- For **smote,** the process was, creating a subset, apply smote to get a smote version of that subset, split the data, check proportions, factor variables before model fitting, fit data to model, calculate the predictions, compute the confusion matrix and accuracy for both the train and test data.

4.7 Neural Networks modelling

For the **Neural Networks** implementation in Jupyter Notebooks , I created four different models. The two dataset models together made up 4 different sets of results. The format to run all the different scripts was the same with the different .csv files being imported at the start.

The relevant packages had to first be imported along with some key libraries including; tensorflow, tensorflow.keras.layers, keras_tuner etc. When the dataset was imported, the target variable was separated from the dataset. For the smote dataset (which is one of the Binary dataset models), I had to drop some redundant columns from the dataframe. Early stopping was implemented as a call back to help stop the neural network from training once all weights had converged. Hyperparameters defined in this were; monitor, patience, mode and min_delta. During the implementation in each notebook, I did a static model example and a model defined within a function which was necessary when applying Keras Tuner. The static model involved adding layers and activation functions. The model would then be compiled using a; loss, optimizer and metric. Before running my model, the dataframes were coverted to tensors. These were applied as parameters to the model.fit function which would run the neural network. Other parameters applied here were, epochs, batch size, callback and validation split. This produced a simple Keras model and parameters could be observed in the output.

In order to incorporate the Keras tuner, the model set up had to be carried out in a different way and I created instead a HyperModel. This was better for the search space. The build method within the function created a Kera's model using the 'hp' argument to define the parameter and return a model in a similar way to the approach I did for the static model. This involved me inputting a layer of neurons, at the start, before defining the number of layers, units and activation function. The output layer was defined and the learning rate was instantiated with the loss, metrics and optimizer specified as parameters.

The Kera's tuner was beneficial as the model could be hooked up to Tensorboard for visualisation analysis. The tensorboard was defined by setting a log directory where the results from the Neural Network model would be stored for visualisation analysis later.

I used the default random search when setting up the Tuner and the hypermodel was created including setting up the following; objective, max trials for Neural Network training, executions per trial and project directory. In the project directory, the trials run was stored which provided a checkpoint per trial folder. The detail instantiated in the hyper model was observed when I used search_space_summary() to see what the Keras tuner would capture during the trial runs. When the Neural Network finished running the information returned included; value accuracy from second last trial, best accuracy overall from the trials and the total elapsed time which can be extensive depending on the data and conditions set.

I ran tensorboard in my Jupyter Notebook by loading the magic keyword for tensorboard and then launched it using the log directory set. Logs could also be cleared from the folder as needed.

4.8 Evaluation:

One the models finished running, the predictions could be collected and I could the get the confusion matrix which is central to classification machine learning problems. The confusion matrix is a common way of representing, True Positives, False Positives, True Negatives and False Negatives. This is calculated on class predictions. The accuracy in some cases can also be collected for the models but for Random Forest this can be manually computed by using the confusion matrix to compute performance. It was also important to figure out what variables had an effect on the model and was shown for the Random forest models. The Neural Networks were analysed based on the information returned by the models and the information from the visualisation about the models through Tensorboard in Jupyter Notebook.

4.9 Deployment:

I ran the models from Jupyter Notebook and RStudio on my laptop. As I had upgraded my laptop to 16Gb RAM I did not have significant performance issues. Experiments that took more time to deploy were those collecting Smote datasets and running Neural Networks. The run time for the R scripts arose from the data being used and their subsets. The run time for the Neural Networks was dependent on hyperparameter settings such as; epochs, batch size etc. Batch sizes were affected by the amount of data being trained in the network meaning the bigger the data size the bigger the batch would be and as a result the training could be slow. Neural network models that used Tensorboard as a callback took longer to run than using the early stopping as a callback, however the purpose was to be able to visualise the data and collect visuals. Tensorboard was launched through the Jupyter Notebook using the notebook magic keyword.

5 Analysis

5.1 Statistical Analysis

I carried out the following **preliminary statistical analysis** as part of the project

Understanding the centre of mean for features:

As part of the initial stage of the project I had carried out histogram analysis on various features from the ’10 Years Diabetes dataset’. The purpose of doing this was to understand

value distribution per feature and to understand the centre of mean in the features based on where the graph peaked. When observing the 'x' and 'y' axis, the frequency on the y-axis indicated the number of times a certain value on the x-axis appeared in the feature.

T-Tests:

I conducted t-tests is to check whether there was a significant correlation between the mean of two same or different groups. The t-test could also help me by explaining whether the results come out by chance or whether there was an actual difference between the means of two groups. The important points to note about the t-tests were the; 'p-value' and the 't-score'. The **'t-score'** was the difference between the averages of the two samples divided by the difference that appears within the two samples. In other words, you give a measure of the difference of one sample from another relative to the differences within the sample itself. The **'df' value** represented the degrees of freedom. The **'p-value'** represented the likelihood of falsely rejecting the null hypothesis. The null hypothesis being a simple statement that tells you that there isn't a significant difference in the mean of the two samples being tested.

Correlations:

I decided to do a correlation check because this was one way of understanding whether features should be retained kept or not. It is possible that sometimes when we have large dimensions of data that not all the features are particularly relevant to the problem that is being solved. Correlations also help by showing patterns between features. However, a positive correlation doesn't mean that one variable affects another. It simply provides an indication of what features should be investigated further.

F**eature Significance:**

Feature significance was an important part of the project as it provided me an indication of whether a feature has high importance. This would mean I would only focus on features that have a p-value less than a particular range for example. Another way of addressing this would be removing features that correlate too high with other features. This would show redundancy that could be problematic for the models.

Chi-Square test:

I selected the Chi-Square test because it is a very common way to understand the level of independence among features. This was more focused on the independent features as opposed to the dependent feature which I was trying to predict. The statistics that were produced during the test were important, such as the p-value as they could affect what hypothesis I would select (Null or Alternative).

5.2 Modelling Analysis

Decision Trees:

I selected the Decision Trees model because of it's ability when it comes to predictive analysis. It works particularly well with categorical output variables. The purpose of using this classification model was to predict whether someone could be deemed diabetic based

on risk factors. The risk factors can be seen in the decision trees below nodes. Stemming from these (nodes) are the chances of whether the risk factor influences someone being diabetic by labelling them yes or no. The further the nodes reduce, the less likely the risk factor is linked to someone being diabetic or at least its less relevant. If the yes outcome weighs higher than the no outcome, then its likely the risk factor is related to someone being diabetic. The tree was built from the root to the leaves. It was built using the training data and the test data was used to measure the model's performance by comparing it's predictions to the actual values.

However, the above is not sufficient to tell whether the decision tree is accurate. The **confusion matrix** can help solve this issue. The four different values on the matrix axes are the True Positives, False Positives, True Negatives and False Negatives. The objective of using the matrix is to understand the performance of the machine learning model on the test data. The reason for the table being a 2x2 is that this is a binary classification problem. The side (Y) axis represents the predicted values and the top axis (X) represents the actual values. For example, a **True Positive** would mean that the total counts having both predicted and actual value are not diabetic. While the False Positive means the total counts having prediction is not diabetic while actually, they are diabetic. **True Negative** means the total counts having both predicted and actual values are diabetic. **False Negative** on the other hand means the total counts having prediction is diabetic while it's not diabetic.

Random Forest:

I selected the Random Forest model because it can point out whether an outcome belongs to either of the two binary classes which my project was about. The key items to consider here were the confusion matrix, the predictions, model and variable importance. The confusion matrix works on the number of cases assigned to the True Positive, False Positive, true Negative and False Negative. The following metrics; **accuracy, precision, recall and F1-score** can also be calculated from the confusion matrix. **Accuracy** measures the performance of the model, **Precision** is a measure of how accurate the models positive predictions are. This is defined as the as the ratio of true positive predictions to the total number of positive predictions made by the model. **Recall** measures the effectiveness of a classification model by identifying all relevant instances from a dataset. The **F1-score** is used to evaluate the overall performance of the classification model. It is the harmonic mean of precision and recall.

The predictions give an indication for each row in the dataset; whether it is a male or female, who is diabetic/non-diabetic. The model provides important details showing how many features are tried at every split in the Random Forest, the "Out of Bag" error rate and the Confusion Matrix. The variable importance tells which variables were likely to impact on the Random Forest model. When the model is plotted the number of trees on the x axis is compared against the error rate (OOB) on the y-axis. The OOB error rate is also helpful for evaluating the model's accuracy.

Naïve Bayes:

Naïve Bayes is a supervised non-linear classification algorithm. Based on the Bayes theorem, it assumes that the occurrence of a feature is independent of the occurrence of another feature. The important areas to note in this analysis are the model, predictions, and confusion matrix. The model shows that the conditional probability for each feature or variable is created by the model separately. The a-priori probabilities are also calculated which indicates the distribution of the data. The confusion matrix is interpreted by doing the total counts per feature and checking how many classified by that feature and others for example; if there were total number of 20 No's, 19 of those counts were classified as No while 1 count was classified as Yes.

Neural Networks:

This was the main application of AI in the project. Neural networks are useful in binary classification problems for making predictions. Some key areas to consider in the analysis of using Neural Networks included utilising early stopping to stop a neural network when all weights had converged, the model creation, Kera's tuner, and tensor board. The early stopping is effective when training a neural network because instead of running the total number of epochs set in hyperparameters of the model, it stops when all weights have converged.

I evaluated the use of static neural networks and the use of a function to create the neural network. The static model could provide an understanding of how a basic neural network works without the use of the tuner and tensor board. The key areas to observe in the model was; accuracy, value accuracy and loss. When it comes to creating a hyper model more analysis can be observed such as the search space size and what the Keras tuner is capturing within the space such as; min_value, max_value, and units. The result of the training provides important details during the running time and afterwards. During the running time it is easy to observe whether the model is running slowly if messages appear during the run time. The number of units captured, activation functions used, the learning rate value and layers used per trial can be observed in a mini data frame above the output space. The best overall accuracy and the previous accuracy of the last trial can be observed and especially the time taken to run the network. Accuracy is a key area as it can help understand whether a network performs well or not based on the conditions set prior to running it.

6.0 Results

6.1 Statistical Results

During the course of the project, I collected results from the different analysis from the preliminary stages to the modelling phase. The following charts and commentary is provided in this section to show the results of the statistical analysis and classification models performed. It reviews the comparison between the models and what the results reflect based on the analysis detailed in the previous section.

Pre-Liminary Stages:

Histogram Analysis

The histogram analysis reflected in **Error! Reference source not found.**2 and 3 explain the center (also known as the mean) of the data. The features provided in these figures originated from the '10 Years Diabetes dataset'. We can observe in 2 that the race category peak happens at Category '3' on the x-axis.

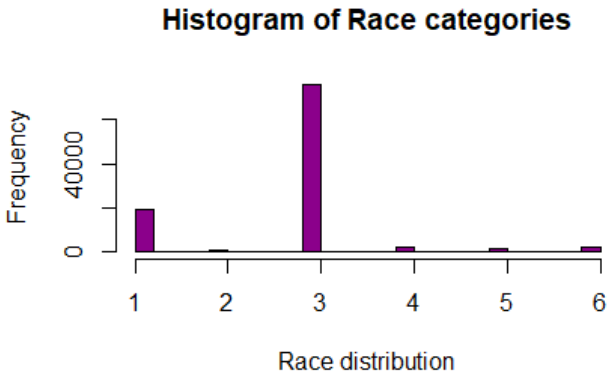


Figure 2: Histogram for count of race categories

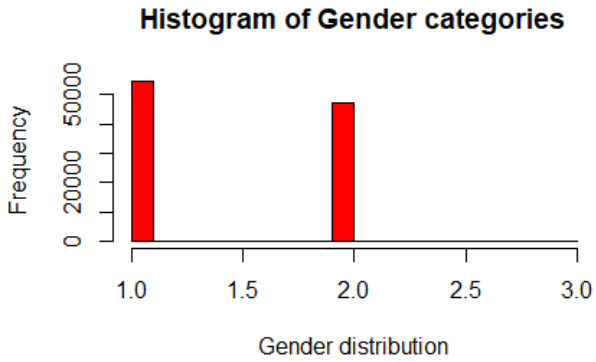


Figure 3: Histogram for count of Gender categories

In **Error! Reference source not found.**3 the frequency is indicated at '1.0' and 2.0 on the x-axis representing the two (M/F) genders.

T-Tests

The tests conducted below provide features from the '**10 Years Diabetes' dataset**. The type of t-test I conducted in this project was a One Sample t-test. This is about a condition that is checked to see whether the data is normally distributed. A particular formula is given for a single mean. In my case, I was using the t-test to help find out what someone's gender was or what race category does a person belong to as detailed in **Error! Reference source not found.** and **Error! Reference source not found.** below. I concluded from both tests that the p-value is significantly low and as a result the null hypothesis was true.

```
        One Sample t-test

data:  correlation_df$gender
t = 935.59, df = 101765, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.459380 1.465507
sample estimates:
mean of x
 1.462443
```

Table 1: T-test for gender column

```
        One Sample t-test

data:  correlation_df$race
t = 885.07, df = 101765, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.726737 2.738841
sample estimates:
mean of x
 2.732789
```

Table 2: T-test for race column

Chi-Square Tests

Pearson's chi-square evaluates how likely it is that any observed difference between two or more groups of categorical variables arose by chance. It can be used to tests for goodness of fit, homogeneity or independence.

I was using the test to test for independence in a contingency table. The test gives a 'p-value' which I could use to conclude whether the hypothesis of independence was true or not. In other words, the p-value is the chance of the values being independent of each other.

Typically, if the value for p is greater than 0.05 it shows there isn't any relationship amongst the values. It can be observed in **Error! Reference source not found.** that the p-value is above the 0.05 threshold meaning the insulin doesn't depend on the gender of a patient.

However, when we observe **Error! Reference source not found.**, the p-value is significantly lower than the threshold meaning that there is a strong relationship between the amount of insulin a patient takes depending on their ethnicity.

The use of the Chi-Square test plays an important role in the field of medicine and helps understand relationships between risk factors for other diseases such as arthritis.

```
> # Chi-square tests: Pearsons chi-squared test
> ch_test_1 = data.frame(correlation_df$gender,correlation_df$insulin)
> ch_test_1 = table(correlation_df$gender,correlation_df$insulin)
> print(ch_test_1)

      1     2     3     4
  1  6593 25497 16608  6010
  2  5625 21883 14241  5306
  3     0     3     0     0
> print(chisq.test(ch_test_1))

        Pearson's Chi-squared test

data:  ch_test_1
X-squared = 5.6899, df = 6, p-value = 0.4588
```

Table 3: Chi-Square test for gender and insulin

In the above **Error! Reference source not found.** we can observe the results from the Chi-square test for the gender and insulin columns in the 10 Years Diabetes Dataset.

```
> ch_test_2 = data.frame(correlation_df$race,correlation_df$insulin)
> ch_test_2 = table(correlation_df$race,correlation_df$insulin)
> print(ch_test_2)

       1     2     3     4
  1 2445  8024  6555  2186
  2   69   333   156    83
  3 9011 36600 22223  8265
  4  253   934   584   266
  5  261   605   384   256
  6  179   887   947   260
> print(chisq.test(ch_test_2))

        Pearson's Chi-squared test

data:  ch_test_2
X-squared = 559.36, df = 15, p-value < 2.2e-16
```

Table 4: Chi-Square test for race and insulin

In the above **Error! Reference source not found.** we can observe the results for the Chi-Square test for race and insulin from 10 Years Diabetes Dataset.

6.2 Modelling Stage Results

As a result of the modelling. I produced 12 sets of results for the **Diabetes Health Indicators dataset** and four sets of results for **the 10 Years Diabetes dataset**. In the following section I have detailed the results for the classification models and these are grouped by dataset usage. The bulk of the classification modelling for the project was completed in RStudio.

6.2.1 Decision Trees modelling

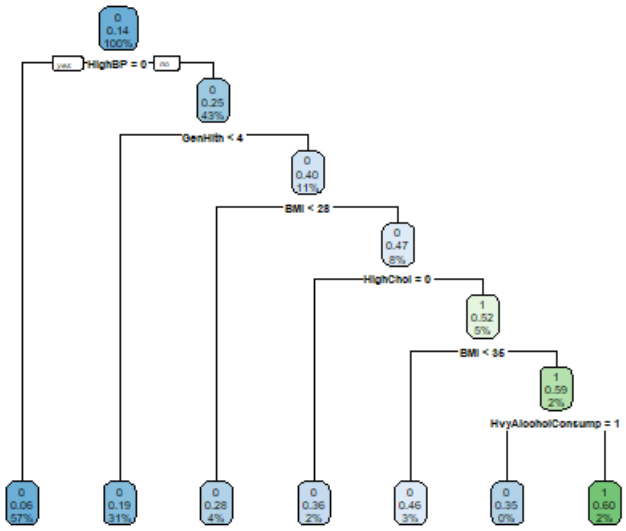
Decision trees summary : A decision tree structure is made up of root nodes, branches and leaf nodes. The branches that stem from the root node feed into the leaf nodes also known as decision nodes. Decision tree learning employs a “divide and conquer” strategy by conducting a greedy search to identify the optimal split points within a tree. The process is repeated in a top down recursive manner until all or the majority of the records have been classified under class labels.

Diabetes Health Indicators Dataset (Decision Trees)

The following figures detail the modelling results using Decision Trees on the Diabetes Health Indicators dataset. The results outline the findings based on the data subsets.

