Table of Contents

[Chapter 1 1](#_Toc82164669)

[Chapter 2 3](#_Toc82164670)

[Chapter 3 7](#_Toc82164671)

[Data Set 7](#_Toc82164672)

[Feature Selection 8](#_Toc82164673)

[Caret() R 8](#_Toc82164674)

[Machine Learning Algorithms 8](#_Toc82164675)

[**Prediction** **Algorithm** 8](#_Toc82164676)

[Chapter 4 11](#_Toc82164677)

[Chapter 5 17](#_Toc82164678)

**Diabetes Prediction Using Machine Learning**

# Chapter 1

**Introduction**

Health is currently one of the most discussed agenda in the world as more effort is being put into the development of new drugs and vaccines to combat various diseases especially after the recent pandemic. Health care systems are being designed to meet the health care and needs of people. As population continues to grow rapidly and hence will there be an ever-increasing demand for acute exhilaration of chronic illness and may routine (Song *et al., 2012*). Health problems always require a prompt action. Health services aimed at improving health or diagnostic treatment and rehabilitation of sick people is continually being developed. The several aspects of health services are actions to organize the inputs necessary for the provision of effective interventions as inclusive of promotion, prevention, cure rehabilitation and palliation efforts as oriented towards either individuals or populations (Kumar *et al., 2015*).

Diabetes is a life long disease that occurs when the pancreas doesn’t produce enough insulin or when the body can not effectively use the insulin that it produces. Insulin is a hormone that regulates the blood sugar. Hyperglycemia, or high blood sugar, is a common effect of uncontrolled diabetes and over time leads to a damage in the body’s systems, especially the nerves and blood vessels. According to a statistic done by the United States government as at 2014 8.5% of adults aged 18 years and above has diabetes. In 2015, diabetes was the direct cause of 1.6 million deaths and in 2012 high blood glucose was the cause of another 2.2 million deaths in the US. In 2015, the United States estimated that the number and percentage of people diagnosed and undiagnosed suffering from diabetes from among the adults aged 18 years and above will rise exponentially as at compared to then where the number of people with diabetes had risen to about 108 million in 1980 and 422 million in 2014. The global prevalence of diabetes has risen sharply from 4.7% in 1980 to 8.5% in 2014. Diabetes prevalence has seen more increase in middle- and low-income countries. Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke, and lower limb amputation. In 2015, an estimated 1.6 million deaths were caused by diabetes. Another 2.2 million deaths were attributed to high blood glucose in 2012. Almost half od all deaths attributable to high blood glucose occur before the age of 70 years. World health organization projects that diabetes will be the seventh leading cause of death by 2030. Accurate and efficient diabetes prediction system is needed to overcome all the deadly statistics that has been mentioned above.

R studio, PowerBI and Weka were used in this research for data analytics. R is an open-source software that was designed based on the S language. Its interface was deigned was so users can clearly view graphs, data tables, R code and output all at the same time. R also offers a feature that allows users to import CSV, Excel, SAS, SPSS, Stata files into it without having to write specific codes to do so. R also offers various libraries for performing clusters, feature selection, data cleaning and also prediction machine learning algorithms. R is basically suited for big data analytics as prediction algorithms such as Random Forest, decision trees and naïve bayes are used in this research.

The paper is divided into the following sections:

* This chapter that introduces the topic
* The next chapter that discusses the various existing health care systems and diabetes in details.
* Chapter 3 discusses the methodology used in this research paper
* Chapter 4 discusses the results and experiments including challenges
* Finally, chapter 5 discusses future work, recommendations and conclusions.

# Chapter 2

**Literature Review**

The field of big data analytics is an ever-growing field that uses machine learning and other available algorithms to find patterns in datasets. Machine learning is one of the most widely used method for finding patterns and making predictions from dataset. Machine learning is a well-established branch of computer science and it plays a huge role in the development of classification and predictive analysis systems. Minimizing the total number of feature present in a multi-dimensional dataset is known as feature selection. Feature fusion and multiple dictionary models can be implemented for higher accuracy, when Euclidean distance between the classification’s response vectors is used the similarity measure it enhances the accuracy of classification. Data mining algorithms perform better in development of intelligent computing systems for making good decisions. Some of the familiar machine learning algorithms include perceptrons, decision trees, naïve bayes and random forest algorithms. Predictive data mining methods are supervised and it focuses on induced models or theories from class-labeled data. The induced models are used for predictions and classifications. Maximum likelihood model is proposed to overcome difficulties in complex performance measure that occur in the predictive models.

Diabetes is a disease that cause an increase in blood sugar level of a person. According to the NHS it is a lifelong disease that can only be managed despite key development in medical science. If diabetes is not properly diagnosed or managed it causes death in its victims. This research seeks to develop a diabetes prediction system based in already existing data and patterns. Patients suffering from diabetes can’t properly process produced insulin as food is typically broken down into sugar (glucose) for digestion and released into the blood stream, patients suffering from this disease either don’t produce enough or don’t use it as well as they should so a lot of the produced insulin remains in the blood stream. As stated earlier this a chronic disease that patients typically suffer from till, they die.

There are basically two types of diabetes categorized as diabetes type 1, and type 3, and gestational diabetes.

