

**SPRING 2020: BAN 693 CAPSTONE PROJECT**

# **Gateway for Disease Control and Prevention**

By

Bhargavi Sankula (NetID: qv4668)

Pooja Bhansali (NetID : gs6262)

**Advisor**

**Dr. Jiming Wu**

**California State University, East Bay**

**MS Business Analytics**

# Table of Contents

CHAPTER 1:Introduction .....	1
CHAPTER 2:Objective / Goal of the work .....	3
CHAPTER 3:Study of Diabetes Classification.....	4
3.1 Data Acquisition and Data extraction .....	4
3.2 Tools used to analyze the dataset.....	5
3.3 Data description .....	5
3.4 Statistical metrics .....	6
3.5 Data profiling .....	7
3.5.1 Distribution of outcome variable .....	7
3.5.2 Distribution of independent variables .....	8
3.6 Data Cleansing / Data filtering .....	8
3.7 Data Partitioning .....	10
3.8 Modeling .....	10
3.8.1 Baseline model.....	11
3.8.2 k-Nearest Neighbors (k- NN).....	12
3.8.3 Logistic Regression.....	13
3.8.4 Decision Trees .....	14
3.8.5 Random Forest .....	15
3.8.6 Neural Networks .....	16
3.9 Evaluation of the models implemented.....	19
CHAPTER 4:Business model for predicting the risk of diabetes in real world .....	20
CHAPTER 5: Study of heart disease classification.....	23
5.1 Data Acquisition and Data extraction .....	23
5.2 Tools used to analyze the dataset.....	23
5.3 Data description .....	23
5.4 Attributes.....	24
5.5 Statistical metrics .....	25

5.6 Data profiling .....	26
5.6.1 Distribution of Outcome Variable .....	26
5.6.2 Distribution of independent Variables .....	26
5.7 Data Cleansing / Data filtering .....	27
5.8 Data Partitioning .....	29
5.9 Modeling .....	29
5.9.1 Baseline model.....	30
5.9.2 Logistic regression .....	30
5.9.3 Decision Tree .....	31
5.9.4 k-Nearest Neighbors (k- NN).....	32
5.9.5 Neural Network.....	33
5.10 Comparison of models .....	34
5.11 Best model for heart disease is LOGISTIC REGRESSION.....	35
5.12 Cross validation score of the best model .....	36
 CHAPTER 6 :Applications and Findings .....	 37
 CHAPTER 7:CONCLUSION .....	 38
 References .....	 39



## **CHAPTER 1**


### **Introduction**

According to the definition of health, it is a state of complete physical, mental and social wellbeing in which diseases and infirmity are absent. Every part of life relies on the good health of an individual and it is a center for everything that follows. However, due to various factors in the lifestyle which include both physical and mental conditions may disturb the health cycle and can cause diseases to humans which will affect life expectancy. These diseases can be of two types i) Acute and ii) Chronic.

Acute illness generally develops and lasts for a short time often only a few days or weeks. Chronic conditions (diabetes & heart diseases etc.) develop slowly and may worsen over an extended period – months to years. Unfortunately, there are few chronic conditions that are affecting humans since past few years and a permanent lifestyle changes, dietary habits need to be followed to overcome or prevent them.

Diabetes is a chronic condition that affects how our body turns food into energy. There is no cure yet for diabetes but losing weight, eating healthy food and regular exercise can really help. Diabetes can cause problems during pregnancy for women and their developing babies. Proper healthcare before and during pregnancy can help prevent birth defects and other health problems. People with a chronic condition of diabetes are more likely to develop a heart disease as diabetes can damage the blood vessels and nerves that controls the heart.

As per World Health Organization statistics, 12 million deaths occur worldwide every year due to heart diseases. The early prognosis of heart disease can aid in making decisions on lifestyle



changes in high risk patients and in turn reduce the complications. Prediction of heart disease is regarded as one of the most important subjects in the section of clinical data analysis.

We would like to apply our theoretical knowledge that we learnt through this program and develop a proof of concept to build an application to identify risk of diabetes for women with pregnancy and help them to control diabetes through eating a healthy diet and exercising regularly. This application is a preliminary step to start their journey without / controlled diabetes during the pregnancy. They can seek advice from their primary physician on the required steps to be followed.