Regular Splitting

Figure 4: Decision tree Diabetes Health Indicator dataset Regular Splitting




```
> # Predicted model----
> predict_model<-predict(fit, Valid[, -2], type="class")
> m_at <- table(Valid$Diabetes_binary, predict_model)
> m_at
   predict_model
        0     1
  0 43332   399
  1  6370   635
```

Table 5:
Confusion matrix for Diabetes Health Indicator dataset regular splitting

```
> # Accuracy of model----
> accuracy <- sum(diag(m_at)) / sum(m_at)
> print(paste('Accuracy for test is found to be', accuracy))
[1] "Accuracy for test is found to be 0.866583885209713"
>
```

Equation 1: Accuracy for Diabetes Health Indicator dataset regular splitting

Downsample

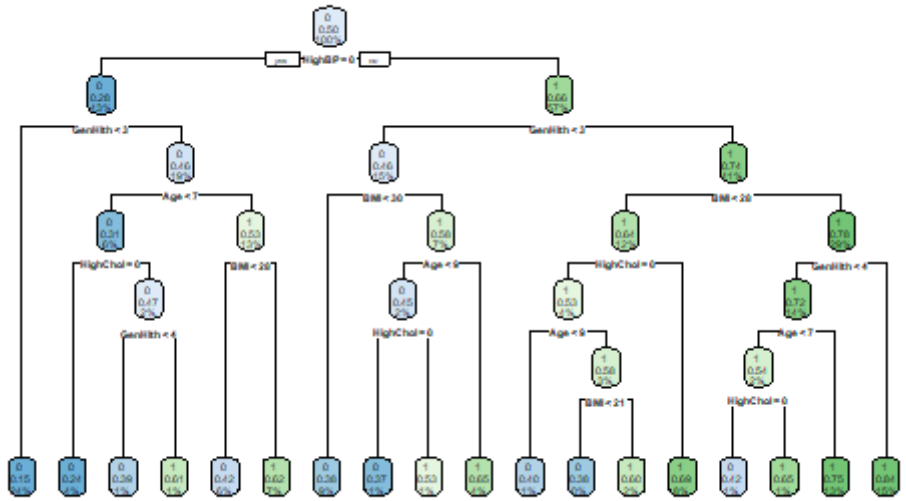


Figure 5: Decision tree for Diabetes Health Indicator dataset down sample

```
> # Predicted model----
> predict_model2<-predict(fit2, Valid2[, -2], type="class")
> m_at2 <- table(Valid2$Diabetes_binary, predict_model2)
> m_at2
   predict_model2
       0    1
  0 5038 2031
  1 1730 5339
```

Table 6: Confusion matrix for Diabetes Indicator dataset down sample

```
> # Accuracy of model----
> accuracy2 <- sum(diag(m_at2)) / sum(m_at2)
> print(paste('Accuracy for test is found to be', accuracy2))
[1] "Accuracy for test is found to be 0.733979346442213"
>
```

Equation 2: Accuracy for Diabetes Health Indicator dataset down sample

Smote

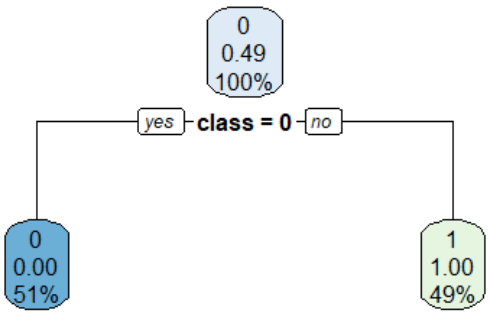


Figure 6: Decision Tree for Diabetes Health Indicator dataset Smote.

```
> # Predicted model----
> predict_model3<-predict(fit3, Valid3[, -2], type="class")
> m_at3 <- table(Valid3$Diabetes_binary, predict_model3)
> m_at3
   predict_model3
        0     1
  0 43663     0
  1     0 42419
```

Table 7: Confusion Matrix for Diabetes Health Indicator dataset Smote.

```
> # Accuracy of model----
> accuracy3 <- sum(diag(m_at3)) / sum(m_at3)
> print(paste('Accuracy for test is found to be', accuracy3))
[1] "Accuracy for test is found to be 1"
>
```

Equation 3: Accuracy for Diabetes Health Indicator dataset Smote.

10 Years Diabetes Dataset (Decision Trees modelling)

The following sections contain figures detailing the modelling results using Decision Trees on the **10 Years Diabetes Dataset.** The results outline the findings based on one data subset.

Regular Splitting

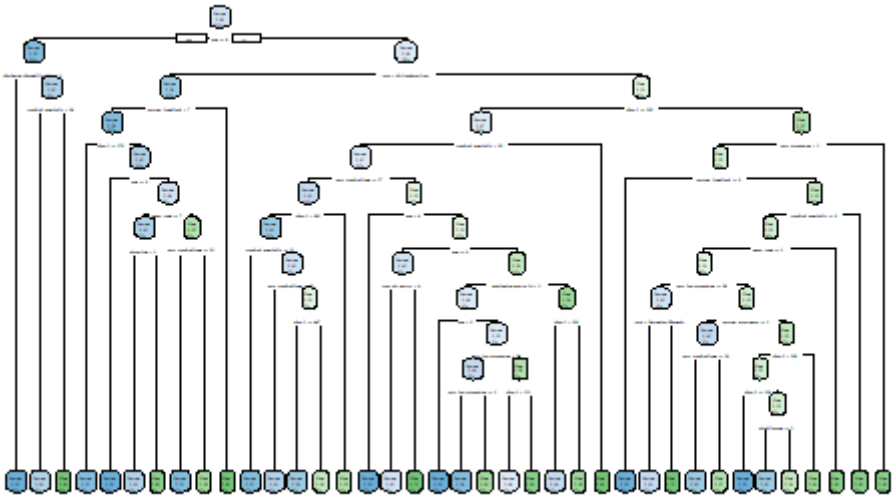


Figure 7:
Decision tree for 10 Years Diabetes dataset regular splitting

```
> # Predicted model----
> predict_model<-predict(fit, test[, -2], type="class")
> m_at <- table(test$gender, predict_model)
> m_at
        predict_model
         Female Male
  Female   1917  948
  Male     1285 1120
```

Table 8: Confusion matrix for 10 Years Diabetes dataset regular splitting

```
> # Accuracy of model----
> accuracy <- sum(diag(m_at)) / sum(m_at)
> print(paste('Accuracy for test is found to be', accuracy))
[1] "Accuracy for test is found to be 0.576280834914611"
>
```

Equation 4: Accuracy for 10 Years Diabetes dataset regular splitting

Evaluation of dataset confusion matrix scores with precision, recall and F1 measures (Decision Trees)

- **Precision =** (TP/TP + FP) : This measures how accurate a model's positive predictions are.
- **Recall =** (TP/TP + FN) : This measures the effectiveness of a classification model.

- **F1-Score =** (2xprecisionxrecall/precision + recall). This score measures the overall performance of a classification model.

For the Diabetes Health Indicators Dataset:

The regular splitting model shows that the number of samples in the confusion matrix is disproportionate. Despite the accuracy being high, I felt that there may be imbalance issues due to the low number of False Positive cases. The cases in the confusion matrix improved in the down sampling experiment and there is a better representation of the cases. The accuracy is also more realistic.

The Smote experiment proved to be very inaccurate due to the high scores from the precision, recall and f1-scores which were calculated from the matrix. There was also a lack of false positive and false negative cases in the matrix. This suggested that there may be oversampling issues with the smote dataset.

The confusion matrix results can be viewed below for all three different subsets below:

Regular Splitting

- **Precision =** 0.9991
- **Recall =** 0.8718
- **F1-Score =** 0.9311

Down Sampling

- **Precision =** 0.7127
- **Recall =** 0.7444
- **F1-Score =** 0.7282

Smote

- **Precision =** 1
- **Recall =** 1
- **F1-Score =** 1

For the 10 years Diabetes Dataset (Decision Trees):

My overall observation from the 10 Years Diabetes Model was that the model didn't perform well. The model was very busy with some 37 root nodes and 40 leaf nodes. This was due to the model having some overplotting issues. The accuracy observed was affected by this issue as well, as can be seen from the low score here. The binary values within the nodes could not be recognised because of this.

Most importantly though it is not very good at predicting the test . The first value in a node is meant to identify the node as being either non-diabetic = 0 or diabetic = 1. The tree is structured so that, at each root node, a binary split is created. When observing the yes and no directions of a split in a tree we can identify which outcome has the higher percentage. This will indicate whether that feature is involved in predicting the target feature.

The confusion matrix indicates that the number of True Positive cases outweigh the number of False Positive cases. The following scores can be calculated from this including; Precision, Recall and F1-Score. These are displayed below.

- **Precision** = 0.6691
- **Recall** = 0.5987
- **F1-Score** = 0.6319

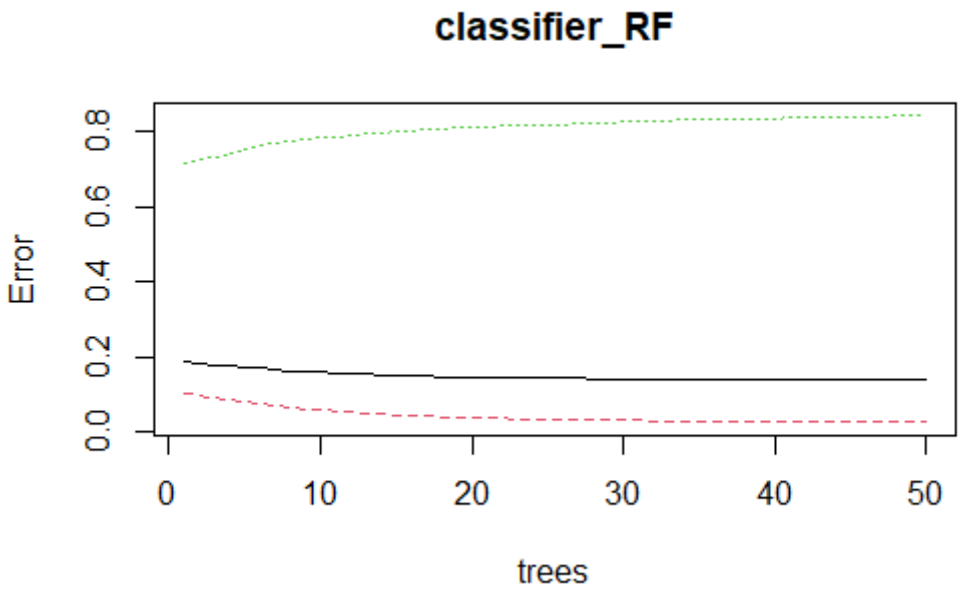
6.2.2. Random Forest modelling

Random Forest Modelling: A random forest is an ensemble of decision trees. This is a supervised learning algorithm. The forest builds an ensemble of decision trees usually trained with the bagging method. Bagging is used because a combination of learning models increases the overall result. Random Forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Random forest adds additional randomness to the model while growing trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features.

Diabetes Health indicators Dataset (Random Forest modelling)

The following sections contain figures detailing the modelling results using Random Forest on the **Diabetes Health Indicators Dataset.**

Regular Splitting



Random forest Chart legend (Diabetes Health Indicators dataset):

- ----- = Non-Diabetic (Red)
- ..... = Diabetic (Green)

Figure 8: Random Forest regular splitting Diabetes Indicator dataset diagram

```
> classifier_RF = randomForest(x = train[-2], y = train$Diabetes_binary, ntree = 50)
> classifier_RF

Call:
 randomForest(x = train[-2], y = train$Diabetes_binary, ntree = 50)
               Type of random forest: classification
                     Number of trees: 50
No. of variables tried at each split: 4

        OOB estimate of  error rate: 13.8%
Confusion matrix:
       0    1 class.error
0 166746 4198  0.02455775
1  23201 4387  0.84098159
```

Figure 9: Random Forest Diabetes Indicator dataset regular splitting model

```
> # Confusion Matrix
> confusion_mtx = table(test[, 2], y_pred)
> confusion_mtx
   y_pred
        0     1
  0 46474   916
  1  6657  1101
```

Table 9: Confusion Matrix Diabetes Indicator dataset regular splitting random forest

Downsample

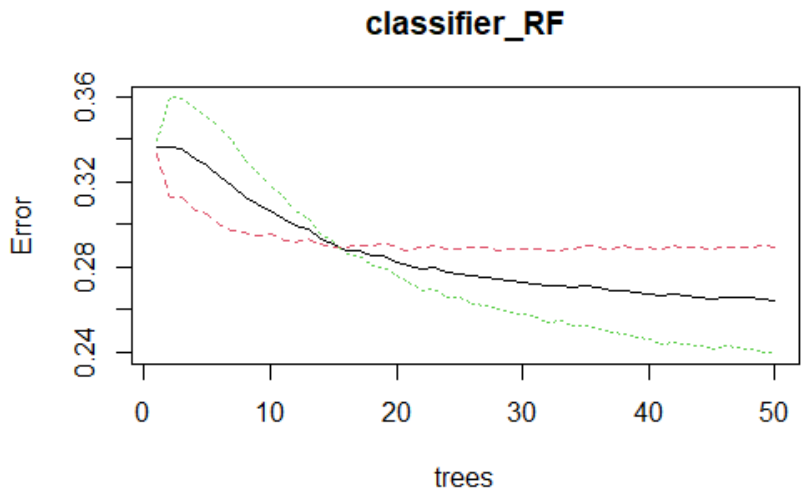


Figure 10: Random Forest Model Diabetes Indicator dataset downsample

```
> classifier_RF = randomForest(x = Train2[-2], y = Train2$Diabetes_binary, ntree = 50)
> classifier_RF

Call:
 randomForest(x = Train2[-2], y = Train2$Diabetes_binary, ntree = 50)
               Type of random forest: classification
                     Number of trees: 50
No. of variables tried at each split: 4

        OOB estimate of  error rate: 26.42%
Confusion matrix:
      0     1 class.error
0 20108  8169   0.2888920
1  6771 21506   0.2394526
> |
```

Figure 11: Random Forest Model Diabetes Indicator dataset downsample

```
> # Confusion Matrix
> confusion_mtx = table(Valid2[, 2], y_pred)
> confusion_mtx
   y_pred
       0    1
  0 5007 2062
  1 1546 5523
```

Table 10: Confusion matrix Random Forest Diabetes Indicator dataset downsample

Smote

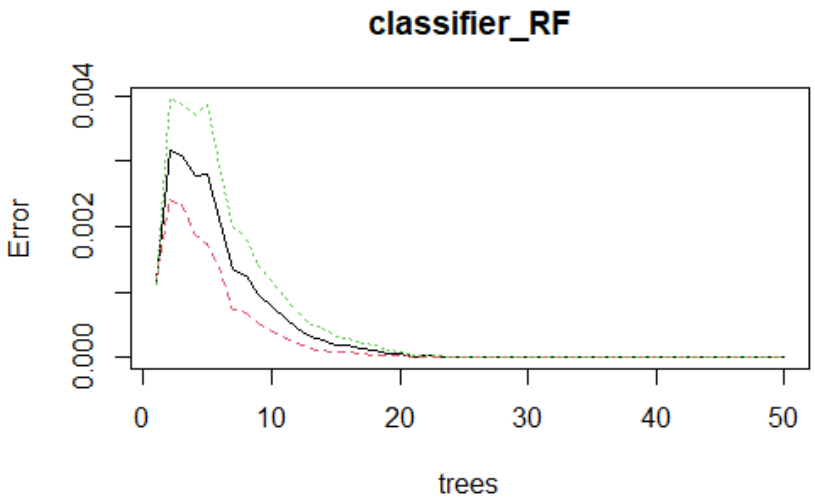


Figure 12: Random Forest model Diabetes Indicator dataset smote

```
> classifier_RF = randomForest(x = Train3[-2], y = Train3$Diabetes_binary, ntree = 50)
> classifier_RF

Call:
 randomForest(x = Train3[-2], y = Train3$Diabetes_binary, ntree = 50)
               Type of random forest: classification
                     Number of trees: 50
No. of variables tried at each split: 4

        OOB estimate of  error rate: 0%
Confusion matrix:
       0      1 class.error
0 174671      0           0
1      0 169657           0
> |
```

Figure 13: Random Forest confusion matrix Diabetes Indicator dataset smote

```
> # Confusion Matrix
> confusion_mtx = table(Valid3[, 2], y_pred)
> confusion_mtx
   y_pred
        0     1
  0 43663     0
  1     0 42419
```

Table 11: Random Forest Model confusion matrix Diabetes Indicator dataset smote

10 Years Diabetes Dataset (Random Forest modelling)

The following section contain figures detailing the modelling results using Random Forest on the **10 Years Diabetes Dataset.** The results outline the findings based on one data subset. **Regular Splitting.**

Random Forest Chart legend (10 Years Diabetes dataset):

- ----- **= Female (Red line)**
- ..... **= Male (Green line)**

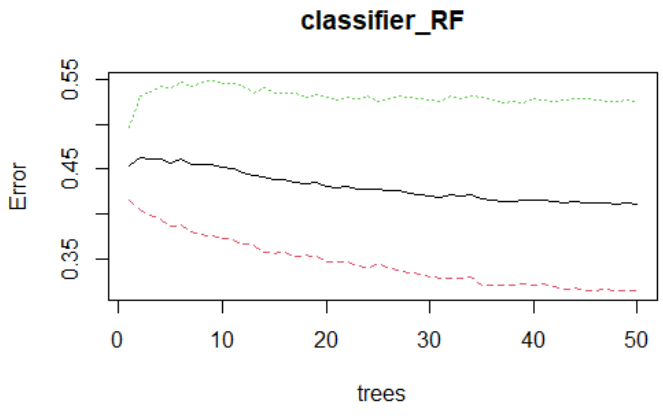


Figure 14: Random Forest Model 10 Years Diabetes dataset dataset Regular Splitting

```
> classifier_RF = randomForest(x = train[-2], y = train$gender, ntree = 50)
> classifier_RF

Call:
 randomForest(x = train[-2], y = train$gender, ntree = 50)
               Type of random forest: classification
                     Number of trees: 50
No. of variables tried at each split: 6

        OOB estimate of  error rate: 41.06%
Confusion matrix:
       Female Male class.error
Female   7844 3586   0.3137358
Male     5062 4568   0.5256490
> |
```

Figure 15: Confusion matrix Random forest model 10 Years Diabetes dataset dataset Regular Splitting

```
> # Confusion Matrix
> confusion_mtx = table(test[, 2], y_pred)
> confusion_mtx
        y_pred
         Female Male
  Female   2196  888
  Male     1341 1270
```

Table 12: Confusion Matrix Random Forest model 10 Years Diabetes dataset dataset Regular Splitting

Evaluation of dataset confusion matrix scores with precision, recall and F1 measures (Random Forest)

For the Diabetes Health Indicators Dataset:

I concluded from the above that the error rate for the classes differed with the three datasets. I observed that there was a big margin in the regular splitting. The error rate for the diabetes class was rather high and the error rate for the non-diabetic class was very low. This correlates with the class error rate in the confusion matrix of **Error! Reference source not found.**.

In the down sampled experiment, the error rates meet each other but there is little difference apart from one another as can be seen in the matrix in **Error! Reference source not found.**. The error rate is poor for Smote, and this is reflected in **Error! Reference source not found.**. Also, the confusion matrix had the same issues as the smote model for decision trees for the confusion matrix which had no classes for the false negative and false positive.

The metrics were inaccurate for the smote dataset, and they were best seen for the down sampled model which performed best. The regular splitting model had similar performance to that which occurred with decision trees. I also calculated the model's accuracy as this wasn't available with the original code. This is done by implementing (TP+TN/TP+TN+FN+FP).

The Confusion matrix metrics for the three subsets can be viewed below:

Regular Splitting

- **Precision =** 0.9807
- **Recall =** 0.8747
- **F1-Score =** 0.9247
- **Accuracy =** 0.9680

Down sampling

- **Precision =** 0.7083
- **Recall =** 0.7641
- **F1-Score =** 0.7351
- **Accuracy =** 0.7448

Smote

- **Precision =** 1
- **Recall =** 1
- **F1-Score =** 1
- **Accuracy = 1**

For the 10 Years Diabetes Health Indicators dataset

I concluded from the above is that the confusion matrix provided a good representation overall of cases. The False Positive cases remained low however, but it performed pretty well when compared to the down sampled model for random forest. The class error rate indicated a sharp rise for male and a low rise for female.

The confusion matrix metrics for this dataset can be viewed below:

- **Precision =** 0.7121
- **Recall =** 0.6209
- **F1-Score =** 0.6634
- **Accuracy =** 0.6086

6.2.3 Naïve Bayes Modelling

Naïve Bayes is a supervised machine learning algorithm which is used for classification tasks like text classification. It is also part of a family of generative learning algorithms which means it seeks to model the distribution of inputs for a given class or category.

For the Diabetes Health Indicators dataset (Naïve Bayes)

The following section shows figures detailing the modelling results using Naive Bayes on the Diabetes Health Indicators dataset. The results outline the findings based on the data subsets.