Diabetes type 1: when suffering from this disease it is notice that the body attacks itself by mistake and stops the production of insulin symptoms of this diabetes type shows very early in people and is usually diagnosed in children, teenagers and young adults.

Diabetes type 2: when suffering from this disease it is typically noticed that the body make proper use of its produced insulin leaving huge amount of insulin in the blood about 90-95% of people diagnosed with diabetes have this type of diabetes. It is the most common type of diabetes and unlike type 1 its symptoms are generally delayed and it’s mostly diagnosed in adults.

Gestational diabetes: is basically diagnosed in pregnant women who have never suffered from diabetes, if this disease is diagnosed during pregnancy that means baby is at high risk of being exposed to disease or health risk after its birth. This disease usually goes away after pregnancy but it increases chances of getting type 2 diabetes later in life for both the mother and the baby.

There are also other forms of diabetes which are inherited are not commonly found in people but these types include monogenic diabetes and cystic fibrosis-related diabetes. This type of diabetes is suffered from birth as it is part of one’s genetic definition and can also only be managed and not cured.

The rate of diabetes as increased spontaneously in recent years as at 2015. According to the world health organization diabetes is commonly found in 1 of 4 people over the age of 65 years. As stated earlier in adults the most found cases of diabetes is diabetes type 2. Diabetes is more likely to be had if there is a family history of diabetes, overweight, physical inactivity, race, and health problems such as high blood also increase the chance of contracting diabetes.

Doctors define a stage known as the prediabetes stage when blood sugar levels are between 100 to 125 milligrams per deciliter of blood. This basically means that sugar level is higher than normal but not as high as it should be for diabetes. The prediabetes and type 2 diabetes have the same risk factors and patients suffering from this do not have the same symptoms as patients suffering from diabetes.

Insulin resistance

Whilst the reason for diabetes type 1 is not known it properly explained by doctors the reason for diabetics type 2 is and it is caused by insulin resistance. Occurs when cells in the body resist, refuse or ignore the signal that the hormone insulin us trying to send out, that is to grab glucose out of the blood stream and put them into our cells. Glucose which is also called blood sugar, is the main of source of energy in the body. Glucose is gotten from the food we eat as food is broken down into glucose and other necessary substance during digestion but the main source of glucose is carbohydrates.

Causes of insulin resistance

As far as genetics, aging and ethnicity play huge roles in the development of insulin sensitivity, the main cause of insulin resistance is basically overweight, lack of exercise, smoking, and even skimping in sleep.

As resistance to insulin develops in the body the body also fights back by producing more insulin. As years and months pass by since insulin resistance silently began in the body the cells responsible for the production of insulin becomes worn out due to being over worn trying to meet body demand for insulin, then years later, blood sugar finally begins to rise and diabetes type 2 is developed in the body. You may also develop non-alcoholic fatty liver disease which is also associated with insulin resistance and causes a high risk of developing heart and liver diseases.

Symptoms of insulin resistance

Insulin resistance is typically triggered by a series of factors some of which are weight, age, genetics, being sedentary and smoking.

A large waist: medical experts basically say that the best way to determine if you are at risk for insulin resistance is a tape measure and a moment of truth before the mirror. A waist that measures 35 inches or more in women and 40 or more in men increases the odds for insulin resistance.

Metabolic syndrome: according to the NHS, if you have one or two of the following listed below to most likely have metabolic syndrome which increases the risk of insulin resistance.

* High triglycerides: level of 150 or higher, or taking medication to treat high blood fats.
* Low HDLs: low density lipoprotein levels of below 50 for women and 40 for men, or taking medication to in HDL levels.
* High blood pressure: diagnosed with a high blood pressure or taking medication to maintain blood pressure level.
* High blood sugar: having sugar level within the prediabetes or diabetes range
* High fasting blood sugar: mildly high blood sugar which may be a sign of early diabetes.

Developing dark skin patches: if insulin resistance becomes severe skin changes are visible. These include patches of dark skin in the back of the neck, knees, elbows, knuckles or armpits.

Insulin resistance can also be gotten through prolonged periods of fast, fasting for 36 hours is one to get insulin resistance faster.

Health condition related to insulin resistance

According to the international diabetes federation, thanks to years of high insulin resistance followed by an onslaught of high blood sugar, people with insulin resistance, prediabetes and type 2 diabetes usually are at risk of contracting cardiovascular disease. Insulin resistance doubles the risk for heart attack and stroke and triples the odds that the heart attack or brain attack will be deadly.

In research by the oxford university, it also shows a string correlation between insulin resistance and memory function decline increasing the risk for Alzheimer’s disease.

How to prevent insulin resistance

Losing weight and exercising regularly reduces chances of insulin resistance by a lot. Also getting a lot of sleep plays a huge factor in increasing insulin sensitivity. In a study by University of New Mexico School of Medicine, overweight people that lost 10% of their weight through exercise and diet saw sensitivity to insulin increase by a whopping 80%, whilst others who lost the same amount of weight through diet alone saw a 38% increase only in sensitivity to insulin and those who didn’t lose much weight and got more exercise saw almost no shift in their sensitivity to insulin. So, it is advisable to not rely on exercising only.