We will be applying statistics, data visualization, data mining, machine learning and deep learning algorithms to analyze the data in hand and implement various models to classify the outcome variable of chronic disease condition. We would like to help pregnant women to early detect diabetes risk and advise them to contact their physician to stay fit and healthy.



## **CHAPTER 2**

### **Objective / Goal of the work**

Pregnancy is an important phase in any woman's life and there are various stages that require utmost care to have a balanced lifestyle and develop a healthy baby. Gestational diabetes is first seen in a pregnant woman who did not have diabetes before she was pregnant. It usually shows up in the middle of pregnancy and doctors most often test for it between 24 and 28 weeks of pregnancy. Often gestational diabetes can be controlled through healthy foods and regular exercise.

We would like to help mothers by answering simple questions to figure out their risk of diabetes during pregnancy using an optimized model to classify the outcome. This hypothetical model / application will help women to make decisions to improve their lifestyle, food habits and exercise routine before consulting with a doctor. And, the early prognosis of heart diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications.

We will further study the most relevant/ risk factors of heart disease as well as predict the overall risk using various models and try to derive conclusions to help diabetic patients to avoid further risk of developing a heart disease.

The hypothetical model for classification of diabetes risk is built from a given dataset of pregnant women, will have a proof of concept developed using python web framework to fetch the inputs (answers from end users) and identify the risk of diabetes. This can be enhanced into a business model with the help of large real time dataset information and a wide scale application (mobile app or iWatch app or a website) to deploy it under healthcare management.

## **CHAPTER 3**

### **Study of Diabetes Classification**

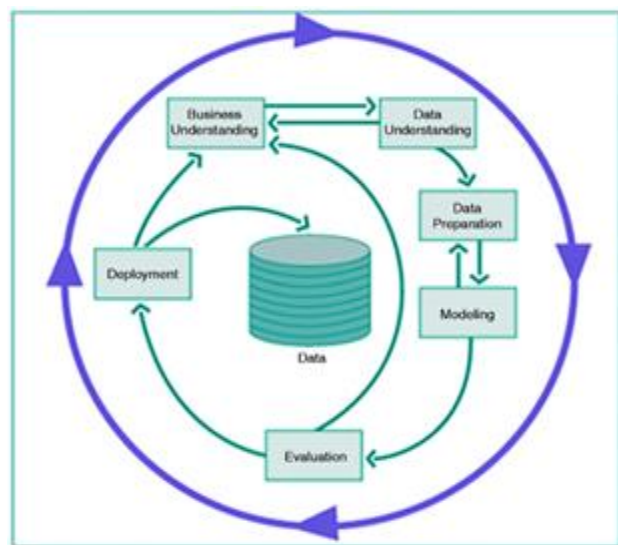
#### **3.1 Data Acquisition and Data extraction**

All the data that we collected will be helpful to achieve our objective of classifying the outcome variable. The appropriate interpretation of the data highly depends on our understanding of the problem that we are trying to solve and its application.

The dataset of pregnant women can be found at Kaggle using the [link](#). This same data set is used to develop various models and to classify the outcome variable as Non-Diabetic (0) or

Diabetic (1) respectively. This collected data can be used for preprocessing, data modelling, model evaluation and model deployment accordingly.

The below diagram describes a high-level view of the entire business model process.



### **3.2 Tools used to analyze the dataset**

We used the open source programming tool Python 3. We also made use of the inbuilt libraries present in python to do all the required analysis. Jupyter notebook is used to do all the required coding.

- Pandas library is used to read our data from a csv file and to do further study.
- Numpy is used to covert data as required that is suitable for our classification model.
- Seaborn and matplotlib are mainly used for visualization.
- Several other libraries are used as required for model building activities

### **3.3 Data description**

In this section, we do the analysis of individual variables using statistical metrics, relations between pairs of variables and visualizing techniques for more complex relationships between the variables.

The acquired dataset contains 15000 records and the following variables to classify whether the pregnant woman has diabetes or not.