Naïve Bayes Chart legend (Diabetes Health Indicators dataset):

- -------- Red line= Non-Diabetic (0)
- ……. Green Line = Diabetic (1)

Regular Splitting

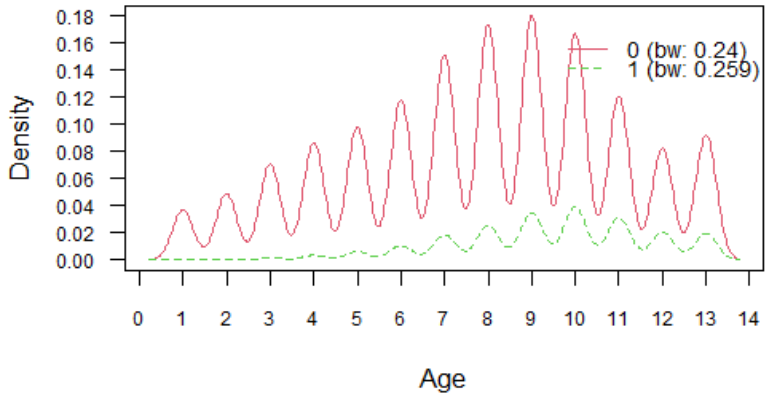


Figure 16: Naive Bayes model 'Age' Diabetes Indicator Dataset Regular Splitting

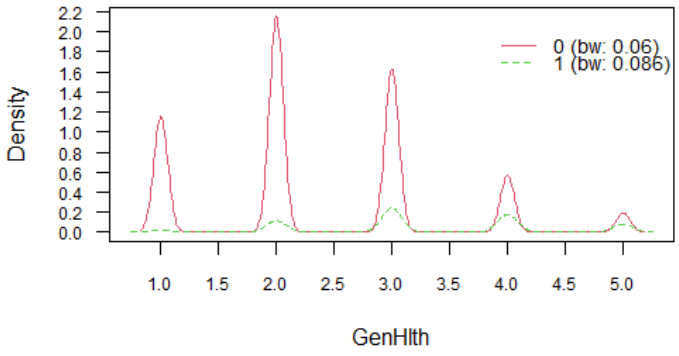


Figure 17: Naive Bayes Model 'General Health' Diabetes Indicator dataset Regular Splitting

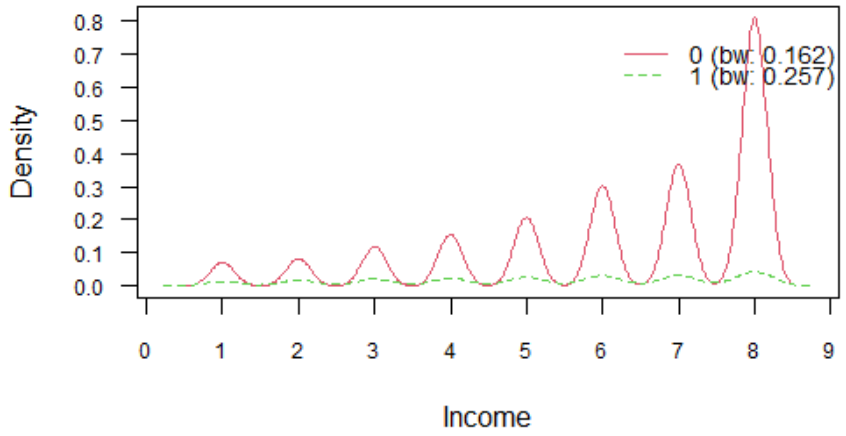


Figure 18: Naive Bayes model 'Income' Diabetes Indicator dataset Regular Splitting

```
> # Predictions----
> p <- predict(model, test_df, type = 'prob')
> head(cbind(p, test[, 2]))
             0           1
[1,] 0.9999533 4.668167e-05 1
[2,] 0.9949844 5.015558e-03 2
[3,] 0.9999921 7.855725e-06 1
[4,] 0.9999995 5.419891e-07 1
[5,] 0.9769925 2.300751e-02 2
[6,] 0.9890815 1.091848e-02 2
```

Table 13: Predictions for Naive Bayes model Regular Splitting Diabetes Indicator dataset

```
> # Confusion matrix for train data----
> p1 <- predict(model, train_df)
> (tab1 <- table(p1, train$Diabetes_binary))

p1       0      1
  0 174659  27974
  1    188    284
```

Table 14: Predictions for train data Diabetes Indicator dataset Regular Splitting

```
> #### This concludes the models accuracy
> 1 - sum(diag(tab1)) / sum(tab1)
[1] 0.1386573
```

Equation 5: Accuracy for train data Naive Bayes model Diabetes Indicator dataset regular splitting

```
> # Confusion matrix for test data----
> p2 <- predict(model, test_df)
> (tab2 <- table(p2, test$Diabetes_binary))

p2      0     1
  0 43444  7006
  1    43    82
```

Table 15: Predictions for test data Naive Bayes model Diabetes Indicator dataset regular splitting

```
> #### This concludes the models accuracy
> 1 - sum(diag(tab2)) / sum(tab2)
[1] 0.1393772
>
```

Equation 6: Accuracy for tests data Naive Bayes model Diabetes Indicator dataset Regular Splitting

Downsample

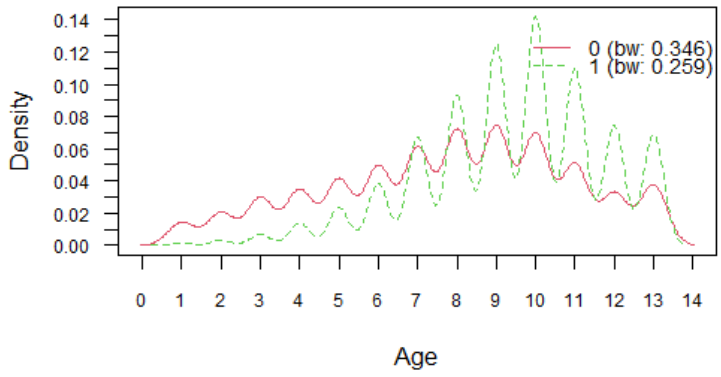


Figure 19: Naive Bayes model Down sample 'Age' Diabetes Indicator dataset

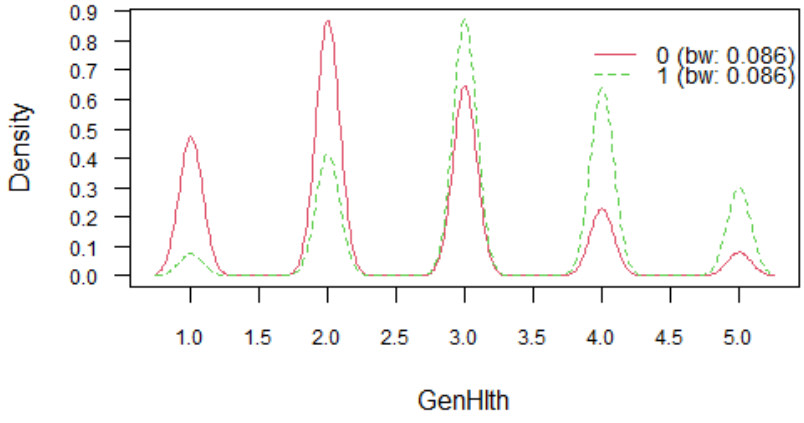


Figure 20: Naive Bayes model 'General health' Down sample Diabetes Indicator dataset

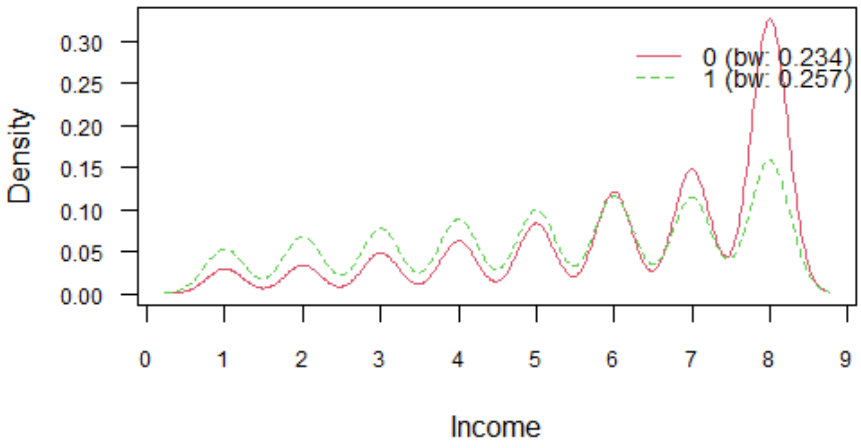


Figure 21: Naive Bayes model 'Income' Down sample Diabetes Indicator dataset

```
> # Predictions----
> p3 <- predict(model2, Valid2_test, type = 'class')
> head(cbind(p3, Valid2[, 2]))
     p3
[1,]  2 1
[2,]  1 1
[3,]  1 1
[4,]  1 1
[5,]  1 1
[6,]  1 1
```

Table 16: Predictions for Naive Bayes model down sample Diabetes Indicator dataset

```
> # Confusion matrix for train data----
> p4 <- predict(model2, Train2[, -2])
> (tab3 <- table(p4, Train2$Diabetes_binary))

p4      0     1
  0 22938 11077
  1  5339 17200
```

Table 17: Confusion matrix for train data Naive Bayes Model down sample Diabetes Indicator dataset

```
> #### This concludes the models accuracy
> 1 - sum(diag(tab3)) / sum(tab3)
[1] 0.2902712
```

Equation 7: Accuracy for Naive Bayes model down sample Diabetes Indicator dataset

```
> # Confusion matrix for test data----
> p5 <- predict(model2, Valid2_test)
> (tab4 <- table(p5, Valid2$Diabetes_binary))

p5     0    1
  0 5762 2766
  1 1307 4303
```

Table 18: Confusion Matrix for test data Naive Bayes model down sample Diabetes Indicator dataset

```
> #### This concludes the models accuracy
> 1 - sum(diag(tab4)) / sum(tab4)
[1] 0.2880888
>
```

Equation 8: Accuracy for test data Naive Bayes model down sample Diabetes Indicator dataset

Smote

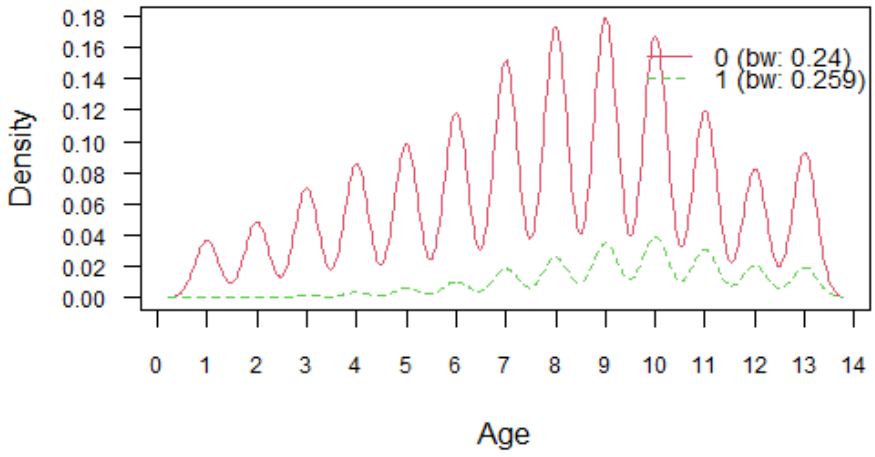


Figure 22: Naive Bayes model smote 'Age' Diabetes Indicator dataset dataset

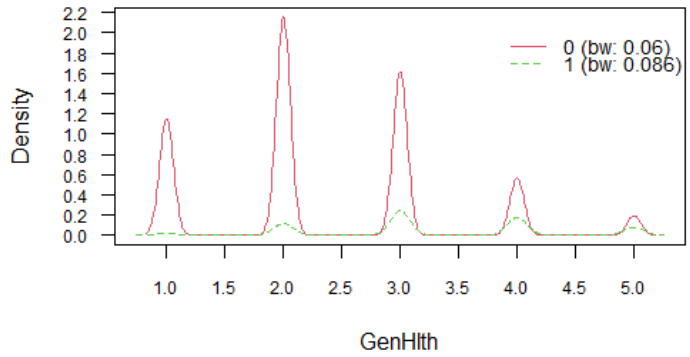


Figure 23: Naive Bayes model smote 'General health' Diabetes Indicator dataset

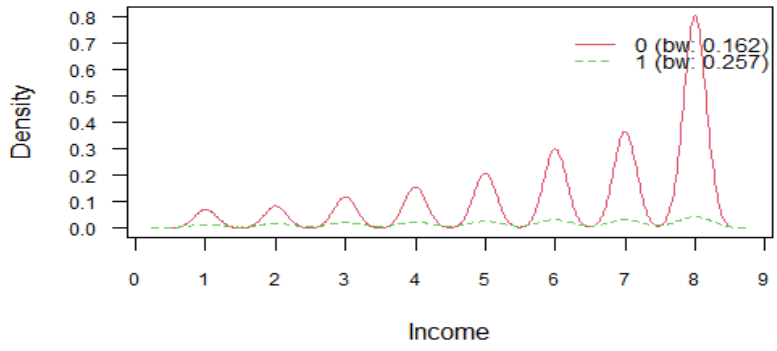


Figure 24: Naive Bayes model Diabetes Indicator dataset 'Income' smote

```
> # Predictions----
> p6 <- predict(model3, Valid3_test, type = 'class')
> head(cbind(p6, Valid3[, 2]))
     p6
[1,]  1 1
[2,]  1 1
[3,]  1 1
[4,]  1 1
[5,]  1 1
[6,]  1 1
```

Table 19: Naive Bayes predictions table smote Diabetes Indicator dataset

```
> # Confusion matrix for train data----
> p7 <- predict(model3, Train3[, -2])
> (tab5 <- table(p7, Train3$Diabetes_binary))

p7       0      1
  0 174407  28037
  1    196    304
```

Table 20: Confusion matrix for train data Naive Bayes Model Diabetes Indicator dataset smote

```
> #### This concludes the models accuracy
> 1 - sum(diag(tab5)) / sum(tab5)
[1] 0.1391172
```

Equation 9: Accuracy for train data Naive Bayes Model Diabetes Indicator dataset smote

```
> # Confusion matrix for test data----
> p8 <- predict(model3, Valid3_test[, -2])
Warning message:
predict.naive_bayes(): only 21 feature(s) out of 22 defined in the naive_bayes object "m
odel3" are used for prediction.

> (tab6 <- table(p8, Valid3$Diabetes_binary))

p8       0     1
  0 43703  6957
  1    28    48
```

Table 21: Confusion matrix for test data Naive Bayes model smote Diabetes Indicator dataset

```
> #### This concludes the models accuracy
> 1 - sum(diag(tab6)) / sum(tab6)
[1] 0.1376734
>
```

Equation 10: Accuracy for test data Naive Bayes model smote Diabetes Indicator dataset

Diabetes Health Indicators Dataset Naive Bayes observations

I observed certain trends among three selected aspects from the Naïve Bayes model: the predictions for the target feature, the training matrix and accuracy and the test matrix and accuracy.

It can be observed from the above analysis that the non-diabetic class has higher density trends than the diabetic class for the regular splitting and smote models. The down-sampling model however performed better and showed a more balanced view of the target class. The objective of using the model for predictions was to see how good the model was at predicting the outcome of certain rows: for example, is row 1 non-diabetic or diabetic?

The Model was tested to see how well it performed on the training and test data. It seemed to perform best when using the training data to create a matrix showing cases for the target features different classes. The accuracy scores were affected by the distribution of cases in the train and test matrix.

10 Years Diabetes Dataset (Naïve Bayes)

The following section contain figures detailing the modelling results using Naïve Bayes on the 10 Years Diabetes Dataset

Naïve Bayes Chart legend (10 Years Diabetes dataset):

- ----- Red Line = Female (0)
- ……. Green Line = Male (1)

Regular Splitting

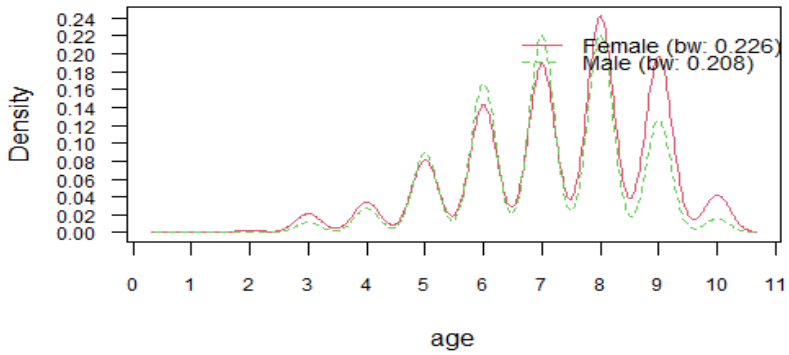


Figure 25: Naive Bayes model 10 Years Diabetes dataset 'Age' regular splitting

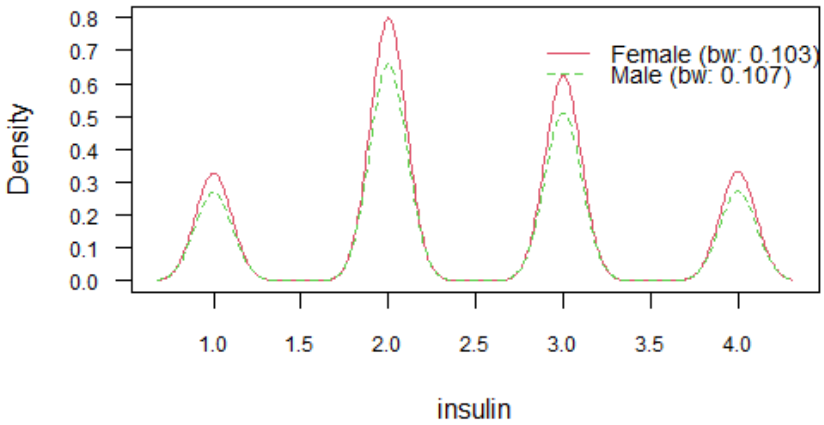


Figure 26: Naive Bayes model 10 Years Diabetes dataset 'Insulin' regular splitting

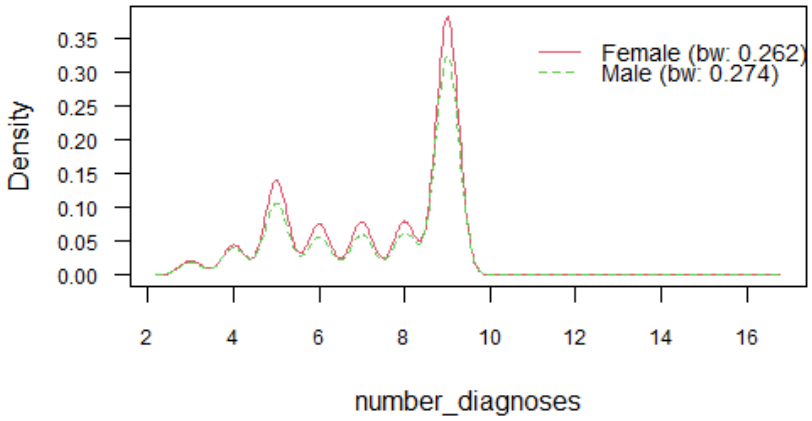


Figure 27: Naive Bayes model 10 Years Diabetes dataset 'Number of diagnosis' regular splitting