Symptoms of diabetes

* Early signs of diabetes
* Hunger and fatigue: this occur because the body converts the food eaten into glucose for the cells to use almost immediately, this causes severe tiredness or fatigue and hunger.
* Peeing: on average an individual pees no more than four to five in a day but diabetes patients go more because the body reabsorbs glucose and passed it through the kidneys, but when diabetes pushes the blood sugar up, the kidneys may not be able to bring it all back in. This causes more urine to be produced in the body and that takes fluids. This leads to a circular effect in which you drink more and pee more.
* Dry mouth and itchy skin: because peeing more often because the norm and more fluids are used for that, there is little or no moisture left for the body to keep the body hydrated causing dehydration and the mouth to feel very dry. Dry skin can make you cry itchy.
* Blurred vision: changing fluid levels in the body could make the lenses in the eyes swell up. They can’t shape and they can’t focus as a result of this.
* Diabetes type 1
* Unplanned weight loss: if the body can’t get its energy from food, it starts burning muscle and fat instead. Weight loss becomes inherent because of this even though feeding habits has not changed.
* Nausea and vomiting: when your body starts resorting to burning fat, it makes ketones. This can build up in the blood to dangerous levels, and possibly cause a life-threatening chronic disease called diabetic ketoacidosis. Ketones make you feel sick to the stomach.
* Diabetes type 2
  + Yeast infection: both men and women with diabetes are susceptible to this disease. Yeast feeds on glucose, so plenty around makes it thrive. Infections can grow in any warm, moist fold of skin including between the fingers and toes, under breast, and in or around sexual organs.
  + Slow healing sore or cuts: blood sugar van affect the flow of the blood in the body and cause nerve damage making it hard for wounds to heal quickly.
  + Pain or numbness in the feet.
* Gestational diabetes

High blood sugar during pregnancy has no symptoms.

Health diseases that diabetes can cause

* Heart diseases: after diabetes as persisted for some time it is advisable to go in for heart checks as it is generally and publicly announced by every health body that diabetes increases chances of stroke and heart diseases.
* Loss of feeling: numbness in the body as diabetes causes severe nerve damages
* Looking after your feet: cause as discussed earlier diabetes causes infections, and reduces blood supply to on the feet. This means that foot injuries won’t heal well and if not properly checked daily can cause sore or injuries to fester. These problems lead to ulcer and infections.
* Eyes checks: the eyes should be checked every year for damage to blood vessels, which can cause sight problems and blindness.
* High blood pressure.

The national Diabetes Statistics Report is a publication of the Centers for Disease Control and Prevention (CDC) in the United States that provides updated statistics about diabetes for a scientific audience. It usually includes information on prevalence and incidences of diabetes, prediabetes, risk factors, acute and long-term complications, death, and costs. An estimated 33.9% of U.S. adults aged 18 years or older (84.1 million people) had prediabetes in 2015, based on their fasting glucose or A1C level. Nearly half (48.3%) of adults aged 65 years or older had prediabetes. Diabetes was the seventh leading cause of death in the United States in 2015.

Secured wireless body area network supports a lot in developing predictive analysis system in health care. Big data was put forward for the first time in 2009, and since then has found application in the field of multiple business and development, especially its mature usage in medical field. In big data, users put forward higher request to the storage service on the availability, reliability and durability of data. Map Reduce based distributed computing framework works efficiently with machine learning algorithms in training the distributed data blocks.

With the advent of social media, Internet of Things and other data sources handling the huge volumes of data has become an ever-growing challenge as more information is being put on the internet even at this moment and due to the wide range of existing data types wider classification and formats of data now exists. From the literature it is clear that an intelligent cloud-based cluster model is effective and efficient for managing the huge volume of data. Due to the availability of huge volume of data from different data sources distributed computing-based cluster along with cloud environment supports in effective handling of data .From literature it is clear that the disease detection and surveillance systems provide epidemiologic intelligence that allows all the persons in the healthcare to deploy preventive measures and help clinic and hospital administrators make automatic and intelligent decisions.

# Chapter 3

**Methodology**

The data set used, data cleaning process, feature selection techniques, and the machine learning algorithms used are discussed in this chapter.

## Data Set

To create a predictive analysis system a comprehensive dataset to learn from and a test dataset that ensure that the machine learning algorithm reaches the expected accuracy standards. Accuracy of classification and clustering algorithms is very crucial and plays a major role in health care systems. In order to develop an accurate Diabetes prediction model the dataset used in this research includes data gather from 130 US hospitals and discusses over 50 features representing patients and hospital outcomes and it includes data from 101, 766 patients. The dataset contains attributes that discusses the drugs administered to the patients which can be very useful for predicting diabetes type, it also includes attributes on ethnicity as discussed earlier in this paper that race as an effect on the dominance of diabetes. Age which also plays a factor is included in the dataset. From the set of available data, a split is performed to test and train data, which is a 50% split i.e., 50% of the data is used for test whilst the remaining 50% is used for training data. The summary function in R was used to find the statistical analysis of the data which includes the mean median and mode function of the data.