- PatientID: This is a unique number generated for each patient
- Pregnancies: Number of times a woman is pregnant
- PlasmaGlucose: Measures a person's blood sugar level after fasting or not eating anything for at least 8 hours.
- DiastolicBloodPressure: The minimum arterial pressure during relaxation and dilation of ventricles of the heart (mm Hg)
- TricepsThickness: Triceps skin fold thickness (mm)



- SerumInsulin: Measures the amount of insulin in the blood (mu U/ml)
- BMI: Body Mass Index (weight in kg/ (height in m)<sup>2</sup>)
- DiabetesPedigree: This value provides a synthesis of diabetes mellitus history in relatives and genetic relationship of those relatives.
- Age: Age in years
- Diabetic: Classification variable (0 if non-diabetic, 1 if diabetic)

Below screenshot represents few records from the Diabetes dataset:

	PatientID	Pregnancies	PlasmaGlucose	DiastolicBloodPressure	TricepsThickness	SerumInsulin	BMI	DiabetesPedigree	Age	Diabetic
0	1354778	0	171	80	34	23	43.509726	1.213191	21	0
1	1147438	8	92	93	47	36	21.240576	0.158365	23	0
2	1640031	7	115	47	52	35	41.511523	0.079019	23	0
3	1883350	9	103	78	25	304	29.582192	1.282870	43	1
4	1424119	1	85	59	27	35	42.604536	0.549542	22	0

### 3.4 Statistical metrics

Analysis and evaluation of these metrics will help us to prepare the data for modelling. The evaluation of low mean values, high standard deviation, minimum, maximum and mean will help us to identify any outliers or low potential for classifying the outcome. Hence, we removed the patientID column from the data set as it is of no use and a unique number.

	count	mean	std	min	25%	50%	75%	max
Pregnancies	15000.0	3.224533	3.391020	0.000000	0.000000	2.000000	6.000000	14.000000
PlasmaGlucose	15000.0	107.856867	31.981975	44.000000	84.000000	104.000000	129.000000	192.000000
DiastolicBloodPressure	15000.0	71.220667	16.758716	24.000000	58.000000	72.000000	85.000000	117.000000
TricepsThickness	15000.0	28.814000	14.555716	7.000000	15.000000	31.000000	41.000000	93.000000
SerumInsulin	15000.0	137.852133	133.068252	14.000000	39.000000	83.000000	195.000000	799.000000
BMI	15000.0	31.509646	9.759000	18.200512	21.259887	31.767940	39.259692	56.034628
DiabetesPedigree	15000.0	0.398968	0.377944	0.078044	0.137743	0.200297	0.616285	2.301594
Age	15000.0	30.137733	12.089703	21.000000	22.000000	24.000000	35.000000	77.000000
Diabetic	15000.0	0.333333	0.471420	0.000000	0.000000	0.000000	1.000000	1.000000

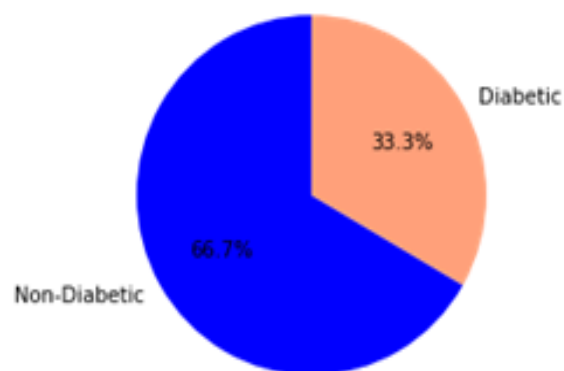
Upon evaluation of these metrics we could not find any variable of no use except patientID. We will be doing more evaluations in the following sections to get more understanding of the data.

### **3.5 Data profiling**

In this section we will look at the data distribution of each variable and observe the central tendency of data in the variables, any potential outliers and presence of any suspicious data.

#### **3.5.1 Distribution of outcome variable**

This indicates there are more non-diabetic records when compared with diabetic records in the data set. Our goal is to identify true-positive (sensitivity) cases for high risk diabetes in pregnant women and avoid false negative cases as much as we can. Because, false positive cases (false detection of diabetes -false alarm) are less costly than the false negative (not detecting diabetes even if it is present). We should be using the higher stake of non-diabetic records from our dataset to achieve a lower number of false negative cases without compromising on the overall model's accuracy on the unseen dataset.