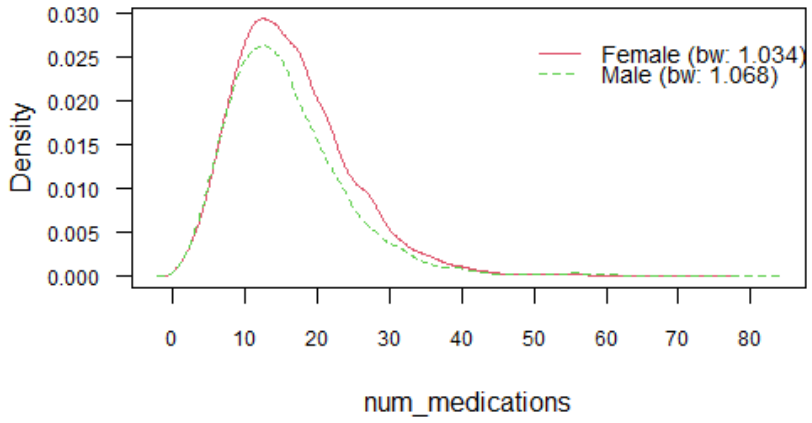


Figure 28: Naive Bayes model 10 Years Diabetes dataset 'Number of Medications' regular splitting

```
> # Predictions----
> p <- predict(model, newdata = test_df, type = 'prob')
> head(cbind(p, test[, 2]))
        Female         Male
[1,] 0.9996201 0.0003799355 1
[2,] 0.9993286 0.0006714443 2
[3,] 0.9993111 0.0006888654 1
[4,] 0.9991457 0.0008542635 1
[5,] 0.9994189 0.0005810702 2
[6,] 0.9966709 0.0033291014 1
>
```

Table 22: Naive Bayes predictions table regular splitting 10 Years diabetes dataset

```
> # Confusion matrix for train data----
> p1 <- predict(model, train_df)
> (tab1 <- table(p1, train$gender))

p1       Female  Male
  Female  11559  9812
  Male        0     9
```

Table 23: Confusion matrix for train data Naive Bayes 10 Years Diabetes Dataset regular splitting

```
> #### This concludes the models accuracy
> 1 - sum(diag(tab1)) / sum(tab1)
[1] 0.4589336
>
```

Equation 11: Accuracy for train data Naive Bayes model regular splitting 10 Years Diabetes dataset

```
> # Confusion matrix for test data----
> p2 <- predict(model, test_df)
> (tab2 <- table(p2, test$gender))

p2       Female Male
  Female   2952 2418
  Male        3    2
```

Table 24: Confusion matrix for test data Naive Bayes 10 Years Diabetes Dataset regular splitting

```
> #### This concludes the models accuracy
> 1 - sum(diag(tab2)) / sum(tab2)
[1] 0.4504186
>
```

Equation 12: Accuracy for test data Naive Bayes model regular splitting 10 Years Diabetes dataset

10 Years Diabetes Dataset Naïve Bayes Observations

It was interesting to observe predictions being carried out to test which rows of the data frame are either non-diabetic or diabetic. It was observed that the accuracy of how the model used the train and test data, to create the confusion matrix by predicting on the target feature, was poor. This is evidenced by the lack of distribution of cases in the confusion matrix. I also observed that the six first rows of the test dataframe have been labelled female due to the probability outcomes displayed for both genders per row. The four charts above show better trend distribution than the other dataset. This is evidenced by the close proximity of the lines for the classes.

6.2.4 Neural Networks modelling

Neural networks are a subset of machine learning and are at the heart of deep learning. Their name and structure are inspired by the human brain mimicking the way biological neurons signal one another. Neural Networks rely on training data to learn and improve accuracy over time.

Set-up of implementations for Neural Network Models: Jupyter Notebook

The Neural Networks analysis was carried out in Jupyter Notebooks using the Kera's packages. In addition to data preparation techniques other implementation aspects

included Early Stopping, Tensor board, Kera's Tuner and a comparison between a static and hyper model neural network.

Early Stopping

The implementation of early stopping as a call back for the neural networks helps stop a neural network during trial runs when all the weights have converged after several epochs run. The main goal of "early stopping" is to stop training when a monitored metric has stopped improving. The main parameters set up for this included: **monitor, val_accuracy, patience, min_delta and mode**. Monitor means quantity to be measured. Patience means the number of epochs with no improvement after which training will stop. This means that the hyper model will move to the next trial run or in the case of a static model, it will cease training altogether. Min delta means minimum change in the monitored quantity to qualify as an improvement. This means that an absolute change of less than min_delta, will count as no improvement. There are different types of modes including min, max and auto. With the max mode, the training will stop once the quantity monitored has stopped increasing.

Figure 29: Early Stopping Implementation

Kera's Tuner

The Kera's tuner was implemented to act on the hyper model and some parameters were set up initially. Important parameters included objective trials to be run and directory linking to project set folder. The project set folder contained the trials run during the algorithm and in each trial a checkpoint was created. I also created a separate project name so the neural network models wouldn't overlap with each other.

The Kera's tuner would (during run time for the model) produce a mini table in the output of the code cell which gives information about the hyper model and its set features. The features I set in the hyper model were units, activation function, learning rate, number of layers and compiling mode.

I also set value conditions and ranges for the units, layers, and activation function. During model runtime the best value for these features will be recorded depending on which trial it occurs from , followed by the value of the previous run. Using the hyper model parameter this will return a model instance as defined in **Error! Reference source not found.**. I set the objective value as the same value used for the monitor hyperparameter in early stopping.

The model ran for 30 trials and the execution per trial was set to 1. The trials were stored at a set project folder which was located depending on the directory location.

```
In [17]: # Defining Keras tuner
         tuner = keras_tuner.RandomSearch(hypermodel=get_model,objective="val_accuracy",
                                          max_trials=30,executions_per_trial=1,overwrite=True,
                                          directory="./benmk/",project_name="diabetes")
```

Figure 30: Kera's Tuner implementation

Tensor Board

Tensor board is a visualisation tool provided with TensorFlow. The callback logs events for TensorBoard. The logs include; Metrics summary plots, Training graph visualisation, weight histograms and sampled profiling.

Implementing this would allow me to view the results from the models after run time. This was also the second call back parameter being used in the model. I set a log directory where the results would be stored after the model finished. I connected tensor board to this directory using a keyword that is usable in Jupyter notebooks. In the log directory a series of folder are available with a unique number each. These contain the train and validation results for each trail run in the model. When using different visualisation modes in Tensorboard it is possible to use different trial folders to observe the results and create comparisons.

Tensorboard

```
In [16]: # Setting tensorboard callback
         tensorboard_callback = tf.keras.callbacks.TensorBoard(log_dir="./logs")
```

Figure 31: Tensor board call back

Tensorboard

```
# Load the extension for tensorboard
%load_ext tensorboard
```

```
The tensorboard extension is already loaded. To reload it, use:
  %reload_ext tensorboard
```

```
# Reload extension for tensorboard
%reload_ext tensorboard
```

```
# Launch tensorboard
%tensorboard --logdir "./logs"
```

Figure 32: Launching Tensorboard

Static Model

Static model

```
In [11]: # Define the keras model
         model = Sequential()
         model.add(Input(shape=(21,)))
         model.add(Dense(21, activation='relu'))
         model.add(Dense(10, activation='relu'))
         model.add(Dense(1, activation='sigmoid'))

In [12]: # Compiling the model
         model.compile(loss = 'binary_crossentropy', optimizer = 'adam', metrics = ['accuracy'])

In [13]: # Converting dataframes to tensor
         trn_new2 = tf.convert_to_tensor(trn_new2)
         y = tf.convert_to_tensor(y)

In [14]: # Fit the keras model on the dataset
         model.fit(trn_new2, y, epochs=1500, batch_size=500, callbacks = [custom_early_stopping], validation_split=0.2)

         Epoch 1/1500
         406/406 [==============================] - 1s 2ms/step - loss: 0.3764 - accuracy: 0.8576 - val_loss: 0.3403 - val_accuracy: 0.8
         597
```

Figure 33: Static Model implementation

Hyper Model

Model function

```
In [15]: # Creating model function
         def get_model(hp):
             model = Sequential()
             model.add(Input(shape=(21,)))
             for i in range(hp.Int("num_layers",1,7)):
                 model.add(Dense(units = hp.Int(f"units_{i}",min_value = 10, max_value = 28, step = 3),
                                 activation = hp.Choice("activation",["relu","tanh"]),))
             model.add(Dense(1, activation='sigmoid'))
             learning_rate = hp.Float("lr",min_value = 1e-4,max_value = 1e-2,sampling="log")
             model.compile(loss = 'binary_crossentropy', optimizer = tf.keras.optimizers.Adam(learning_rate = learning_rate),
                           metrics = ['accuracy'])
             return model
```

Figure 34: Hyper Model implementation

```
In [19]: # Fit the keras model on the dataset
         tuner.search(trn_new2, y, epochs=1500, batch_size=500,
                      callbacks = [custom_early_stopping, tensorboard_callback], validation_split=0.2)

         Trial 30 Complete [00h 00m 42s]
         val_accuracy: 0.8649479746818542

         Best val_accuracy So Far: 0.8654603958129883
         Total elapsed time: 00h 20m 50s
         INFO:tensorflow:Oracle triggered exit
```

Figure 35: Hyper Model execution

Diabetes Health Indicators Dataset (Neural Networks)

Regular Splitting

When comparing the Static Model against the Hyper Model, the networks are analysed in the following areas. Accuracy, loss, time, performance etc. The static model performed well and the accuracy for the epochs during the execution time remained at a similar level. It was noticeable that the loss value slowly decreased per run.

Having defined my hyper model with a little more detail as mentioned in the analysis section, the overall accuracy was proven to be higher showing that the hyper model

performed better than the static model. The difference was around 6% between the static model's accuracy and the hyper model. Also, the fact that the hyper model was running more trials meant it performed better, whereas the hyper model finished more quickly.

Downsampled

In the downsampled model, I also compared the static and hyper model performance. The accuracy of the hyper model performed a lot better when compared to the static model. The static model also had more value loss than the regular splitting static model discussed above. The hyper model finished more quickly because the dataset was a lot smaller than the regular splitting model. The accuracy for the hyper model was less than the regular splitting hyper model which was related to the batch size for the models and the amount of data being trained. Typically, the greater amount of data a model gets during training the better the accuracy will be.

Smote

In the smote model, the static model performed less well when compared to the hyper model in terms of overall performance. The static model did gain a higher accuracy than the smote model. Both models however performed better than the down sampled and regular splitting models due to the higher batch size.

10 Years Diabetes Dataset (Neural Networks)

Regular Splitting

I concluded from the examination of both models, that the hyper model performed better than the static model. The accuracy however (for this analysis) on either model was lower than the accuracy of the static and hyper models in the other dataset. This dataset also had the shortest running network which could lead to poor performance. Overall, on the Neural Networks I concluded that there are many factors that could influence good or bad performance including batch size, dataset size, construction of neural networks and call-back usage.

7.0 Conclusions

During the course of the project, I encountered many issues which had to be overcome and were opportunities to learn in addition to the successful implementation aspects of the project. I have summarised below what was achieved during the project and have provided reflections on the advantages/disadvantages and strengths/limitations of the project approach taken.

Top Performing models for project

Having analysed the results and models in areas such as accuracy and prediction of the project objective (i.e., to predict the incidence of diabetes), the three top performing models for this project objective were:

- Neural Networks Smote Hyper Model,
- Random Forest Down sampled model
- Down sampled Decision Tree model.

These were picked based on model predictions, train and test predictions and overall performance. When I analyse them based on overall model performance, the **Neural Network Smote model** performed the best having achieved an accuracy > 90%. This beats the other model's accuracy performance which were between 70-75%.

This also illustrates that the use of AI over normal machine learning models to predict the chances of someone being diabetic is superior to normal machine learning models (such as Decision Trees). It was also worth noting that machine learning models sometimes perform well on smaller batches of data. For example, the down sampled models performed better than the regular splitting and smote models (exc. Neural Networks).

Datasets

During the modelling phase two binary features were analysed, with one from each dataset, given that separate analysis was performed on each. From the Neural network analysis, the Diabetes Health Indicators dataset proved to be better, because of a larger amount of data. The target feature in the Diabetes Health Indicators dataset was 'Diabetes Binary' and in the 10 Years diabetes dataset it was 'Gender'. Both had their merits as possible ways of predicting the gender and possible chance of someone being diabetic.

Analysis

The statistical analysis carried out in the project was helpful. In particular the chi-square tests and centre of mean analysis. Understanding the centre of the mean was helpful to identify where the data peaked at class values even though it didn't help explain the distribution.

Also, the chi-square could have been used on all the features in the datasets to help check for independence. This is particularly important when conducting algorithms like Naïve Bayes which only tend to perform well if features in a dataset don't depend on each other. The other type of distribution tests I would have conducted if more time was allowed included checking normal distribution which can be provided by a package available in R.

Project approach – Advantages

- The project approach proved to be capable of using classification models to analyse outcomes of binary problems.
- It was beneficial to see how well the project approach highlighted relative performance depending on the models created and the data provided.
- AI was proven to be very useful in completing binary classification tasks.

Project approach – Disadvantages

- The project was constrained by time, and I would like to have explored the chart outcomes a little bit more in the results section to see what insights could be gained from them.
- Further statistical analysis would have been beneficial in deciding what binary classification models could have been picked for the task at the beginning of the project.

Strengths

- The project demonstrated the use of models and tools in solving real world health problems and understanding how certain diseases could be controlled. Statistic tests such as the chi-square test are very popular in the healthcare industry.

Limitations

The project doesn't capture transitional disease data such as pre-diabetes as it was only designed to handle binary outcomes and whether someone is diagnosed with a disease or not. This however would be an interesting area to look in more detail.

8.0 Further Development or Research

With additional time and resources, which direction would this project take?

Further Development:

Looking back at the exploratory and statistical analysis I could search for other possible methods that would be appropriate for my area of study. I could try to make sure they are only suitable for binary classification problems as this the focus topic.

With the statistical analysis, I would like to have more visuals created as I feel it would help with understanding the values and help highlight their importance. This would be beneficial for a lay audience. I would also like to have included metrics that would support the machine learning models such as: Accuracy, F1-Score, Precision, Sensitivity, Specificity and the classification report.

As some of the machine learning took quite a bit of time to run (such as Neural network), I felt it would be interesting to look at alternative platforms to run the Jupyter Notebooks such as Google Collab. The cloud provider may have saved some time for running the models.

For the project two platforms were primarily used which were RStudio and Jupyter Notebooks. There were other platforms such as Spyder that could be considered in future work .

Related Research Papers

I have referenced a number of research papers below which contain ideas that could contribute to improving certain parts of the project such as the pre-liminary analysis and modelling section.

Pre-liminary analysis

- In a paper from Elsevier "Evaluating the Incidence and Risk Factors Associated With Mild and Severe Hypoglycaemia in Insulin-Treated Type 2 Diabetes", [3] the data that was collected was part of a prospective, noninterventional, epidemiological study. The data was collected from 1 hospital centre. The statistical statistics [3] was carried out using descriptive stats, mean, Standard Deviation, Median and Interquartile range. Negative binomial regression analysis was used to access the predictors of the frequency of non-severe and severe hypoglycaemic events. SPSS statistics was also used to calculate results.

Modelling

- In a chapter from the book "Expert Systems with Applications" [1] the authors outlined in their approach similar classifiers that which I proposed and which I would use in my future modelling stage. These included the KNN classifier. They apply different methods to those which I used such as; PNNR, LMPNN and MVMCNN. Data imputation methods were applied to deal with the missing values, and they used a train test split on the dataset with a 5-10 FCV. The datasets used in the study was sourced from [1] UCI data Machinery and Diabetes Type from the Data World repository. To solve the missing data problem, they used Linear regression and correlation techniques to remove the values.
- In a chapter from the book "Computers in Biology and Medicine" [2] a variety of classification models were implemented and these included SVM, KNN, Bagging and Stacking. They collected their data from King Fahad University Hospital in Saudi Arabia [2]. For visualisation of outliers, they applied boxplots, whereas I used Histogram charts to check for outliers.
- In a paper from Elsevier "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," [6] the dataset collected was the Pima Indian dataset, it contained around 9 features, and one output column for predicting diabetes. The architectures used in this study included: Logistic Regression, Naïve Bayes, Random Forest, and Proposed Ensemble soft voting classifier [6].

9.0 References

Research Articles:

[1] S. Suyanto, S. Meliana, T. Wahyuningrum and S. Khomsah," A new nearest neighbor-based framework for diabetes detection," Indonesia, Elsevier, 2022.

[2] M. Gollapalli, A. Alansari, H. Alkhorasani, M. Alsubaii, R. Sakloua, R. Alzahrani, M. Al-Hariri, M. Alfares, D. Alkhafaji, R. Al Argan and W. Albaker," A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM," Saudi Arabia, Elsevier, 1982.

[3] A. Chantzaras and J. Yfantopoulos," Evaluating the Incidence and Risk Factors Associated With Mild and Severe Hypoglycaemia in Insulin-Treated Type 2 Diabetes," Greece, Elsevier, 2021.

[4] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," India, 2013.

State of the art:

[5] Z. Stojanoski, M. Kalendar and H. Gjoreski, "Comparative Analysis of Machine Learning Models for Diabetes Prediction," *Proceedings of International Conference on Applied Innovation in IT.*, vol. 11, no. 1, pp. 75-80, Mar. 2023.

[6] J. Pradeep Kandhasamy and S. Balamurali, "Performance Analysis of Classifier Models to Predict Diabetes Mellitus," *Procedia Computer Science.*, vol. 1, no. 47, pp. 45-51, 2015.

Dataset sources:

[7] Kaggle, "10 Years Dataset," 2022. [Online]. Available: 10 Years Diabetes Dataset | Kaggle [Accessed on: Dec. 18, 2022].

[8] Kaggle, "Diabetes Health Indicators Dataset," 2022. [Online]. Available: Diabetes Health Indicators Dataset | Kaggle [Accessed on: Dec. 17, 2022].

10.0 Appendices

National College of Ireland

Project Proposal

Title: An analysis of the causes and prevention of diabetes

Date: 30th October 2022

Programme Name: BSc. Honours in Data Science
Specialisation: BSHDS4
Academic Year: 2022/2023
Student Name: Benjamin Kelly

Student Number: x19370681
Student Email: x19370681@student.ncirl.ie

Contents

1.0 Project Objectives

The proposed project is focused on an analysis of global healthcare data in the area of Diabetes. It will explore the incidence of diabetes and predict the main changes that can be made to healthcare services to deal with the growing incidence of diabetes in the future.

Diabetes is a medical condition that is characterised by high blood sugar levels. One in ten people around the world are living with diabetes which is a lifelong condition and one of the top ten leading causes of deaths globally. The project objectives include the identification of trends in diabetes internationally and the factors are causing them to occur, it will explore where diabetes is growing and what methods can be used to prevent and manage the disease. The disease is split into Type 1 and Type 2 diabetes with the latter being more of a medical condition as opposed to a disease. Specifically, the project will explore data the following three areas to make predictions and draw conclusions:

- **Causes of diabetes:** The cause of Type 1 diabetes is unknown, but it mainly happens when antibodies attack the pancreatic cells in the pancreas producing insulin. Genetic inheritance in families can be a major issue since children who inherit the same genes are likely to pick it up at some point. Pancreatic diseases may also cause Type 1 diabetes in some people. Type 2 diabetes is caused by cells not taking the insulin produced from the pancreas or the pancreas doesn't produce enough insulin. Diet and obesity are a major cause of Type 2 diabetes which varies across different countries globally. It is also a disease that can be inherited through families but affects people who are 45 or older. People can also be more at risk through poor lifestyle choices and a lack of exercise. The above are explored in the project to identify high risk factors.
- **Incidence of diabetes on a global scale:** Diabetes Type 1 is a disease that ranges in different countries depending on the prevalence of the disease and the incidence rates. Countries that have a long history of the disease include Scandinavia and Italy. In the USA, Caucasians are more likely to pick up the disease than African Americans or Hispanic Americans. In Asian countries remain relatively low but China suffers from Type 2 diabetes and is one of the countries in the world with a high incident rate. The incidence will be explored through exploratory analysis and statistical analysis.
- **Preventing the disease:** Type 2 diabetes can be prevented if an individual chooses to adjust their lifestyle by; losing weight, being more active, quitting smoking and making changes to their diet. Type 1 diabetes unlike Type 2 can't be prevented as easily by adjusting diet and lifestyle. However, a person may be able to prevent further damage to other organs by managing the disease.