## Feature Selection

As discussed earlier 50 features are available in the dataset, and dataset is also highly diversified due to huge number of patients considered in the dataset. not all available attributes will contribute to the prediction of patient’s diabetes condition therefore feature selection is used to choose from among the many attributes available. Including the less important features in prediction is time consuming and reduces efficiency of the system. Different statistical techniques have been discussed in literature but the function used in this paper is the R correlation function known as Caret() that helps remove redundant feature and also ranks said feature in order of importance.

### Caret() R

The caret function in R is a function that helps remove redundant features based on their correlation with each other that is similarities between the features. The caret package provides the findCorrelation which analyzes a correlation matrix of the dataset’s attributes and reports of which attributes can be removed.

In this paper dataset was first normalized using the factor function in R before being passed into the caret function and a seed of 10 was set. The importance of the features was also determined using the traincontrol method.

## Machine Learning Algorithms

### **Prediction** **Algorithm**

It’s used to make decisions. It is a type of supervised learning that uses historical and existing data to forecast likely future outcomes. It works on a number of existing data learns from them and makes future forecasts. Prediction models are regularly updated to suit changes to enhance accuracy of the models. Predictive models make decisions or forecasts based on what is happening now and what has happened in the past. Prediction models finish their calculations in real time.

There are five types of prediction algorithms but only two are used in this research that is the most commonly used clustering and classifying models.

#### Clustering Model

This is an unsupervised learning that is used to find similarity between various attributes in dataset. After data is segmented various clusters are produced, all objects in a cluster are similar. In data mining clusters are used to find similar data attributes.

##### Simple Kmeans

The most popular and widely used clustering algorithm is the Kmeans algorithm. Kmeans clusters datasets based on a number of clusters K after calculating the distance between each attribute and variables. Kmeans basically works in 3 steps which are:

* Selecting the K values
* Initializing the centroid
* Select group and finds the average

There are two methods to select the right value for K which is namely the Elbow method and Silhouette method. The elbow method is an empirical method for selecting the right value for K it simply does this by picking the range of values and selecting the best among them. It calculates the sum of the square of the points and finds the average distance.

The formula above is the formula for calculating the elbow method.

The silhouette method like the elbow method also picks up the range of values for K but unlike the elbow method it draws the silhouette graph for these values i.e., it calculates the average distance of points within its cluster and the average distance to the closest cluster.

Therefore, its formula becomes

Where a represents distance inside cluster and b represents distance to the nearest other cluster.

For better and accurate results another clustering algorithm was used to build compares between both clustering algorithms.

##### Hierarchical Cluster

In this type of clustering clusters are built by forming hierarchies between the various attributes present in a dataset. There are two types of hierarchical clusters and they include

* Agglomerative: clusters are built from bottom-up, i.e., the algorithm continues merging the dataset until desired clusters is formed. It has a time complexity of and memory requirement of which makes it slow for even a simple dataset. Its method of merging is usually greedy.
* Divisive: this typically starts with one cluster and then divides into various clusters possible. This follows a top-bottom approach. It has time complexity of .

#### Classification Model

It’s the simplest model and it classifies attributes based on simple query questions and answers an example is “Is this patient on diabetic medication?”

The classifiers used in this paper is Random Forest, Naïve Bayes and J48 decision trees.

##### Random Forest

Is easy to implement with R and it’s used by calling the library random forest and mlbench and caret R package libraries respectively. Random Forest is based on decision trees and it uses the regression technique for its predictions, it combines various classifying techniques to arrive at its decisions this is possible because of its utilization of ensemble learning. A random forest consists of many decision trees and the ‘forest’ it generates its trained through bootstrapping or bagging aggregating. It also establishes it outcomes based on the decision tree; it uses the mean output of its decision trees. The more tree it generates the higher the accuracy of its prediction.