### **3.5.2 Distribution of independent variables**

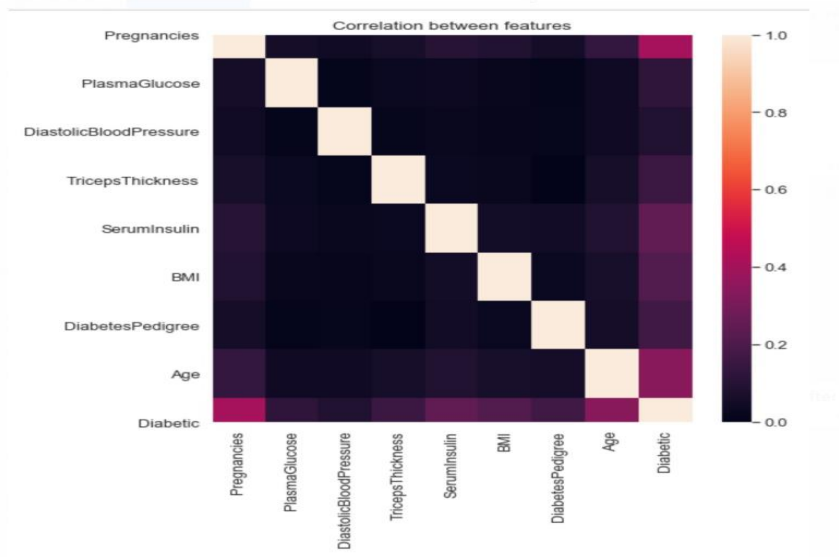
- By looking into the distribution of values for pregnancies, we can see that they are not normally distributed and have right-skewness. We can apply log transformation to a normal mode (bell-curve). However, their variable values will be normalized to make it symmetric.
- The distribution of plasma glucose and blood sugar variables are almost symmetric with minimal skewness and we will be applying the normalization function to have a symmetric distribution.
- The distribution of insulin, BMI, Diabetes pedigree & Age variables are also having right-skewness with outliers detected. The symmetric distribution of data with no effect of outliers on the model implementation can be applied by using normalized function.

### **3.6 Data Cleansing / Data filtering**

In this section, we have checked for the presence of any null values, applied the correlation function, and drew the pair plots to identify and eliminate if there is any dependency between them.

- Using “is null” function we can find out if there are any missing or NaN values in the data set and specifically checking for null values to replace them or eliminate them from the data set if any. In our data set, we do not have any missing or null values and hence proceeding ahead without replacing any of the values provided from Kaggle data set.
- The correlation function helps us to determine the correlation coefficient between two variables and +1 indicates strong linear relationship and -1 indicates a weak linear relationship. The heatmap generated for the input parameters with strong correlation greater than 0.5 should be considered as inter-parameter dependency and should be avoided

for model implementation. We do not see any correlation between our variables greater than 0.5 and hence all the variables are considered for our model implementation.



- Using the pair plots, we can identify the outliers in the dataset with the help of Pregnancy and Age values. For example, a person with age of 21 cannot have pregnancies greater than 4 and age 22 cannot have pregnancies greater than 5 and so on (we presumed that the minimum age for pregnancy was 18). We have eliminated these outliers with reference to age and pregnancies (total number of eliminated records with outliers: 2127).
- Since many parametric statistical models assume that the effects of each variable on the target are linear and operate at their best, we have normalized these variables. The normalization also eliminates the effect of outliers especially in SerumInsulin, Diabetes pedigree & Age variables.

### **3.7 Data Partitioning**

- In our supervised classification model, our goal is to find a model that accurately assigns data to separate predefined classes. To test the effectiveness of a model, an independent test data and its accuracy of classification helps us to get the quality of the classifier. Too simple parameters lead to under-fitting and too complex parameters at the same time lead to overfitting. Hence, we need to have a tradeoff between bias and variance and test different assumptions we made. To address the overfitting problem, we divide our data and develop our models using only one of the partitioning i.e. training data and we try to see how it performs on the other partitioned data set i.e. validation data.
- Our data set after data cleansing, got partitioned into two sets 1) Training data set, used iteratively for optimizing the parameters of the chosen classifier and 2) Validation data set, used to validate the generalization performance of the final classifier.
- The training data set consists of 80% of the entire data and remaining 20 % got partitioned into validation data.

### **3.8 Modeling**

In this section, we deal with selecting different modeling techniques, building the models, assessing the models and evaluating them. These steps will help us to develop different modeling algorithms and finally evaluate them to fulfill our business use case.