2.0 Background – why this project

In my third-year internship I worked on public health and social care data (NCS-National Childcare Scheme) with POBAL in Dublin. This gave me an insight into how data can be used to inform public policy in health and social care. As a result, I wanted to do a project in the healthcare area to help understand how data can assist the government agencies in controlling a big disease area such as diabetes. I have an uncle and cousin who suffer from Type 1 diabetes which has given me a personal interest as I have discussed with them how

they manage and measure the food they ingest to control blood sugar levels. Type 1 diabetes in a lifelong condition that must be controlled regularly and have insulin shots taken daily. The other condition Type 2 is not as severe but can be controlled through management of lifestyle such as exercise, diet, smoking, weight, sleep, stress etc.

I believe the project is a new area of focus for a college student project in NCI, yet more than half a billion adults are now living with diabetes worldwide and it is a disease which puts a growing strain on the healthcare system in Ireland and around the world. Healthcare is also an area I am interested in exploring in the project because there is a lot of data research is being constantly carried out.

Meeting the objectives

Through extensive dataset research, **datasets that highlight the causes** such as diet, lifestyle, etc. for Type 2 Diabetes will be considered to help me identify how each cause factor effects the growth of the disease. For Type 1 diabetes health issues, family history and genetic information will be examined to identify trends in the disease.

Similar research will be carried out into the **global incidence levels of diabetes** and examine the comparisons between race, country and gender data.

I will carry out predictive analysis of the available data to assist in **prevention of the disease** by trying to **predict the likely reduction in incidence** or risk if people were to carry out some adjustments to their lifestyle. This will be based on a deep analysis of diabetes and lifestyle data to provide a clearer understanding of the links between the incidence of diabetes and lifestyle and other related factors.

3.0 State of the Art – similar analysis by others

Given the growing prevalence of the disease, there is a considerable amount of data available in the area of diabetes research around the world. As part of this project research, I have identified to date similar studies which analyse the links between incidence of diabetes and lifestyle.

Examples include the NHS (Nurses' Health Study) analysing the risk factors for Type 2 diabetes including Obesity, physical activity, smoking and sleep. This identified that obesity was a major cause of the diabetes and confirmed that where an individual remains obese their likelihood of getting type 2 diabetes increases. It also had conclusions relating to fat as a cause of type 2 diabetes and the impact of fibre in the diet.

Another example paper is the 'New England Journal of Medicine' who carried out lifestyle trials on patients that were at high risk from diabetes. Most of the diabetes research identified to date is focused on clinical data measuring the impact of specific treatments, cohorts of patients, symptoms or treatment areas. I believe that this project is different because I will attempt to pull together data from diverse geographic areas and draw conclusions.

I hope that my conclusions can be used to predict the areas where Ireland could improve services by identifying the lifestyle, diet and healthcare services key data attributes which have most impact on meeting the needs of the patient population. Using this data, I will draw conclusions and help predict what changes will best close the gap for services in Ireland and elsewhere.

What makes the project stand out

The proposed project covers a wide range of data on prevalence, prevention and causes internationally whereas much of the research is on clinical treatments. I plan to make strong use of visuals to help offer a comparison between prevalence, prevention and causes in different countries. In this way it would be beneficial to readers who wish to understand international trends and draw conclusions for health service planning in the future. Being able to predict how the disease can be controlled in the future is an important aspect I haven't seen in many papers, and this will be explored in the project. I feel that this project could be useful for policy makers like the HSE, the department of health and advocacy groups such as diabetes Ireland. They need this data to help them decide how best to use limited resources in the future ,ideally benchmarked on "best practices" outside Ireland.

I believe this is a **worthwhile project** because, with an ageing population and pandemics, healthcare services around the world are under pressure and this pressure will continue to grow. Policy makers, healthcare funders, and healthcare service providers will all need this information to support decision making. With over half a billion people worldwide living with diabetes and another half a billion at high risk of type 2 diabetes due to lifestyle and related reasons, this represents a huge drain on resources into the future.

The proposed project will be **challenging** because, while a lot of data is available in the area of diabetes, much of it is treatment focused in very specific areas. As such sourcing and combining suitable datasets and preparing the data to compare, analyse and make predictions will be challenging.

4.0 Data required for the project

As part of the project research to date I have identified data from several dataset sources including Kaggle and data.gov.ie. In the project I will identify other suitable datasets from other sources including OECD, UCI and CSO to gather the relevant information on prevalence of diabetes.

The dataset requirement for the project will contain information on diabetes incidence and causes, it will include global data covering country, race, and gender as part of the global analysis and comparison work. Various datasets identified to date include yearly diabetes, health indicators datasets and hospital case data for diabetes as medical history is an important part of the project.

Lifestyle datasets covering areas such as obesity have been identified addition to health indicators datasets. These are important datasets as they help link lifestyle and diet information about those with diabetes as outlined in the first objective of the project. This dataset will help predict whether a patient is likely to be diagnosed with diabetes. The lifestyle factors data will include age/exercise/diet /smoking/sugar intake.

Health and hospital data also provides information such as the race, gender and age of patients and their medical details. The health services data will include diabetes incidence /how many hospitalised/for how long/ with what symptoms/ how much spent on treatment by age group/gender/country and also demographic and geographic data. Other possible dataset information that will be considered for further research will be in the areas of "pre-diabetes" which will assist with prediction as diabetes is often missed before someone is diagnosed with it.

Accessing and compiling the necessary data

The data will be accessed online from various data sources including Kaggle, data.gov, Stats, OECD and many more. They will be mainly downloaded through a .csv format as this format is easiest to manage in preparation for programming platforms.

The objective is to compile a number of datasets which may be used together assuming they have common variables. These will be accessed via excel following download for detailed analysis and identification of key columns.

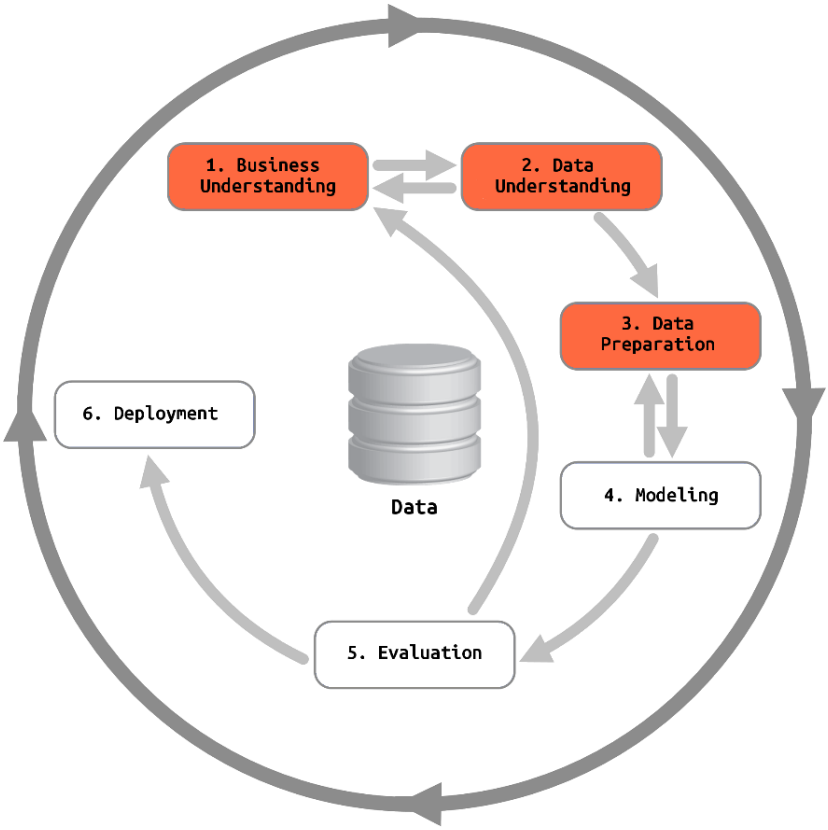
Pre-processing will be a critical next step in the process once I have done doing the exploratory analysis of the dataset to understand the variables and correlation. Pre-processing would involve; getting rid of null values/useless data, creating functions to clean the data, encoding the data (possibly depending on the task), checking for class imbalances etc.

5.0 Methodology & Analysis

I plan to use a selection of tools to prepare and analyse this data as part of the project and these could include visualisation, statistical analysis, machine learning modelling etc to predict how policymakers might best use healthcare resources in the future. The project will follow an established analysis methodology (CRISP DM) and will employ technologies such as Python and the R programming language.

Methodology to be used and why

For my methodology I am using the CRISP -DM methodology. (Cross Industry Standard Process for Data Mining). This methodology is more suitable for the project than for example KDD method as the problem can be refined as we learn from the modelling and there is a focus on defining the business problem.



By using the CRISP DM approach, the project is broken out into the stages outlined in the image opposite. As we implement the modelling, we can evaluate results and refine the understanding of the data and repeat the exercise.

It is a methodology composed of six different stages**: Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment.** The order of the stages is flexible and often it is necessary to move back and forth between the different stages. After the solution is deployed, the data mining process continues, and the lessons learned during the process can trigger new, more focused business questions allowing the next data mining processes to benefit from the experience of previous ones.

Business Understanding: The objectives for the project are as laid out in Section 1.0. I will also be able to refine the diabetes analysis objectives as I move through the project. I wish to analyse the subject of diabetes and draw conclusions connecting the causes the global incidence and how it may be prevented. I have found an initial range of datasets that can be possibly used in the project from data sources and hubs. I am expecting from a data perspective that by carrying out various analysis models I can link them to my goals and

form conclusions. These conclusions will be checked to see if they have satisfied the objectives of the project.

I intend that a range of technology tools will be used during the project which include language, platforms, and packages. The tools will vary based on the tasks to be carried out. For my visualisation task I plan to use the PowerBI tool which I found useful during my work placement. For statistical objectives I intend to use tools such as Excel, Spyder, RStudio and IBM SPSS. From a programming perspective I will use the Python or R programming languages and the platforms used here could be RStudio, Jupyter Notebook or Spyder. Planning and brain maps will be carried out in draw.io for thinking purposes.

Data Understanding: When the data has been sourced for it will be loaded into excel as csv files for analysis purposes. The data will be kept in this format so that it can be imported into Spyder, Jupyter Notebook or RStudio for coding purposes. It will be put into excel first to carry out specific tasks such as identifying quantity of data, variable inspection and checking for null data. Following this, the data will be imported to PowerBI for visualisation purposes and finding correlations between variables. After this a statistical analysis will be carried out in either Excel, RStudio or Spyder to help understand relationships and statistical points in the data depending on what task is being solved.

Data Preparation: Once the data has been explored and understood, getting it prepared for the modelling will be the next phase of the project. From the datasets I have explored, I will be picking the ones that will be of most useful to the project. Subsets may be used from several datasets depending on the relevance of variables. Certain statistical tests that involve around preparation as opposed to exploration such as checking for outliers within variables will be conducted to see if they would pose a problem to the training and testing stage later in the project. If variables are similar between datasets these may be combined in one dataset. Variables may be encoded depending on whether the task is classification or regression related.

Modelling: The modelling task that I have in mind for this section is a supervised regression machine learning problem. This is due to the type of data collected which is presented in numerical format. The algorithms that may be used here would be Linear/Logistic regression, Support Vectors Regression, Random Forest, Decision Trees, and Neural Networks. The split will be mainly train and test, but validation may be used for optimising the models' parameters. Parameters may be adjusted during model tests to deliver better results.

Evaluation: When the models have been carried out and completed the results will be checked for each. At this point, the objective will be to see which models performed well and which didn't. If results seem to be too good to be true then certain measures will be taken such as checking the preparation steps taken earlier. Otherwise, certain models will be picked if they have satisfied the objectives of the project. If other relationships are identified based on patterns between variables at a later stage of the project, then the exploratory stage will be expanded, and I will repeat the steps as per the CRISP-DM model.

Deployment: Checks will be run regularly before the models take in the data. Some data may not be ready for models so once it is ready, they will be put into the modelling stage while other models will be running or have run at that time. Figuring out when they are ready to be put into the report is important. A report detailing the work completed and results collected will be summarised in IEEE word document. This will show briefly the steps taken, the goals achieved during the project, the methodology analysis, what the results show, what could have been done differently and what viewers will expect from the document.

Approach to be taken in analysis

In the **exploratory stage**, I will be finding charts within Power BI that will represent the data I'm trying to show. Visuals can range from bar charts to wheels etc. I will be noting how visual each correlation is so viewers can understand the picture in context. Different variables will be picked in pairs for different visuals created. In Excel, some visualisation tasks will be carried out using the tools available in it.

In **the statistical stage,** I will be finding what regression measures are best suitable for understanding the statistics of the data. Some of these may be present on charts and plots to help understand their value and meaning. Any results will be carried forward in the project toward the modelling phase of the project.

In the **preparation stage**, I will be considering the correlations among variables from the earlier stages and including them in the processing stage. These will be used for the modelling as they have a positive impact on the models.

In the **evaluation stage**, the results that I feel made a positive impact on the objectives mentioned at the beginning will be carried forward for final conclusions. Statistical results and visuals that represent the use of variables in models will also be used for explanation purposes. Other results from other models will be kept as extra for background work done during the modelling phase.

In the deployment stage, analysis of other variables that may be possible candidates for models will be considered if they are likely to have a good correlation with other variables. Report will be reviewed regularly for anything missed.

Breaking down the project work into project tasks, activities and milestones

Using help of the CRISP-DM methodology I going to break down the work into steps. Each project task, be it the exploratory or statistical analysis, will be given a time frame to complete it and will be recorded in the project plan. Some tasks may be one at the same time if multitasking allows. From taking in the data to getting results and a report this will be explained further in section 7.0.

6.0 Technical Details

Technical development to be carried out as part of this project.

Applications to be used

I plan to use a wide variety of applications during the project. A few of the obvious ones that will be used include Jupyter Notebook, Spyder, RStudio, PowerBI, Excel and possibly IBM SPSS statistics.

PowerBI is one of the main exploratory tools for the exploratory phase of the project and will be important when looking for correlations among variables. Other tools that will aid in visual tasks will be excel, Spyder, RStudio and Jupyter notebook. These will though, mostly focus on visuals. For practical implementation of coding, Jupyter Notebook, RStudio and Spyder will be the primary platforms for intensive coding during the project and where most of the analysis will take place. The above applications will also be capable of carrying out a range of statistics as well.

During the project, I don't intend to use any cloud software, or VirtualBox as the I will be using my main computer for the work. This could however change depending on the requirements of the data taken in.

I will be using the Python language and R language during coding phase. The R language will be widely used for statistics and the python language will be used a lot for the deep learning phase.

Algorithms or approaches under consideration

The algorithms I propose to use for the machine learning stage are Logistic/linear regression, Random Forest, Decision Trees, Neural Networks, and support vector machines. Other methods may be considered for the modelling phase.

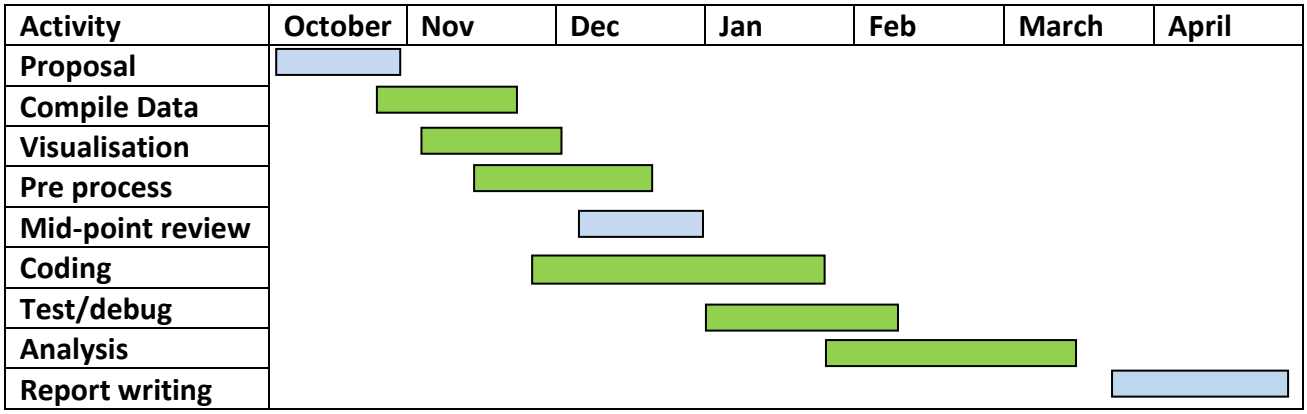
Accuracy metrics and measures will be used to analyse the performance of the algorithms. Some of these include using; RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) and Adjusted R2.

Linear regression is a very efficient algorithm and would be good from a correlation point of view using some metrics. Random forest and decision trees would be good for large quantities of data and getting different data samples which would help improve the performance of the algorithm. Neural Networks is widely used for predictive tasks as it can predict future trends which would be helpful on aspects of the diabetes project objectives.

During the project, a prediction engine may be used with the help of neural networks to answer the prediction question mentioned in section 2.0.

7.0 Project Plan

High Level Project Gannt Chart: The following is a high-level overview of the key project stages set out over semester one and two of my final year.



Detailed Activity Table

The following table will list in detail the activities which will be carried out as part of the project. This will also be updated during the project with the actual activities and dates as they are completed for inclusion in the project documentation.

Note MB = Assigned Academic Supervisor Michael Bradford.

Task Name	Who	Start Date	End Date	Duration
Concept for project	BK	19/09/2022	31/10/2022	43 days
Sourcing data	BK	30/10/2022	9/11/2022	11 days
Exploratory data analysis	BK	9/11/2022	23/11/2022	14 days
Data pre-processing	BK	27/11/2022	5/12/2022	9 days
Reflective Journal 1	BK	31/10/2022	31/10/22	½ day
Meeting 1 - Project discussion with academic Supervisor	BK and MB	8/11/2022	8/11/2022	30 minutes
Revisions to project idea	BK	31/10/2022	9/11/2022	8 days
Detailed activity / task planning	BK	9/11/2022	17/11/2022	8 days
Statistics analysis	BK	23/11/2022	27/11/2022	4 days
Initial analysis Power BI	BK	9/11/2022	23/11/2022	14 days
Visualisation graphics	BK	9/11/2022	30/11/2022	21 days
Reflective journal 2	BK	30/11/2022	30/11/2022	½ day
Research ….	BK	22/10/2022	9/11/2022	18 days
Finalise Ethics / Project Proposal paper	BK and MB	5/11/2022	20/12/2022	45 days
Meeting 2 academic supervisor	BK and MB	TBC	TBC	TBC
Data preparation	BK	27/11/2022	5/12/2022	8 days
Data modelling and Analysis	BK	2/12/2022	TBC	TBC
Reflective journal 3	BK	31/12/2022	31/12/22	½ day
Meeting 3 academic supervisor	BK and MB	TBC	TBC	TBC
Coding……	BK	23/11/2022	TBC	TBC
December mid project review draft report	BK and MB	9/12/2022	22/12/2022	13 days
Further analysis	BK	TBC	TBC	TBC
Debugging and testing	BK	TBC	TBC	TBC

Conclusions	BK	TBC	TBC	TBC
Draft project report	BK	TBC	TBC	TBC
Draft project report review with supervisor	BK and MB	TBC	TBC	TBC
Report writing	BK	TBC	TBC	TBC

7.1. Ethics Approval Application

OCTOBER 2022 Ethics Forms

National College of Ireland

DECLARATION OF ETHICS CONSIDERATION

School of Computing

Student Name: …Benjamin Kelly……

Student ID: …x19370681……..……

Programme … BSHDS4…… Year: …4……..