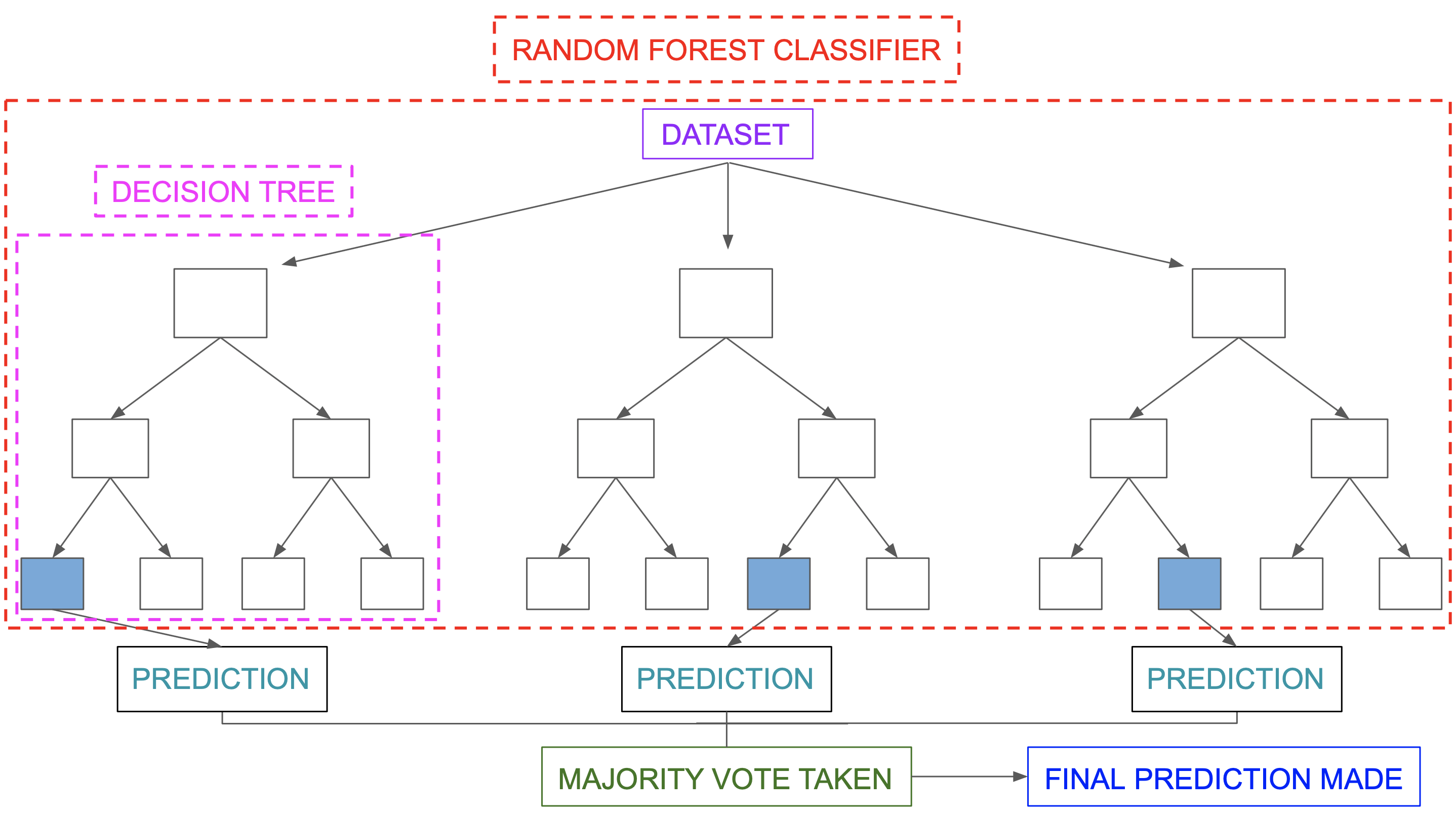


Fig 1. Showing how random forest makes predictions.

##### J48

Also known as C4.5 makes its decisions by a top-down, recursive, divide and conquer strategy as it searches for a new instance to split for at each stage of the tree. In J48 accuracy of the algorithm can be expanded by fitting i.e., the attribute extreme standardized data gained is utilized. The minor subsets are returned by the algorithm. The split strategies stops if a subset has a place with a similar class in all the instances. It develops a decision node utilizing the expected estimations of the class and can deal with particular characteristics, lost or missing attribute estimations of the data and varying attribute costs.

##### Naïve Bayes

It is the most popular and most widely used prediction and classification algorithm. It uses probability in making its predictions based on applying the Bayes’ theorem and they have strong independence assumptions between the features that is it’s termed as naïve. It is simplest of the Bayesian network. It is highly scalable and requires a number of parameters linear in the number of variables to learn a problem. It is a conditional probability model and its formula is

This equation simply means

In practice there is interest only in the numerator of the fraction, because the denominator does not depend on C and the values of the features are given, so the denominator is effectively constant. This equation is combined with a decision rule when applied for classification such that it picks the most probable hypothesis.