Our business use case is a classification problem of supervised machine learning i.e. the classification operation is based on the relationship between a known class assignment and characteristics of the entity to be classified. There are two general kinds of supervised classification

problems i) **binary classification** and ii) multiple classification. Our target variable i.e. diabetic or Non-diabetic is a binary classifier.

Modeling techniques: We prepare our data to choose different modeling algorithms and modeling architecture i.e. we might need to perform standardization or normalization or feature ranking of the independent variables before building the models.

Model building: Depending upon the modeling algorithm selected above , various types of models can be built based on the parameters selected .

Assessing the model: Using confusion matrix, lift charts, ROC curves and normal probability charts etc. are different tools available to assess a given model and we will analyze these models developed to tell how good the model is.

Evaluation of models: Evaluation of different models that are generated will help us to include a list of modeling goals for the future and modeling approaches to accomplish them. By evaluation and identification of the best model, a business use case model can be implemented accordingly.

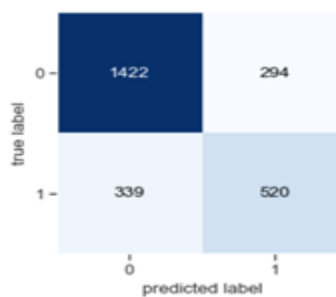
### **3.8.1 Baseline model**

As we know, the baseline model will help us to spot check the algorithms to compare the results. This baseline model is a meaningful reference point to which we can compare. A baseline model can be a simple prediction / classifying model and we can select the class that has most observations and use that class as the result of all predictions.

However, we took **Naïve Bayes classifier** as our baseline model as it follows the pattern of human thinking more closely than the classical statistical analysis.

This Naïve Bayes classifier works on the principle that the presence of one feature does not affect the other feature and they are independent of each other. We chose **Gaussian Naïve Bayes** as the predictors are continuous variables (after normalization).

From Naïve Bayes model, we got the baseline accuracy of 75.42 % and sensitivity of 0.605. And we would like to explore the other models in the following section to see if the accuracy and sensitivity values are better than the baseline model.



0	1422	294
1	339	520
	0	1
	predicted label	

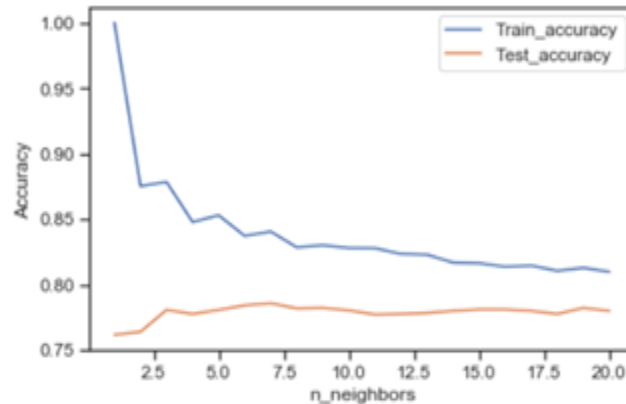
$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FN} + \text{FP}) = 1942 / 2575 = 0.7542$$

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) = 520 / (520 + 339) = 0.6053$$

### **3.8.2 k-Nearest Neighbors (k- NN)**

The next simplest algorithm that we used for classification of our target variable is k-Nearest Neighbors approach (k-NN). This algorithm can be used for classification or prediction using the method that relies on finding “similar” records in the training data. These neighbors are then used to derive a classification for the new record by voting. However, this algorithm does not make assumptions about the form of the relationship between class membership and predictors.

We chose the k with best classification performance i.e. by classifying the records in the validation data for computing accuracy for various choices of k. Typically, the values of k fall in the range of 1 to 20 and better to choose an odd number to avoid ties. For our data set k-NN model is at the best accuracy of 79% with k value of 7 as shown in the graph below.



### **3.8.3 Logistic Regression**

Logistic regression is used more widely in classification rather than numerical prediction and therefore we implemented this model for our classification dataset. This algorithm is used to model the nonlinear relationship between outcome and combined effects of the independent predictor variables. This algorithm applies a function called logit to predict the outcome and can be mapped back to probability of the outcome.