Module: …Computing Project…….……

Project Title: An analysis of the causes and prevention of diabetes internationally and lessons for Irish healthcare

Please circle (or highlight) as appropriate

This project involves human participants	No

Introduction

Secondary data refers to data that is collected by someone other than the current researcher. Common sources of secondary data for social science include censuses, information collected by government departments, organizational records and data originally collected for other research purposes. Primary data, by contrast, is collected by the investigator conducting the research.

A project that does not involve human participants requires ONLY completion of Declaration of Ethics Consideration Form and submission of the form on module's Moodle page

A project that involves human participants requires ethical clearance and an Ethics Application Form must be submitted through the module's Moodle page. Please refer to and ensure compliance with the ethical principles stated in NCI Ethics Form available on the Moodle page.

The following decision table will assist you in deciding if you have to complete the Declaration of Ethics Consideration Form or/and the Ethics Application Form.

Public Data	Y	Y	Y	Y	N	N	N	N

Private Data	Y	Y	N	N	Y	Y	N	N
Human Participants	Y	N	Y	N	Y	N	Y	N
Declaration of Ethics Consideration Form	x	X	x	X	X	X	x	
Ethics Application Form	X		X		X		X	

Please circle (or highlight) as appropriate

The project makes use of secondary dataset(s) created by the researcher	Yes
The project makes use of public secondary dataset(s)	Yes
The project makes use of non-public secondary dataset(s)	No
Approval letter from non-public secondary dataset(s) owner received	N/A

Sources of Data:

It is students' responsibility to ensure that they have the correct permissions/authorizations to use any data in a study. Projects that make use of data that does not have authorization to be used, will not be graded for that portion of the study that makes use of such data.

Public Data

A project that makes use of public secondary dataset(s) does not need ethics permission, but needs a letter/email from the copyright holder regarding potential use.

Some websites and data sources allow their data sets to be used under certain conditions. In these cases, a letter/email from the copyright holder is NOT necessary, but the researcher should cite the source of this permission and indicate under what conditions the data are allowed to be used. See Appendix I for examples of permissions granted by Fingal Open Data, and Eurostat website.

Where websites or data sources indicate that they do not grant permission for data to be used, you will still need a letter/email from the copyright holder. For example, see Appendix II for an example from the Journal of Statistics Education.

Private Data

A project that makes use of non-public (private) secondary dataset(s) must receive data usage permission from School of Computing.

An approval letter/email from the owner (e.g. institution, company, etc.) of the non-public secondary dataset must be attached to the Declaration of Ethics Consideration. The letter/email must confirm that the dataset is anonymised and permission for data processing, analysis and public dissemination is granted.

Evidence for use of secondary dataset(s)

Include dataset(s) owner letter/email or cite the source for usage permission

Not applicable

CHECKLIST

Non-public/private secondary dataset(s) -Owner letter/email is attached to this form *OR* Citation and link to the web site where permission is granted – provided in this form	N/A N/A

ETHICS CLEARANCE GUIDELINES WHEN HUMAN PARTICIPANTS ARE INVOLVED

The Ethics Application Form must be submitted on Moodle for approval prior to conducting the work.

Considerations in data collection

Participants will not be identified, directly or through identifiers linked to the subjects in any reports produced by the study

Responses will not place the participants at risk of professional liability or be damaging to the participants' financial standing, employability or reputation

No confidential data will be used for personal advantage or that of a third party

Informed consent

Consent to participate in the study has been given freely by the participants

participants have the capacity to understand the project goals.

Participants have been given information sheets that are understandable

Likely benefits of the project itself have been explained to potential participants

Risks and benefits of the project have been explained to potential participants

Participants have been assured they will not suffer physical stress or discomfort or psychological or mental stress

The participant has been assured s/he may withdraw at any time from the study without loss of benefit or penalty

Special care has been taken where participants are unable to consent for themselves (e.g children under the age of 18, elders with age 85+, people with intellectual or learning disability, individuals or groups receiving help through the voluntary sector, those in a subordinate position to the researcher, groups who do not understand the consent and research process)

Participants have been informed of potential conflict of interest issues

The onus is on the researcher to inform participants if deception methods have to be used in a line of research

I have read, understood, and will adhere to the ethical principles described above in the conduct of the project work.

Signature: ……Benjamin Kelly……

Date: …31 Oct 2022……

Fingal Open Data: http://data.fingal.ie/About

Licence

Citizens are free to access and use this data as they wish, free of charge, in accordance with the Creative Commons Attribution 4.0 International License (CC-BY).

Note: From November 2010 to July 2015, data on Fingal Open Data was published in accordance with the PSI general licence.

Use of any published data is subject to Data Protection legislation.

Licence Statement

Under the CC-BY Licence, users must acknowledge the source of the Information in their product or application by including or linking to this attribution statement: "Contains Fingal County Council Data licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence".

Multiple Attributions

If using data from several Information Providers and listing multiple attributions is not practical in a product or application, users may include a URI or hyperlink to a resource that contains the required attribution statements.

Eurostat: https://ec.europa.eu/eurostat/about/policies/copyright

COPYRIGHT NOTICE AND FREE RE-USE OF DATA

Eurostat has a policy of encouraging free re-use of its data, both for non-commercial and commercial purposes. All statistical data, metadata, content of web pages or other dissemination tools, official publications and other documents published on its website, with the exceptions listed below, can be reused without any payment or written licence provided that:

the source is indicated as Eurostat

when re-use involves modifications to the data or text, this must be stated clearly to the end user of the information

Journal of Statistics Education: http://jse.amstat.org/jse_users.htm

JSE Copyright and Usage Policy

Unlike other American Statistical Association journals, the Journal of Statistics Education (JSE) does not require authors to transfer copyright for the published material to JSE. Authors maintain copyright of published material. Because copyright is not transferred from the author, permission to use materials published by JSE remains with the author. Therefore, to use published material from a JSE article the requesting person must get approval from the author.

November 2022 Ethics Forms

National College of Ireland

DECLARATION OF ETHICS CONSIDERATION

School of Computing

Student Name: …Benjamin Kelly……

Student ID: …x19370681……..……

Programme … BSHDS4…… Year: …4……..

Module: …Computing Project……..……

Project Title: An analysis of the causes and prevention of diabetes internationally and lessons for Irish healthcare

Please circle (or highlight) as appropriate

This project involves human participants	No

Introduction

Secondary data refers to data that is collected by someone other than the current researcher. Common sources of secondary data for social science include censuses, information collected by government departments, organizational records and data originally collected for other research purposes. Primary data, by contrast, is collected by the investigator conducting the research.

A project that does not involve human participants requires ONLY completion of Declaration of Ethics Consideration Form and submission of the form on module's Moodle page

A project that involves human participants requires ethical clearance and an Ethics Application Form must be submitted through the module's Moodle page. Please refer to and ensure compliance with the ethical principles stated in NCI Ethics Form available on the Moodle page.

The following decision table will assist you in deciding if you have to complete the Declaration of Ethics Consideration Form or/and the Ethics Application Form.

Public Data	Y	Y	Y	Y	N	N	N	N
Private Data	Y	Y	N	N	Y	Y	N	N
Human Participants	Y	N	Y	N	Y	N	Y	N
Declaration of Ethics Consideration Form	x	X	x	X	X	X	x	
Ethics Application Form	X		X		X		X	

Please circle (or highlight) as appropriate

The project makes use of secondary dataset(s) created by the researcher	Yes
The project makes use of public secondary dataset(s)	Yes
The project makes use of non-public secondary dataset(s)	No
Approval letter from non-public secondary dataset(s) owner received	N/A

Sources of Data:

It is students' responsibility to ensure that they have the correct permissions/authorizations to use any data in a study. Projects that make use of data that does not have authorization to be used, will not be graded for that portion of the study that makes use of such data.

Public Data

A project that makes use of public secondary dataset(s) does not need ethics permission, but needs a letter/email from the copyright holder regarding potential use.

Some websites and data sources allow their data sets to be used under certain conditions. In these cases, a letter/email from the copyright holder is NOT necessary, but the researcher should cite the source of this permission and indicate under what conditions the data are allowed to be used. See Appendix I for examples of permissions granted by Fingal Open Data, and Eurostat website.

Where websites or data sources indicate that they do not grant permission for data to be used, you will still need a letter/email from the copyright holder. For example, see Appendix II for an example from the Journal of Statistics Education.

Private Data

A project that makes use of non-public (private) secondary dataset(s) must receive data usage permission from School of Computing.

An approval letter/email from the owner (e.g. institution, company, etc.) of the non-public secondary dataset must be attached to the Declaration of Ethics Consideration. The letter/email must confirm that the dataset is anonymised and permission for data processing, analysis and public dissemination is granted.

Evidence for use of secondary dataset(s)

Include dataset(s) owner letter/email or cite the source for usage permission

Not applicable

CHECKLIST

Non-public/private secondary dataset(s) -Owner letter/email is attached to this form *OR* Citation and link to the web site where permission is granted – provided in this form	N/A N/A

The Ethics Application Form must be submitted on Moodle for approval prior to conducting the work.

Considerations in data collection

Participants will not be identified, directly or through identifiers linked to the subjects in any reports produced by the study

Responses will not place the participants at risk of professional liability or be damaging to the participants' financial standing, employability or reputation

No confidential data will be used for personal advantage or that of a third party

Informed consent

Consent to participate in the study has been given freely by the participants

participants have the capacity to understand the project goals.

Participants have been given information sheets that are understandable

Likely benefits of the project itself have been explained to potential participants

Risks and benefits of the project have been explained to potential participants

Participants have been assured they will not suffer physical stress or discomfort or psychological or mental stress

The participant has been assured s/he may withdraw at any time from the study without loss of benefit or penalty

Special care has been taken where participants are unable to consent for themselves (e.g children under the age of 18, elders with age 85+, people with intellectual or learning disability, individuals or groups receiving help through the voluntary sector, those in a subordinate position to the researcher, groups who do not understand the consent and research process)

Participants have been informed of potential conflict of interest issues

The onus is on the researcher to inform participants if deception methods have to be used in a line of research

I have read, understood, and will adhere to the ethical principles described above in the conduct of the project work.

Signature: ……Benjamin Kelly……

Date: …30 Nov 2022……

Fingal Open Data: http://data.fingal.ie/About

Licence

Citizens are free to access and use this data as they wish, free of charge, in accordance with the Creative Commons Attribution 4.0 International License (CC-BY).

Note: From November 2010 to July 2015, data on Fingal Open Data was published in accordance with the PSI general licence.

Use of any published data is subject to Data Protection legislation.

Licence Statement

Under the CC-BY Licence, users must acknowledge the source of the Information in their product or application by including or linking to this attribution statement: "Contains Fingal County Council Data licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence".

Multiple Attributions

If using data from several Information Providers and listing multiple attributions is not practical in a product or application, users may include a URI or hyperlink to a resource that contains the required attribution statements.

Eurostat: https://ec.europa.eu/eurostat/about/policies/copyright

COPYRIGHT NOTICE AND FREE RE-USE OF DATA

Eurostat has a policy of encouraging free re-use of its data, both for non-commercial and commercial purposes. All statistical data, metadata, content of web pages or other dissemination tools, official publications and other

documents published on its website, with the exceptions listed below, can be reused without any payment or written licence provided that:

the source is indicated as Eurostat

when re-use involves modifications to the data or text, this must be stated clearly to the end user of the information

Journal of Statistics Education: http://jse.amstat.org/jse_users.htm

JSE Copyright and Usage Policy

Unlike other American Statistical Association journals, the Journal of Statistics Education (JSE) does not require authors to transfer copyright for the published material to JSE. Authors maintain copyright of published material. Because copyright is not transferred from the author, permission to use materials published by JSE remains with the author. Therefore, to use published material from a JSE article the requesting person must get approval from the author.

December 2022 Ethics Forms

National College of Ireland

DECLARATION OF ETHICS CONSIDERATION

School of Computing

Student Name: …Benjamin Kelly……

Student ID: …x19370681……..……

Programme … BSHDS4…… Year: …4……..

Module: …Computing Project…….……

Project Title: An analysis of the causes and prevention of diabetes internationally and lessons for Irish healthcare

Please circle (or highlight) as appropriate

This project involves human participants	No

Introduction

Secondary data refers to data that is collected by someone other than the current researcher. Common sources of secondary data for social science include censuses, information collected by government departments, organizational records and data originally collected for other research purposes. Primary data, by contrast, is collected by the investigator conducting the research.

A project that does not involve human participants requires ONLY completion of Declaration of Ethics Consideration Form and submission of the form on module's Moodle page

A project that involves human participants requires ethical clearance and an Ethics Application Form must be submitted through the module's Moodle page. Please refer to and ensure compliance with the ethical principles stated in NCI Ethics Form available on the Moodle page.

The following decision table will assist you in deciding if you have to complete the Declaration of Ethics Consideration Form or/and the Ethics Application Form.

Public Data	Y	Y	Y	Y	N	N	N	N
Private Data	Y	Y	N	N	Y	Y	N	N
Human Participants	Y	N	Y	N	Y	N	Y	N
Declaration of Ethics Consideration Form	x	X	x	X	X	X	x	
Ethics Application Form	X		X		X		X	

Please circle (or highlight) as appropriate

The project makes use of secondary dataset(s) created by the researcher	Yes
The project makes use of public secondary dataset(s)	Yes
The project makes use of non-public secondary dataset(s)	No
Approval letter from non-public secondary dataset(s) owner received	N/A

Sources of Data:

It is students' responsibility to ensure that they have the correct permissions/authorizations to use any data in a study. Projects that make use of data that does not have authorization to be used, will not be graded for that portion of the study that makes use of such data.

Public Data

A project that makes use of public secondary dataset(s) does not need ethics permission, but needs a letter/email from the copyright holder regarding potential use.

Some websites and data sources allow their data sets to be used under certain conditions. In these cases, a letter/email from the copyright holder is NOT necessary, but the researcher should cite the source of this permission and indicate under what conditions the data are allowed to be used. See Appendix I for examples of permissions granted by Fingal Open Data, and Eurostat website.

Where websites or data sources indicate that they do not grant permission for data to be used, you will still need a letter/email from the copyright holder. For example, see Appendix II for an example from the Journal of Statistics Education.

Private Data

A project that makes use of non-public (private) secondary dataset(s) must receive data usage permission from School of Computing.

An approval letter/email from the owner (e.g. institution, company, etc.) of the non-public secondary dataset must be attached to the Declaration of Ethics Consideration. The letter/email must confirm that the dataset is anonymised and permission for data processing, analysis and public dissemination is granted.

Evidence for use of secondary dataset(s)

Include dataset(s) owner letter/email or cite the source for usage permission

Not applicable

CHECKLIST

Non-public/private secondary dataset(s) -Owner letter/email is attached to this form *OR* Citation and link to the web site where permission is granted – provided in this form	N/A N/A

ETHICS CLEARANCE GUIDELINES WHEN HUMAN PARTICIPANTS ARE INVOLVED

The Ethics Application Form must be submitted on Moodle for approval prior to conducting the work.

Considerations in data collection

Participants will not be identified, directly or through identifiers linked to the subjects in any reports produced by the study

Responses will not place the participants at risk of professional liability or be damaging to the participants' financial standing, employability or reputation

No confidential data will be used for personal advantage or that of a third party

Informed consent

Consent to participate in the study has been given freely by the participants

participants have the capacity to understand the project goals.

Participants have been given information sheets that are understandable

Likely benefits of the project itself have been explained to potential participants

Risks and benefits of the project have been explained to potential participants

Participants have been assured they will not suffer physical stress or discomfort or psychological or mental stress

The participant has been assured s/he may withdraw at any time from the study without loss of benefit or penalty

Special care has been taken where participants are unable to consent for themselves (e.g children under the age of 18, elders with age 85+, people with intellectual or learning disability, individuals or groups receiving help through the voluntary sector, those in a subordinate position to the researcher, groups who do not understand the consent and research process)

Participants have been informed of potential conflict of interest issues

The onus is on the researcher to inform participants if deception methods have to be used in a line of research

I have read, understood, and will adhere to the ethical principles described above in the conduct of the project work.

Signature: ……Benjamin Kelly……

Date: …20 Dec 2022……

Fingal Open Data: http://data.fingal.ie/About

Licence

Citizens are free to access and use this data as they wish, free of charge, in accordance with the Creative Commons Attribution 4.0 International License (CC-BY).

Note: From November 2010 to July 2015, data on Fingal Open Data was published in accordance with the PSI general licence.

Use of any published data is subject to Data Protection legislation.

Licence Statement

Under the CC-BY Licence, users must acknowledge the source of the Information in their product or application by including or linking to this attribution statement: "Contains Fingal County Council Data licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence".

Multiple Attributions

If using data from several Information Providers and listing multiple attributions is not practical in a product or application, users may include a URI or hyperlink to a resource that contains the required attribution statements.

Eurostat: https://ec.europa.eu/eurostat/about/policies/copyright

COPYRIGHT NOTICE AND FREE RE-USE OF DATA

Eurostat has a policy of encouraging free re-use of its data, both for non-commercial and commercial purposes. All statistical data, metadata, content of web pages or other dissemination tools, official publications and other documents published on its website, with the exceptions listed below, can be reused without any payment or written licence provided that:

the source is indicated as Eurostat

when re-use involves modifications to the data or text, this must be stated clearly to the end user of the information

Journal of Statistics Education: http://jse.amstat.org/jse_users.htm

JSE Copyright and Usage Policy

Unlike other American Statistical Association journals, the Journal of Statistics Education (JSE) does not require authors to transfer copyright for the published material to JSE. Authors maintain copyright of published material. Because copyright is not transferred from the author, permission to use materials published by JSE remains with the author. Therefore, to use published material from a JSE article the requesting person must get approval from the author.

January 2023 Ethics Forms

National College of Ireland

DECLARATION OF ETHICS CONSIDERATION

School of Computing

Student Name: …Benjamin Kelly……

Student ID: …x19370681……..……

Programme … BSHDS4…… Year: …4……..

Module: …Computing Project…….……

Project Title: An analysis of the causes and prevention of diabetes internationally and lessons for Irish healthcare

Please circle (or highlight) as appropriate

This project involves human participants	No

Introduction

Secondary data refers to data that is collected by someone other than the current researcher. Common sources of secondary data for social science include censuses, information collected by government departments, organizational records and data originally collected for other research purposes. Primary data, by contrast, is collected by the investigator conducting the research.