# Chapter 4

**Results and Experiments**

#### Data description and Source

The dataset used in this research was gotten from an open-source data depository known as Kaggle. 50 attributes are present in dataset and that includes Encounter ID, Patient Number, Race etc. the data was collected 130 hospitals across the UK.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature Name | Type | Description | % Missing | Values |
| Encounter ID | Numeric | Describes meeting with a patient. | 0 |  |
| Patient Number | Numeric | Used to uniquely identify a patient | 0 |  |
| Race | Character | Describes the ethnicity of patients | 2 | Caucasian, Asian, African American, Hispanic and other |
| Gender | Character |  | 0 | Male, female |
| Age | Character | Grouped in 10-year intervals | 0 | 0-10, 10-20,…,90-100 |
| Weight | Numeric | Weight is in pounds | 97 |  |
| Admission type | Character | Shows the condition on which patient was attended to | 0 | Emergency, urgent, elective, newborn, and not available |
| Discharge disposition | Character | Under which condition was patient discharged from hospital | 0 | 29 distinct values, for example, discharged to home, expired, and not available |
| Admission Source | Character | Did you come in to the department yourself or where transferred there from elsewhere | 0 | 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital |
| Time in Hospital | Numeric | How long where you admitted for or number of hours spent in the hospital | 0 | time |
| Payer code | Character | How were hospital bills paid? | 52 | 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay |
| Medical Specialty | Character | Showing the specialization of attending doctor | 53 | Medical specialty Nominal Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon |
| Number of lab procedures | Numeric | Number of lab tests performed during your stay at the hospital | 0 |  |
| Number of procedures | Numeric | Apart from lab test any other procedures performed on patient during that encounter | 0 |  |
| Number of medications | Numeric | Number of distinct generic names administered during the encounter | 0 |  |
| Number of outpatient visits | Numeric | The number of patients in the year preceding the encounter | 0 |  |
| Number of emergency visits | Numeric | Number of emergency visits of the patient in the year preceding the encounter | 0 |  |
| Number of inpatient visits | Numeric | Number of inpatient visits of the patient in the year preceding the encounter | 0 |  |
| Diagnosis 1 (diag\_1) | Character | The primary diagnosis (coded as first three digits of ICD9) | 0 | 848 distinct values |
| Diagnosis 2 (diag\_\_2) | Character | Secondary diagnosis (coded as first three digits of ICD9) | 0 | 923 distinct values |
| Diagnosis 3 (diag\_3) | Character | Additional secondary diagnosis (coded as first three digits of ICD9) | 1 | 954 distinct values |
| Number of diagnoses | Numeric | Number of diagnoses entered to the system | 0 |  |
| Glucose serum test result | Character | Indicates the range of the result or if the test was not taken. | 0 | >200, >300, normal, and none |
| A1c test result | Character | Indicates the range of the result or if the test was not taken | 0 | >8, >7, normal, and none. |
| Change of medication | Character | if there was a change in diabetic medications (either dosage or generic name). | 0 | Change and no change |
| Diabetes medications about 24 different attributes for the various diabetic medications | Character | Indicates if there was any diabetic medication prescribed the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. | 0 | Up, down, and steady for when dosage remained the same. |
| Readmitted | Character | if the patient was readmitted in more than 30 days, and “No” for no record of readmission. | 0 | 30, no |

Table 1. showing data attributes and their descriptions.

Summary of dataset using R summary function is shown in table 2 below.

#### Experimental Outcomes

When working on the dataset all experiments were done using R studio and then added to an R dashboard.

##### Expected Outcomes

I expected before performing the clustering algorithm and applying it on the dataset, I expected for cluster results to predict that diabetes was more prominent in adults between the age of 90 has from literature it shows that most adults typically get diagnosed a 70. I expected its precision and accuracy to be higher than that of Hierarchical cluster as their means of clustering are different. I had an expectation that patients referred to the diabetes which is made known via the admission source would mostly be diagnosed with diabetes and have their medication changed, however this was not the case as is shown in the diagrams below. I also expected Asians to have the leading number of diabetic patients as it is earlier discussed in this paper that ethnicity also plays a role in diabetic distribution but the people with the leading number of diabetic patients are Caucasians.

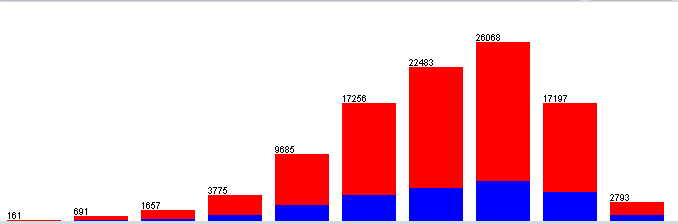


Fig 2. Showing the age distributions of diabetic patients.

The age group with the least case of diabetic patients being the 0-10 and the age group with the most diabetic patients is 0-80. Type 2 diabetics is the most common type of diabetics and it is usually almost never diagnosed early.

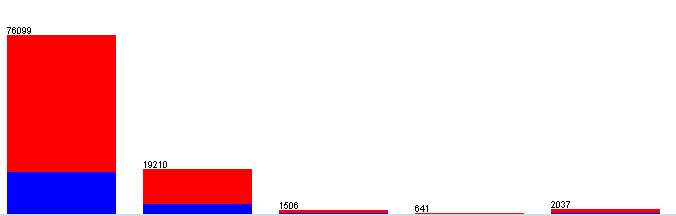


Fig 3. Showing the distribution of diabetic patients by race with Caucasians leading the way and Asians having the least number of diabetic patients.

##### Clustering algorithms

Models were built based on the attribute diabetics medications.

|  |  |
| --- | --- |
| algorithms | Results |
| Simple Kmeans | Surprisingly the results showed that it is most dominantly found in adults between the ages of 70-80 years of age after reading through literature again to confirm results it is said that most diabetic patients typically die before the age of 90(WHO, 2012). Its correctly classified instances are 89%. percentage splits of 50% was performed on the algorithm and little difference was shown in its correctly classified instances. |
| Hierarchical Cluster | When this was performed on the dataset its correctly classified instances were 87 %. Percentage splits was also performed on the dataset at 50% and its correctly classified instance dropped to 78%. |

Table 3. showing the hierarchical clustering and simple Kmeans clustering results

##### Classification Algorithms

Models were also built based on the attribute diabetics medication. Cross validation of 10 folds was done on the dataset and a 50% percentage split was also performed.