A project that does not involve human participants requires ONLY completion of Declaration of Ethics Consideration Form and submission of the form on module's Moodle page

A project that involves human participants requires ethical clearance and an Ethics Application Form must be submitted through the module's Moodle page. Please refer to and ensure compliance with the ethical principles stated in NCI Ethics Form available on the Moodle page.

The following decision table will assist you in deciding if you have to complete the Declaration of Ethics Consideration Form or/and the Ethics Application Form.

Public Data	Y	Y	Y	Y	N	N	N	N
Private Data	Y	Y	N	N	Y	Y	N	N
Human Participants	Y	N	Y	N	Y	N	Y	N
Declaration of Ethics Consideration Form	x	X	x	X	X	X	x	
Ethics Application Form	X		X		X		X	

Please circle (or highlight) as appropriate

The project makes use of secondary dataset(s) created by the researcher	Yes
The project makes use of public secondary dataset(s)	Yes
The project makes use of non-public secondary dataset(s)	No
Approval letter from non-public secondary dataset(s) owner received	N/A

Sources of Data:

It is students' responsibility to ensure that they have the correct permissions/authorizations to use any data in a study. Projects that make use of data that does not have authorization to be used, will not be graded for that portion of the study that makes use of such data.

Public Data

A project that makes use of public secondary dataset(s) does not need ethics permission, but needs a letter/email from the copyright holder regarding potential use.

Some websites and data sources allow their data sets to be used under certain conditions. In these cases, a letter/email from the copyright holder is NOT necessary, but the researcher should cite the source of this permission and indicate under what conditions the data are allowed to be used. See Appendix I for examples of permissions granted by Fingal Open Data, and Eurostat website.

Where websites or data sources indicate that they do not grant permission for data to be used, you will still need a letter/email from the copyright holder. For example, see Appendix II for an example from the Journal of Statistics Education.

Private Data

A project that makes use of non-public (private) secondary dataset(s) must receive data usage permission from School of Computing.

An approval letter/email from the owner (e.g. institution, company, etc.) of the non-public secondary dataset must be attached to the Declaration of Ethics Consideration. The letter/email must confirm that the dataset is anonymised and permission for data processing, analysis and public dissemination is granted.

Evidence for use of secondary dataset(s)

Include dataset(s) owner letter/email or cite the source for usage permission

Not applicable

CHECKLIST

Non-public/private secondary dataset(s) -Owner letter/email is attached to this form *OR* Citation and link to the web site where permission is granted – provided in this form	N/A N/A

ETHICS CLEARANCE GUIDELINES WHEN HUMAN PARTICIPANTS ARE INVOLVED

The Ethics Application Form must be submitted on Moodle for approval prior to conducting the work.

Considerations in data collection

Participants will not be identified, directly or through identifiers linked to the subjects in any reports produced by the study

Responses will not place the participants at risk of professional liability or be damaging to the participants' financial standing, employability or reputation

No confidential data will be used for personal advantage or that of a third party

Informed consent

Consent to participate in the study has been given freely by the participants

participants have the capacity to understand the project goals.

Participants have been given information sheets that are understandable

Likely benefits of the project itself have been explained to potential participants

Risks and benefits of the project have been explained to potential participants

Participants have been assured they will not suffer physical stress or discomfort or psychological or mental stress

The participant has been assured s/he may withdraw at any time from the study without loss of benefit or penalty

Special care has been taken where participants are unable to consent for themselves (e.g children under the age of 18, elders with age 85+, people with intellectual or learning disability, individuals or groups receiving help through the voluntary sector, those in a subordinate position to the researcher, groups who do not understand the consent and research process)

Participants have been informed of potential conflict of interest issues

The onus is on the researcher to inform participants if deception methods have to be used in a line of research

I have read, understood, and will adhere to the ethical principles described above in the conduct of the project work.

Signature: ……Benjamin Kelly……

Date: …31 Jan 2023……

Fingal Open Data: http://data.fingal.ie/About

Licence

Citizens are free to access and use this data as they wish, free of charge, in accordance with the Creative Commons Attribution 4.0 International License (CC-BY).

Note: From November 2010 to July 2015, data on Fingal Open Data was published in accordance with the PSI general licence.

Use of any published data is subject to Data Protection legislation.

Licence Statement

Under the CC-BY Licence, users must acknowledge the source of the Information in their product or application by including or linking to this attribution statement: "Contains Fingal County Council Data licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence".

Multiple Attributions

If using data from several Information Providers and listing multiple attributions is not practical in a product or application, users may include a URI or hyperlink to a resource that contains the required attribution statements.

Eurostat: https://ec.europa.eu/eurostat/about/policies/copyright

COPYRIGHT NOTICE AND FREE RE-USE OF DATA

Eurostat has a policy of encouraging free re-use of its data, both for non-commercial and commercial purposes. All statistical data, metadata, content of web pages or other dissemination tools, official publications and other documents published on its website, with the exceptions listed below, can be reused without any payment or written licence provided that:

the source is indicated as Eurostat

when re-use involves modifications to the data or text, this must be stated clearly to the end user of the information

Journal of Statistics Education: http://jse.amstat.org/jse_users.htm

JSE Copyright and Usage Policy

Unlike other American Statistical Association journals, the Journal of Statistics Education (JSE) does not require authors to transfer copyright for the published material to JSE. Authors maintain copyright of published material. Because copyright is not transferred from the author, permission to use materials published by JSE remains with the author. Therefore, to use published material from a JSE article the requesting person must get approval from the author.

National College of Ireland

February 2023 Ethics Forms

National College of Ireland

DECLARATION OF ETHICS CONSIDERATION

School of Computing

Student Name: …Benjamin Kelly……

Student ID: …x19370681……..……

Programme … BSHDS4…… Year: …4……..

Module: …Computing Project…….……

Project Title: An analysis of the causes and prevention of diabetes internationally and lessons for Irish healthcare

Please circle (or highlight) as appropriate

This project involves human participants	No

Introduction

Secondary data refers to data that is collected by someone other than the current researcher. Common sources of secondary data for social science include censuses, information collected by government departments, organizational records and data originally collected for other research purposes. Primary data, by contrast, is collected by the investigator conducting the research.

A project that does not involve human participants requires ONLY completion of Declaration of Ethics Consideration Form and submission of the form on module's Moodle page

A project that involves human participants requires ethical clearance and an Ethics Application Form must be submitted through the module's Moodle page. Please refer to and ensure compliance with the ethical principles stated in NCI Ethics Form available on the Moodle page.

The following decision table will assist you in deciding if you have to complete the Declaration of Ethics Consideration Form or/and the Ethics Application Form.

Public Data	Y	Y	Y	Y	N	N	N	N
Private Data	Y	Y	N	N	Y	Y	N	N
Human Participants	Y	N	Y	N	Y	N	Y	N
Declaration of Ethics Consideration Form	x	X	x	X	X	X	x	
Ethics Application Form	X		X		X		X	

Please circle (or highlight) as appropriate

The project makes use of secondary dataset(s) created by the researcher	Yes
The project makes use of public secondary dataset(s)	Yes
The project makes use of non-public secondary dataset(s)	No
Approval letter from non-public secondary dataset(s) owner received	N/A

Sources of Data:

It is students' responsibility to ensure that they have the correct permissions/authorizations to use any data in a study. Projects that make use of data that does not have authorization to be used, will not be graded for that portion of the study that makes use of such data.

Public Data

A project that makes use of public secondary dataset(s) does not need ethics permission, but needs a letter/email from the copyright holder regarding potential use.

Some websites and data sources allow their data sets to be used under certain conditions. In these cases, a letter/email from the copyright holder is NOT necessary, but the researcher should cite the source of this permission and indicate under what conditions the data are allowed to be used. See Appendix I for examples of permissions granted by Fingal Open Data, and Eurostat website.

Where websites or data sources indicate that they do not grant permission for data to be used, you will still need a letter/email from the copyright holder. For example, see Appendix II for an example from the Journal of Statistics Education.

Private Data

A project that makes use of non-public (private) secondary dataset(s) must receive data usage permission from School of Computing.

An approval letter/email from the owner (e.g. institution, company, etc.) of the non-public secondary dataset must be attached to the Declaration of Ethics Consideration. The letter/email must confirm that the dataset is anonymised and permission for data processing, analysis and public dissemination is granted.

Evidence for use of secondary dataset(s)

Include dataset(s) owner letter/email or cite the source for usage permission

Not applicable

CHECKLIST

Non-public/private secondary dataset(s) -Owner letter/email is attached to this form *OR* Citation and link to the web site where permission is granted – provided in this form	N/A N/A

ETHICS CLEARANCE GUIDELINES WHEN HUMAN PARTICIPANTS ARE INVOLVED

The Ethics Application Form must be submitted on Moodle for approval prior to conducting the work.

Considerations in data collection

Participants will not be identified, directly or through identifiers linked to the subjects in any reports produced by the study

Responses will not place the participants at risk of professional liability or be damaging to the participants' financial standing, employability or reputation

No confidential data will be used for personal advantage or that of a third party

Informed consent

Consent to participate in the study has been given freely by the participants

participants have the capacity to understand the project goals.

Participants have been given information sheets that are understandable

Likely benefits of the project itself have been explained to potential participants

Risks and benefits of the project have been explained to potential participants

Participants have been assured they will not suffer physical stress or discomfort or psychological or mental stress

The participant has been assured s/he may withdraw at any time from the study without loss of benefit or penalty

Special care has been taken where participants are unable to consent for themselves (e.g children under the age of 18, elders with age 85+, people with intellectual or learning disability, individuals or groups receiving help through the voluntary sector, those in a subordinate position to the researcher, groups who do not understand the consent and research process)

Participants have been informed of potential conflict of interest issues

The onus is on the researcher to inform participants if deception methods have to be used in a line of research

I have read, understood, and will adhere to the ethical principles described above in the conduct of the project work.

Signature: ……Benjamin Kelly……

Date: …03 Feb 2023……

Fingal Open Data: http://data.fingal.ie/About

Licence

Citizens are free to access and use this data as they wish, free of charge, in accordance with the Creative Commons Attribution 4.0 International License (CC-BY).

Note: From November 2010 to July 2015, data on Fingal Open Data was published in accordance with the PSI general licence.

Use of any published data is subject to Data Protection legislation.

Licence Statement

Under the CC-BY Licence, users must acknowledge the source of the Information in their product or application by including or linking to this attribution

statement: “Contains Fingal County Council Data licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence”.

Multiple Attributions

If using data from several Information Providers and listing multiple attributions is not practical in a product or application, users may include a URI or hyperlink to a resource that contains the required attribution statements.

Eurostat: https://ec.europa.eu/eurostat/about/policies/copyright

COPYRIGHT NOTICE AND FREE RE-USE OF DATA

Eurostat has a policy of encouraging free re-use of its data, both for non-commercial and commercial purposes. All statistical data, metadata, content of web pages or other dissemination tools, official publications and other documents published on its website, with the exceptions listed below, can be reused without any payment or written licence provided that:

the source is indicated as Eurostat

when re-use involves modifications to the data or text, this must be stated clearly to the end user of the information

Journal of Statistics Education: http://jse.amstat.org/jse_users.htm

JSE Copyright and Usage Policy

Unlike other American Statistical Association journals, the Journal of Statistics Education (JSE) does not require authors to transfer copyright for the published material to JSE. Authors maintain copyright of published material. Because copyright is not transferred from the author, permission to use materials published by JSE remains with the author. Therefore, to use published material from a JSE article the requesting person must get approval from the author.

March 2023 Ethics Forms

National College of Ireland

DECLARATION OF ETHICS CONSIDERATION

School of Computing

Student Name: …Benjamin Kelly……

Student ID: …x19370681……..……

Programme … BSHDS4…… Year: …4……..

Module: …Computing Project…….……

Project Title: An analysis of the causes and prevention of diabetes.

Please circle (or highlight) as appropriate

This project involves human participants	No

Introduction

Secondary data refers to data that is collected by someone other than the current researcher. Common sources of secondary data for social science include censuses, information collected by government departments, organizational records and data originally collected for other research purposes. Primary data, by contrast, is collected by the investigator conducting the research.

A project that does not involve human participants requires ONLY completion of Declaration of Ethics Consideration Form and submission of the form on module's Moodle page

A project that involves human participants requires ethical clearance and an Ethics Application Form must be submitted through the module's Moodle page. Please refer to and ensure compliance with the ethical principles stated in NCI Ethics Form available on the Moodle page.

The following decision table will assist you in deciding if you have to complete the Declaration of Ethics Consideration Form or/and the Ethics Application Form.

Public Data	Y	Y	Y	Y	N	N	N	N
Private Data	Y	Y	N	N	Y	Y	N	N
Human Participants	Y	N	Y	N	Y	N	Y	N

Declaration of Ethics Consideration Form	x	X	x	X	X	X	x	
Ethics Application Form	X		X		X		X	

Please circle (or highlight) as appropriate

The project makes use of secondary dataset(s) created by the researcher	Yes
The project makes use of public secondary dataset(s)	Yes
The project makes use of non-public secondary dataset(s)	No
Approval letter from non-public secondary dataset(s) owner received	N/A

Sources of Data:

It is students' responsibility to ensure that they have the correct permissions/authorizations to use any data in a study. Projects that make use of data that does not have authorization to be used, will not be graded for that portion of the study that makes use of such data.

Public Data

A project that makes use of public secondary dataset(s) does not need ethics permission, but needs a letter/email from the copyright holder regarding potential use.

Some websites and data sources allow their data sets to be used under certain conditions. In these cases, a letter/email from the copyright holder is NOT necessary, but the researcher should cite the source of this permission and indicate under what conditions the data are allowed to be used. See Appendix I for examples of permissions granted by Fingal Open Data, and Eurostat website.

Where websites or data sources indicate that they do not grant permission for data to be used, you will still need a letter/email from the copyright holder. For example, see Appendix II for an example from the Journal of Statistics Education.

Private Data

A project that makes use of non-public (private) secondary dataset(s) must receive data usage permission from School of Computing.

An approval letter/email from the owner (e.g. institution, company, etc.) of the non-public secondary dataset must be attached to the Declaration of Ethics Consideration. The letter/email must confirm that the dataset is anonymised and permission for data processing, analysis and public dissemination is granted.

Evidence for use of secondary dataset(s)

Include dataset(s) owner letter/email or cite the source for usage permission

Not applicable

CHECKLIST

Non-public/private secondary dataset(s) -Owner letter/email is attached to this form *OR* Citation and link to the web site where permission is granted – provided in this form	N/A N/A

ETHICS CLEARANCE GUIDELINES WHEN HUMAN PARTICIPANTS ARE INVOLVED

The Ethics Application Form must be submitted on Moodle for approval prior to conducting the work.

Considerations in data collection

Participants will not be identified, directly or through identifiers linked to the subjects in any reports produced by the study

Responses will not place the participants at risk of professional liability or be damaging to the participants' financial standing, employability or reputation

No confidential data will be used for personal advantage or that of a third party

Informed consent

Consent to participate in the study has been given freely by the participants

participants have the capacity to understand the project goals.

Participants have been given information sheets that are understandable

Likely benefits of the project itself have been explained to potential participants

Risks and benefits of the project have been explained to potential participants

Participants have been assured they will not suffer physical stress or discomfort or psychological or mental stress

The participant has been assured s/he may withdraw at any time from the study without loss of benefit or penalty

Special care has been taken where participants are unable to consent for themselves (e.g children under the age of 18, elders with age 85+, people with

intellectual or learning disability, individuals or groups receiving help through the voluntary sector, those in a subordinate position to the researcher, groups who do not understand the consent and research process)

Participants have been informed of potential conflict of interest issues

The onus is on the researcher to inform participants if deception methods have to be used in a line of research

I have read, understood, and will adhere to the ethical principles described above in the conduct of the project work.

Signature: ……Benjamin Kelly……

Date: …31 March 2023……

Fingal Open Data: http://data.fingal.ie/About

Licence

Citizens are free to access and use this data as they wish, free of charge, in accordance with the Creative Commons Attribution 4.0 International License (CC-BY).

Note: From November 2010 to July 2015, data on Fingal Open Data was published in accordance with the PSI general licence.

Use of any published data is subject to Data Protection legislation.

Licence Statement

Under the CC-BY Licence, users must acknowledge the source of the Information in their product or application by including or linking to this attribution statement: "Contains Fingal County Council Data licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence".

Multiple Attributions

If using data from several Information Providers and listing multiple attributions is not practical in a product or application, users may include a URI or hyperlink to a resource that contains the required attribution statements.

Eurostat: https://ec.europa.eu/eurostat/about/policies/copyright

COPYRIGHT NOTICE AND FREE RE-USE OF DATA

Eurostat has a policy of encouraging free re-use of its data, both for non-commercial and commercial purposes. All statistical data, metadata, content of web pages or other dissemination tools, official publications and other documents published on its website, with the exceptions listed below, can be reused without any payment or written licence provided that:

the source is indicated as Eurostat

when re-use involves modifications to the data or text, this must be stated clearly to the end user of the information

Journal of Statistics Education: http://jse.amstat.org/jse_users.htm

JSE Copyright and Usage Policy

Unlike other American Statistical Association journals, the Journal of Statistics Education (JSE) does not require authors to transfer copyright for the published material to JSE. Authors maintain copyright of published material. Because copyright is not transferred from the author, permission to use materials published by JSE remains with the author. Therefore, to use published material from a JSE article the requesting person must get approval from the author.

April 2023 Ethics Forms

National College of Ireland

DECLARATION OF ETHICS CONSIDERATION

School of Computing

Student Name: …Benjamin Kelly……

Student ID: …x19370681……..……

Programme … BSHDS4…… Year: …4……..

Module: …Computing Project…….……

Project Title: An analysis of the causes and prevention of diabetes.

Please circle (or highlight) as appropriate

This project involves human participants	No

Introduction

Secondary data refers to data that is collected by someone other than the current researcher. Common sources of secondary data for social science include censuses, information collected by government departments, organizational records and data originally collected for other research purposes. Primary data, by contrast, is collected by the investigator conducting the research.

A project that does not involve human participants requires ONLY completion of Declaration of Ethics Consideration Form and submission of the form on module's Moodle page

A project that involves human participants requires ethical clearance and an Ethics Application Form must be submitted through the module's Moodle page. Please refer to and ensure compliance with the ethical principles stated in NCI Ethics Form available on the Moodle page.

The following decision table will assist you in deciding if you have to complete the Declaration of Ethics Consideration Form or/and the Ethics Application Form.

Public Data	Y	Y	Y	Y	N	N	N	N

Private Data	Y	Y	N	N	Y	Y	N	N
Human Participants	Y	N	Y	N	Y	N	Y	N
Declaration of Ethics Consideration Form	x	X	x	X	X	X	x	
Ethics Application Form	X		X		X		X	

Please circle (or highlight) as appropriate

The project makes use of secondary dataset(s) created by the researcher	Yes
The project makes use of public secondary dataset(s)	Yes
The project makes use of non-public secondary dataset(s)	No
Approval letter from non-public secondary dataset(s) owner received	N/A

Sources of Data:

It is students' responsibility to ensure that they have the correct permissions/authorizations to use any data in a study. Projects that make use of data that does not have authorization to be used, will not be graded for that portion of the study that makes use of such data.

Public Data

A project that makes use of public secondary dataset(s) does not need ethics permission, but needs a letter/email from the copyright holder regarding potential use.

Some websites and data sources allow their data sets to be used under certain conditions. In these cases, a letter/email from the copyright holder is NOT necessary, but the researcher should cite the source of this permission and indicate under what conditions the data are allowed to be used. See Appendix I for examples of permissions granted by Fingal Open Data, and Eurostat website.

Where websites or data sources indicate that they do not grant permission for data to be used, you will still need a letter/email from the copyright holder. For example, see Appendix II for an example from the Journal of Statistics Education.

Private Data

A project that makes use of non-public (private) secondary dataset(s) must receive data usage permission from School of Computing.

An approval letter/email from the owner (e.g. institution, company, etc.) of the non-public secondary dataset must be attached to the Declaration of Ethics Consideration. The letter/email must confirm that the dataset is anonymised and permission for data processing, analysis and public dissemination is granted.

Evidence for use of secondary dataset(s)

Include dataset(s) owner letter/email or cite the source for usage permission

Not applicable

CHECKLIST

Non-public/private secondary dataset(s) -Owner letter/email is attached to this form *OR* Citation and link to the web site where permission is granted – provided in this form	N/A N/A

ETHICS CLEARANCE GUIDELINES WHEN HUMAN PARTICIPANTS ARE INVOLVED

The Ethics Application Form must be submitted on Moodle for approval prior to conducting the work.

Considerations in data collection

Participants will not be identified, directly or through identifiers linked to the subjects in any reports produced by the study

Responses will not place the participants at risk of professional liability or be damaging to the participants' financial standing, employability or reputation

No confidential data will be used for personal advantage or that of a third party

Informed consent

Consent to participate in the study has been given freely by the participants

participants have the capacity to understand the project goals.

Participants have been given information sheets that are understandable

Likely benefits of the project itself have been explained to potential participants

Risks and benefits of the project have been explained to potential participants

Participants have been assured they will not suffer physical stress or discomfort or psychological or mental stress

The participant has been assured s/he may withdraw at any time from the study without loss of benefit or penalty

Special care has been taken where participants are unable to consent for themselves (e.g children under the age of 18, elders with age 85+, people with intellectual or learning disability, individuals or groups receiving help through the voluntary sector, those in a subordinate position to the researcher, groups who do not understand the consent and research process)

Participants have been informed of potential conflict of interest issues

The onus is on the researcher to inform participants if deception methods have to be used in a line of research

I have read, understood, and will adhere to the ethical principles described above in the conduct of the project work.

Signature: ……Benjamin Kelly……

Date: …1st May 2023……

Fingal Open Data: http://data.fingal.ie/About

Licence

Citizens are free to access and use this data as they wish, free of charge, in accordance with the Creative Commons Attribution 4.0 International License (CC-BY).

Note: From November 2010 to July 2015, data on Fingal Open Data was published in accordance with the PSI general licence.

Use of any published data is subject to Data Protection legislation.

Licence Statement

Under the CC-BY Licence, users must acknowledge the source of the Information in their product or application by including or linking to this attribution statement: "Contains Fingal County Council Data licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence".

Multiple Attributions

If using data from several Information Providers and listing multiple attributions is not practical in a product or application, users may include a URI or hyperlink to a resource that contains the required attribution statements.

Eurostat: https://ec.europa.eu/eurostat/about/policies/copyright

COPYRIGHT NOTICE AND FREE RE-USE OF DATA

Eurostat has a policy of encouraging free re-use of its data, both for non-commercial and commercial purposes. All statistical data, metadata, content of web pages or other dissemination tools, official publications and other documents published on its website, with the exceptions listed below, can be reused without any payment or written licence provided that:

the source is indicated as Eurostat

when re-use involves modifications to the data or text, this must be stated clearly to the end user of the information

Journal of Statistics Education: http://jse.amstat.org/jse_users.htm

JSE Copyright and Usage Policy

Unlike other American Statistical Association journals, the Journal of Statistics Education (JSE) does not require authors to transfer copyright for the published material to JSE. Authors maintain copyright of published material. Because copyright is not transferred from the author, permission to use materials published by JSE remains with the author. Therefore, to use published material from a JSE article the requesting person must get approval from the author.

7.2. Reflective Journals

The following section contains my monthly reflective journals for the project.

Supervision & Reflection Journal - October

Student Name	Benjamin Kelly
Student Number	x19370681
Course	BSHDS4

Month: October

Reflect on what has happened in your project this month?

During the month of October, I attended the introductory lecture on the 4th year Project which provided me with valuable guidelines on the scope of the project, the timelines, the marking system and support available at NCI. I also had the opportunity to consider the optional areas which I might look at in the project and initially looked at a range of options including Diabetes, a project analysing Covid data and trends in Ireland and a possible project in fraud detection which I had been thinking about as an extension of my work at Pobal.

I carried out some initial research in all three areas and finally decided to run with the diabetes option because of the depth of data available and the flexibility it offered to narrow down the focus of the project as the year progresses. Covid data is relatively new, and the fraud area is one where I had difficulty finding suitable wider datasets to work with as the Pobal work was very focused on one dataset and grant scheme.

I put together the project summary video which was a very useful exercise for me as it forced me to focus on the area and problem I was hoping to solve.

So What? Consider what that meant for your project progress. What were your successes? What challenges still remain?

As I have further researched the project idea and the datasets available for the area of diabetes I have come to various conclusions:

- There is a lot of clinical data associated with the condition and I do not intend to focus too much on clinical data as this would require me to become more expert on the condition than I am trained for.
- I also feel that if I can focus on lifestyle, causes, and global incidence comparison data this will be easier to understand but it is equally important data for analysis to inform the healthcare sector
- I feel that I still have a lot of data to research and review and there will be challenges in finding common variables to allow me to make sensible comparisons.
- I also think that the goal and objectives of the project may need to be fine tuned as I do this research and the project question may change a little during the year.

However, I am confident that there is a wide enough range of research projects, data and variables to allow me to change as I learn and still deliver interesting results and conclusions at the end of the project. • My successes to date have been in getting a broad understanding of the condition of diabetes, which is very interesting. I also am getting a better idea of the sort of data which is available from a wide range of sources internationally and in Ireland on the subject.
Now What? What can you do to address outstanding challenges? • I am looking forward to reviewing the project proposal with my academic supervisor and discussing how it might be refined • I also need to spend more time really looking at the data available and discussing with my academic supervisor any fine tuning of the objectives and the proposed analysis using the final datasets identified • I can then move on to the actual pre-processing activities on the selected datasets

Student Signature	**Benjamin Kelly**

Supervision & Reflection November

Student Name	Benjamin Kelly
Student Number	X19370681
Course	BSc. Honours in Data Science (BSHDS4)
Supervisor	Michael Bradford

Month: November 2022

What I have been doing in November

Project administration

During the month of November, I carried out a review of my timetable as the project overlaps with other subject tasks and deadlines. As a result, I needed to restructure my priorities to catch up on project work. I discussed this with my project supervisor, and we agreed that putting in place project milestones would be very helpful for me to prioritise and manage my schedule. This meant identifying a series of small goals around which I could build a path up to my mid-point implementation deadline in December .

To assist me with my time management I also had some planning sessions with the college some project to assist with my task organisation in the run up to the December deadlines.

Data sets

I have identified a number of possible useful datasets which provide relevant data for the proposed project relating to Diabetes, lifestyle and clinical data from a range of data sources. I am beginning to the process of exploring these datasets and cleaning them for the project. I have commenced my first cleaning and exploration task in RStudio with one of the diabetes datasets. My supervisor and I discussed ideas around searching for other useful information and datasets of relevance to my topic of interest in other source areas such as the CSO. This is one of my next tasks on the project. My supervisor and I also discussed project structure using CRISP-DM.

My intention is to continue to search the web for more datasets, but I will not be focusing as much time on this research due to the fact that I have some data to work with and due to limitations on the time available in December with other projects at hand.

So What? - Consider what that meant for your project progress. What were your successes? What challenges still remain?

I am learning that while I can spend all available hours, seven days each week working on my various tasks, I can be more efficient with my time. I am improving my project administration by using short term milestones which is crucial to managing multiple parallel tasks within the available time.

Due to the multiple parallel tasks, I missed out on a couple of supervisor meetings which meant that my supervisor did not always have an up to date picture with how was progressing with my project work. I plan to communicate with my supervisor more often to address this issue in the future months of the project.

Now What? - What can you do to address outstanding challenges?

I am setting up a project timetable which details all my project milestones/goals and this will allow me to plan and allocate appropriate timeslots to each of these. This can also be flexible and change due to the impact of other assignments I am undertaking. My supervisor and I have agreed to put in place more regular meetings to allow us both to keep on top of my project progress. I have also done the following to help my project administration:

1. I have reorganised all of my project files on explorer into a tidy folder tree
2. I have set up a project planning area on my laptop using “onenote tabs” covering the following:
    a. My project goals
    b. My notes from meetings with my supervisor
    c. Project documentation + links the documentation
    d. Timetable
    e. To do list
    f. Data and other resources tabs

Student Signature	Benjamin Kelly

Supervision & Reflection December

Student Name	Benjamin Kelly
Student Number	X19370681
Course	BSc. Honours in Data Science
Supervisor	Michael Bradford

Month: December

Reflection on what has happened in my project this month

- During the month I had issues finding time to allocate my project up to mid-December due to the number of other December deadlines and assignments.
- I carried out multiple exploration tasks on my chosen datasets. These tasks included using RStudio for statistical analysis to using Power BI for exploratory analysis.
- I implemented the above for two Diabetes datasets from Kaggle.
- The tasks included; Statistical analysis, Dependent and Independent variable analysis, Independent feature redundancy and merging data frames to create one data frame that could be used to address the problem being solved.
- I had pulled together a report for the mid-point implementation.

So What? Considering what this has meant for my project progress. What are my successes and what challenges still remain.

- The statistical analysis provided me with a better understanding of the features within the datasets.
- The exploratory analysis in Power BI helped me focus on the correlations of pairs.
- I plan to carry out further structuring and analysis of these results. This will be an important step to help me have a more complete understanding of my chosen datasets.
- My task to merge the data frames is still in progress and is not complete yet. Once I get this working it will have a binary feature as a dependent within my working data.
- A few other statistical methods need to be applied to help explain the relationships of the features.
- Some literature review needs to be explored and explained for the project. I plan to use the library resources to collect such papers.
- I have a few areas of the project that I need to look into such as the modelling stage.

Now What? What I can do to address outstanding challenges

- Understanding and pulling together all of the analysis to form a picture will be an outstanding challenge. However, I plan to use mind-maps to lay out all the steps taken as I progress through the project.
- As part of my merging process, I have two datasets. One contains a binary feature which can be used to predict whether someone is diabetic. This will possibly need to be removed along with other features and combined to another dataset.

• I have planned out the areas of the project such as modelling, evaluations etc.. through the project documentation.

Student Signature	Benjamin Kelly

Supervision & Reflection January

Student Name	Benjamin Kelly
Student Number	X19370681
Course	BSc. Honours in Data Science
Supervisor	Michael Bradford

Month: January

Reflection on what has happened in my project this month

- I looked over what was accomplished during my first semester. This was done by reviewing my practical work, reports and monthly journals.
- I noticed a few areas that could be improved in my coding.
- I reviewed and tweaked my previous work in the area of statistical analysis and visualisation in order to get better results.
- The plan in the future will involve looking for suitable machine learning models that would work for the various tasks which need to be carried out.

So What? Considering what this has meant for my project progress. What are my successes and what challenges still remain.

- I am in the progress of setting up a weekly schedule to meet my supervisor to discuss project progress and changes committed.
- I am still planning and researching my data so it is suitable for the machine learning phase.
- Further research into other datasets may happen depending on results collected from the previous data.

Now What? What I can do to address outstanding challenges

- I plan to structure my tasks into a timetable on the run up to each meeting with my supervisor.

Student Signature	Benjamin Kelly

Supervision & Reflection February

Student Name	Benjamin Kelly
Student Number	X19370681
Course	BSc. Honours in Data Science
Supervisor	Michael Bradford

Month: February

Reflection on what has happened in my project this month

- Following discussions with my supervisor, I began the Modelling stage of the project.
- Some of the models being currently worked on include decision trees and random forest.
- Various de-bugging issues had to be completed to proceed with the modelling.
- The two datasets prepared previously in past journals are being used for the project.
- Once the models are prepared and analysed multiple times I plan to then use Neural Networks and AI based methods to formulate predictions.
- Some of the NN libraries which I plan to use later include Kera's Tuner and possibly Tensor-board.
- This work will primarily be carried out using Python as the preferred language for deep learning.
- I have set milestones involving four different models completed this month for decision trees and random forest.
- I have to date completed three different classification models; Random Forest, Naïve Bayes and Decision trees. I am also working on using Support Vector Machines.
- I have applied down sampling techniques to my datasets as I suspected that the models may be fitting poorly.

So What? Considering what this has meant for my project progress. What are my successes and what challenges still remain.

- I am sorting through some code issues now which are hindering progress with the models.
- Clearly outlining target and sample features in both datasets remains to be solved.
- There is a strong likelihood that class imbalance may occur within my coding and this will need to be solved using a range of methods.
- I observed the results of my models and concluded that the models performed differently depending on the dataset used. The decision trees remained at a low-level performance whereas Naïve Bayes performed well especially for the binary dataset.

Now What? What I can do to address outstanding challenges

- I plan to use the help manual in R documentation to get assistance in areas where there are issues.
- I need to scan the datasets and pick out the target variables. I plan to use most of them as sample variables.
- The methods that I am considering, to deal with class imbalance, include 'Smote' which I have used in the past for similar problems.
- Based on the results given from the models I am continuing to look for ways to improve the performance as much as possible.

Student Signature	Benjamin Kelly

Supervision & Reflection March

Student Name	Benjamin Kelly
Student Number	X19370681
Course	BSc. Honours in Data Science
Supervisor	Michael Bradford

Month: March

Reflection on what has happened in my project this month

- This month was very busy with other assessments from other modules which ran up to 19th March. This limiting the available time for the project.
- My project work included initially trying to complete models and then get a suitable set of results. This involved carrying out the following activities: Splitting the data after reading in the data frame, fitting my data to a model , specifying the target feature from the data frame, predicting test results, implementing the confusion matrix, plotting the model, identifying the importance of features, variance and visualisation analysis, assessing accuracy for confusion matrixes and accuracy of the model.
- The different classification models used were: Decision Trees, Random Forest and Naïve Bayes. I am also planning to implement a Neural Networks model in addition to these.
- I decided to try to get three different models working for one dataset first and then move onto the other dataset. I prioritised the Binary diabetes dataset. This was agreed with my supervisor in a review. The reason for choosing this dataset was that the class imbalance remained very high with this one, unlike the other dataset.
- It was important to solve the proportion distribution for the two binary classes in the target variable for the dataset. I noticed that there were more non-diabetic cases than diabetic cases in the proportion distribution. I carried out a few sampling strategies to address this issue. These included down sampling, stratified sampling and SMOTE. So far to date, the down sampling strategy worked best but I am also in the process of implementing SMOTE to see if results would improve.

My thinking is to apply a sampling strategy and then do data partitioning afterwards. Here I experimented with different ratio splits and found the split of either 7:3 or 8:2 to be best.

- Many errors needed to be debugged due to unexpected package issues and code errors.
- The idea of sampling is to make the binary classes similar in size (Reduce the majority class size) which would affect the data frame and potentially making it smaller.
- Proportion checks were carried out to see if the result changed for the class sizes.
- Several different results will be produced from the data modelling stage of the project. These results will only apply to the binary dataset and not to the non-binary dataset. There are three different datasets being produced from the binary dataset; Original with regular splitting, Down sample with balance checks and SMOTE with balance checks. Applying four different models (Neural Networks, Random Forests, Decision Trees and Naïve Bayes Classifier) to these three datasets will give 12 different sets of results.

So What? Considering what this has meant for my project progress. What are my successes and what challenges still remain.

- I was able to successfully implement at least one Random Forest Model, Naïve Bayes Model and Decision trees Model on both binary and non-binary datasets. Due to the problems and issues arising over class imbalance with the binary dataset I decided to focus on just getting results for the binary dataset first.
- I was able to complete a Random Forest Model and Decision Trees Model with down sampling applied to the majority class of the target variable. This improved the results and made it more realistic. I have yet to do this for the Naïve Bayes Model.
- There are error and debugging issues that will need to be addressed over the course of the modelling stage.
- I plan to complete the 12 different results for the binary dataset by early April.
- The same strategy will have to be applied to the other dataset. There may be other issues which arise as the approach that works on one dataset doesn't mean it will work on the other dataset.
- I will need to do some updates on the final report implementation. I will also need to update the modelling section covering the work I did to date.
- I am aware that I will need to be able to balance this workload with other assessments in the coming month and this will be part of the challenge.

Now What? What I can do to address outstanding challenges

- I need to ensure, when doing the data modelling, each of the three different datasets are applied to all classification algorithms which I intend to use.
- Errors in coding will need to be tackled as every algorithm is built differently.
- Time will be needed to get the other dataset working in order to move to the next stage of the project.

- For the Neural Networks aspect, issues such as running time and set-up will need to be sorted.
- The report will be updated multiple times to reflect the different datasets and models completed.
- A clear structured time-table with milestones will be important.

Student Signature	Benjamin Kelly

Supervision & Reflection April

Student Name	Benjamin Kelly
Student Number	X19370681
Course	BSc. Honours in Data Science
Supervisor	Michael Bradford

Month: April

Reflection on what has happened in my project this month

- Due to other assessment deadlines, I struggled with making a lot of progress on my project prior to 25th April. I had informed my supervisor about this and arranged more frequent catch-up conversations after that date.
- I continued to implement Neural network models for one of my three datasets which I had created earlier last month.
- Once all the models were run for the dataset, I planned to create the neural network models in Python. This would allow me to see whether they performed better than they did with R. I used a simple Neural Net package to create the models in R. As I expected, these models took greater time to run than the other classification models.
- I used the Kera's package to create the Neural networks In Python.
- One of the neural networks for the Smote dataset is ongoing and has yet to be completed.
- I was able to successfully run a simple neural network using the Kera's library in Python.
- I commenced the Kera's tuner for one of my neural network models. The benefit of using this is that it picks the optimal set of hyperparameters for the TensorFlow program.
- I had to sort through numerous coding scripts I had used during the course of the project.

So What? Considering what this has meant for my project progress. What are my successes and what challenges still remain.

- The report (Final Implementation) has areas to still be written up and I need to update a number of sections such as: evaluation, results, process, and conclusions.
- My intention is that when I start my Neural Network in Python it would contain 3 hidden layers, each of which would contain at least 10 neurons. This would be a simple sequential model and I plan to downsize the number of records in the dataset if time is too excessive.
- In order to save time, I decided to import my clean dataset from R into Python and then run the models.
- If the running of the Neural networks is successful for all three datasets in Python then I plan to use Keras Tuner which would allow me to hook onto the Tensorboard and do some visualisations.
- The models completed to date will need to be run for the non-binary dataset, and I will need to set aside some time for this.
- The neural network implementation in Python can be adjusted and tuned to improve accuracy.

Now What? What I can do to address outstanding challenges

- Good project organisation skills will be important to ensure the results are collected from both diabetes datasets. At the moment, only one dataset, which has three subsets created, is being implemented.
- I will need to schedule more conversations with my supervisor over the remaining few weeks leading up to the deadline in May.
- The Neural Network can improve when parameters like batch sizer increase, adding validation and early stopping.
- Early stopping helps by stopping a Neural network from continuously running once all weights have converged.
- Validation split allows me to reserve some of the training data for validation.

Student Signature	Benjamin Kelly