|  |  |
| --- | --- |
| Algorithms | Results |
| Random Forest | Its correctly classified instances were 87% using training set. After a cross validation fold of 10was applied the number of correctly classified instances were 85%, and after a percentage split of 50% the number of correctly classified instances was 87% |
| J48 | Its correctly classified instances were 88% using training set. After a cross validation fold of 10 was applied the number of correctly classified instances were 86%, and after a percentage split of 50% the number of correctly classified instances was still 70% |
| Naïve Bayes | Its correctly classified instances were 89% when training set was used on the data. After a cross validation of 10-fold was applied its number of correctly classified instances were 84%, and after a percentage split of 50% its correctly classified instances were still 85%. |

Table 4. discussing the classification algorithms

When algorithm was run in Weka

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | TPR | FPR | Precision | Recall | F-Measure |
| Random Forest | 1.000 | 0.039 | 0.979 | 0.979 | 0.976 |
| J48 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Naïve Bayes | 0.989 | 0.036 | 0.989 | 0.989 | 0.989 |

Time taken for J48 5.47 seconds. Time taken to build model 0.32 seconds Naïve Bayes. Time taken for Random Forest 6.8 seconds.

##### Result Analysis

From the results shown in the experiments conducted above literature research is proven to be true as dataset shows that Random Forest takes longer to build model than the other classification algorithms used. From literature we know Random Forest build every tree by instance and its precision is determined by the amount of trees it builds. It took the longest time to build and its precision value was not much different from the others. From the clustering algorithms we see that the number of correctly clustered instances are high and very low when run on Weka, Simple Kmeans had a 58% correctly clustered instances and hierarchical cluster was the same.

#### Challenges

To compare experimental results and ensure I’ve not made a mistake in coding the same algorithms were applied on dataset in Weka there was a slight difference in results which was first a cause for alarm but after checking dataset again I realized I was working with a different data set, in R all values were converted to nominal giving a much lower precision rate than in Weka where some values were left as numerical attributes and other nominal values. I had memory capacity limitations as dataset is very huge and running on my machine proved to be difficult and algorithm might have been slower to build due to that. I also had challenges converting the variables to nominal value in R.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data | Random Forest | J48 | Naïve Bayes | Simple Kmeans | Hierarchical Clustering |
| Building time | 6.8 | 5.47 | 0.32 | 0.33 | 0.34 |
| Accuracy(%) |  | 88 | 89 | 89 | 87 |

# Chapter 5

**Recommendations and Conclusion**

The field of data analytics and data mining uses algorithms to extract patterns in datasets and use them to make a prediction or forecast on what the situation that dataset records will look like in the future. This study discusses diabetes prediction using machine learning techniques. As is already discussed in earlier chapters some obtained results when outside of expected result outcomes not in a negative way but because of a bias. It is recommended that other machine learning techniques be explored in predicting diabetes and the diabetic types also be predicted based on dataset. The developed models each had their strengths and weaknesses which has already been discussed in the above sections. Finally, data mining has seen its useful well lived in the health sector most especially but other sectors as well.

# References

1. Song, T.M., 2012. Efficient utilization of big data on healthcare and welfare area. In *Healthc Welf Forum* (Vol. 193, pp. 68-76).
2. Kumar, S. and Chakravarty, A., 2015. ABC–VED analysis of expendable medical stores at a tertiary care hospital. *Medical journal armed forces india*, *71*(1), pp.24-27.
3. Huang, W., Wang, H., Zhang, Y. and Zhang, S., 2019. A novel cluster computing technique based on signal clustering and analytic hierarchy model using hadoop. *Cluster Computing*, *22*(6), pp.13077-13084.
4. Saravanan, N. and Gayathri, V., 2019. A PERFORMANCE OF CLASSIFICATION ACCURACY IN MODIFIED J48 ALGORITHM USING SPEARMAN’S. *Infokara Research*, *8*, pp.98-107.
5. Mazid, M.M., Ali, S. and Tickle, K.S., 2010, February. Improved C4. 5 algorithm for rule based classification. In *Proceedings of the 9th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases* (pp. 296-301). World Scientific and Engineering Academy and Society (WSEAS).
6. Chary, S.N. and Rama, B., 2017. A survey on comparative analysis of decision tree algorithms in data mining. In *International Conference On Innovative Applications In Engineering and Information Technology (ICIAEIT-2017)* (Vol. 3, No. 1, pp. 91-95).
7. Nikam, S.S., 2015. A comparative study of classification techniques in data mining algorithms. *Oriental journal of computer science & technology*, *8*(1), pp.13-19.
8. Joshi, S., Pandey, B. and Joshi, N., 2015. Comparative analysis of Naive Bayes and J48 Classification Algorithms. *International Journal of Advanced Research in Computer Science and Software Engineering*, *5*(12), pp.813-817.
9. Patil, T.R., 2013. MSSS Performance analysis of naive bayes and J48 classification algorithm for data classification. Intl. *Journal of Computer Science and Applications*, *6*(2).
10. <https://www.futurelearn.com/info/courses/data-mining-with-weka/0/steps/25391>
11. Saravana, N. and Gayathri, D.V., 2018. Performance and classification evaluation of J48 algorithm and Kendall’s based J48 algorithm (KNJ48). *Int. J. Comput. Trends Technol.(IJCTT)--Volume*, *59*.
12. <https://www.netsuite.com/portal/resource/articles/financial-management/predictive-modeling.shtml>
13. <https://machinelearningmastery.com/clustering-algorithms-with-python/>
14. <https://www.guru99.com/r-k-means-clustering.html>
15. <https://www.journaldev.com/46263/factor-in-r>
16. <https://www.journaldev.com/46265/categorical-variables-in-r>
17. <https://stat.ethz.ch/R-manual/R-devel/library/foreign/html/read.arff.html>
18. <https://machinelearningmastery.com/feature-selection-with-the-caret-r-package/>
19. <https://www.geeksforgeeks.org/decision-tree-in-r-programming/>
20. <https://www.geeksforgeeks.org/clustering-in-r-programming/>
21. Rimal, Y., 2021. Reproducible Academic Writing and Interactive Data Visualization Using R Markdown (R Programming Flex-Dashboard: Flex\_Dashboard Packages). In *Rising Threats in Expert Applications and Solutions* (pp. 603-615). Springer, Singapore.
22. Tippmann, S., 2015. Programming tools: Adventures with R. *Nature News*, *517*(7532), p.109.
23. Konda, P., Kumar, A., Ré, C. and Sashikanth, V., 2013. Feature selection in enterprise analytics: a demonstration using an r-based data analytics system. *Proceedings of the VLDB Endowment*, *6*(12), pp.1306-1309.
24. Prajapati, V., 2013. *Big data analytics with R and Hadoop*. Packt Publishing Ltd.
25. <https://www.kaggle.com/brandao/diabetes?select=diabetic_data.csv>
26. <https://www.cdc.gov/features/diabetes-statistic-report/index.html>
27. Huang, W., Wang, H., Zhang, Y. and Zhang, S., 2019. A novel cluster computing technique based on signal clustering and analytic hierarchy model using hadoop. *Cluster Computing*, *22*(6), pp.13077-13084.
28. Bakshi, S., Sa, P.K., Wang, H., Barpanda, S.S. and Majhi, B., 2018. Fast periocular authentication in handheld devices with reduced phase intensive local pattern. *Multimedia Tools and Applications*, *77*(14), pp.17595-17623.
29. Yuvaraj, N. and Sabari, A., 2016. Performance analysis of supervised machine learning algorithms for opinion mining in e-commerce websites. *Middle-East J. Sci. Res*, *1*(1), pp.341-345.
30. Haindl, M., Somol, P., Ververidis, D. and Kotropoulos, C., 2006, November. Feature selection based on mutual correlation. In *Iberoamerican Congress on Pattern Recognition* (pp. 569-577). Springer, Berlin, Heidelberg.
31. <https://www.webmd.com/diabetes/guide/understanding-diabetes-symptoms>
32. <https://www.endocrineweb.com/conditions/type-2-diabetes/insulin-resistance-causes-symptoms>
33. Timonen, M., Salmenkaita, I., Jokelainen, J., Laakso, M., Härkönen, P., Koskela, P., Meyer-Rochow, V.B., Peitso, A. and Keinänen-Kiukaanniemi, S., 2007. Insulin resistance and depressive symptoms in young adult males: findings from Finnish military conscripts. *Psychosomatic Medicine*, *69*(8), pp.723-728.
34. Schwartz, M.W. and Kahn, S.E., 1999. Insulin resistance and obesity. *Nature*, *402*(6764), pp.860-861.
35. Clark, N.G., Fox, K.M. and Grandy, S., 2007. Symptoms of diabetes and their association with the risk and presence of diabetes: findings from the Study to Help Improve Early evaluation and management of risk factors Leading to Diabetes (SHIELD). *Diabetes Care*, *30*(11), pp.2868-2873.
36. Ramachandran, A., 2014. Know the signs and symptoms of diabetes. *The Indian journal of medical research*, *140*(5), p.579.
37. <https://www.nhs.uk/conditions/type-2-diabetes/health-problems/>
38. Centers for Disease Control and Prevention, 2020. National diabetes statistics report, 2020. *Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services*, pp.12-15.
39. <https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf>
40. World Health Organization, 2004. Diabetes action now: an initiative of the World Health Organization and the International Diabetes Federation.
41. Wendland, E.M., Torloni, M.R., Falavigna, M., Trujillo, J., Dode, M.A., Campos, M.A., Duncan, B.B. and Schmidt, M.I., 2012. Gestational diabetes and pregnancy outcomes-a systematic review of the World Health Organization (WHO) and the International Association of Diabetes in Pregnancy Study Groups (IADPSG) diagnostic criteria. *BMC pregnancy and childbirth*, *12*(1), pp.1-13.
42. Harris, M.I., Eastman, R.C., Cowie, C.C., Flegal, K.M. and Eberhardt, M.S., 1997. Comparison of diabetes diagnostic categories in the US population according to 1997 American Diabetes Association and 1980–1985 World Health Organization diagnostic criteria. *Diabetes care*, *20*(12), pp.1859-1862.
43. TIME, G., 2016. ON DIABETES.
44. Oza-Frank, R., Ali, M.K., Vaccarino, V. and Narayan, K.V., 2009. Asian Americans: diabetes prevalence across US and World Health Organization weight classifications. *Diabetes care*, *32*(9), pp.1644-1646.