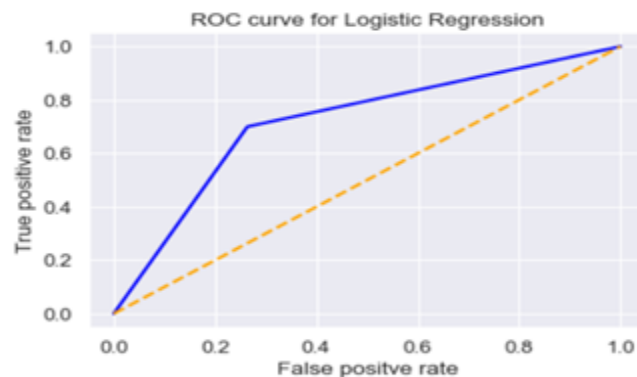
The logit function fits many binary classification problems and expresses much of the non-linear effects of the predictors (from correlation coefficients represented in the heatmap suggest there are no linear relationships between the predictors and the outcome).

The ROC curve plotted between TPR (True positive rate = sensitivity) and FPR (False positive rate = 1-specificity) gives us the tradeoff between sensitivity and specificity. The AUC (area under



the curve) value of the ROC curve will help us to determine the model's capability to achieve better sensitivity and specificity values. However, the ideal coordinates for ROC space is (1,1) i.e. 100 % sensitivity and 100 % specificity and AUC value of 1.

The below ROC curve is plotted with FPR on the x-axis and TPR on the y-axis. And, we can conclude from the graph that the model is better if the ROC curve is closer to ideal co-ordinate.



The AUC value obtained for the ROC curve is 0.7, which indicates that the logit model needs to be fine-tuned to get better sensitivity (True positives) and accuracy values (~74%) of the validation data. However, we will proceed with other models to see if we get better accuracy and sensitivity values.

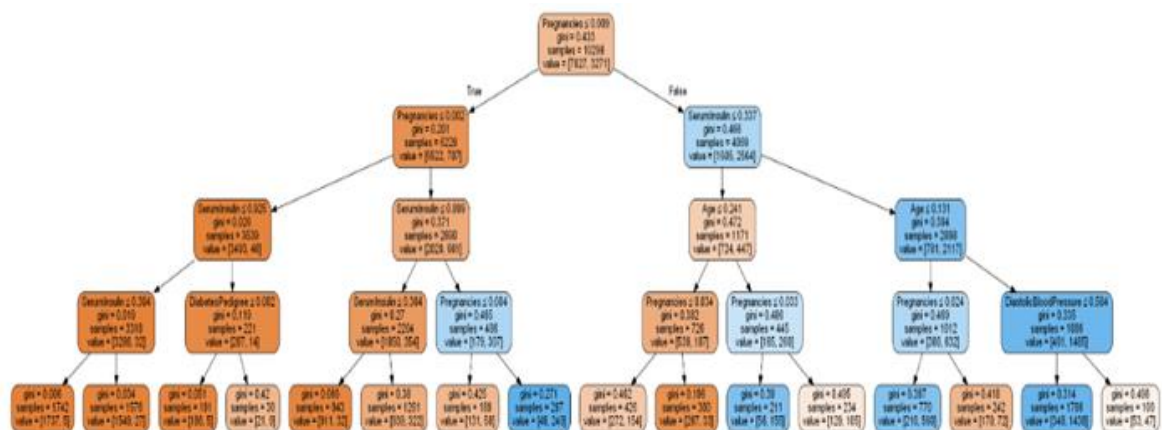
### **3.8.4 Decision Trees**

The decision tree / classification and regression tree algorithm are structured as a sequence of simple questions and the answers to these questions determine what next question (split) if any. The result is a network of questions that forms a tree like structure. Once a best split is found the algorithm repeats the search process for each node below, until either further splitting is stopped by a criterion or splitting is impossible.

The decision trees are grown larger than they need to be and then pruned back to find the best tree. Decision tree determines the best tree by using the validation data. One of the major advantages of this decision trees is its simple structure and can be used for variable selection.

Initial settings in this algorithm are common to most classification procedures and the first setting is the prior probability of target variables. Another important setting is to select the measure of impurity (Gini Index) to use in evaluating split points. The Gini Index is based on the relative frequency of sub ranges in the predictor variables.

The decision tree displayed below has a maximum depth of 4 and the corresponding Gini Index values displayed. The accuracy value for the decision tree is approximately 85.75%, which is better than the baseline model's accuracy of 75.42%.



### 3.8.5 Random Forest

The random forest algorithm is an advanced form of decision trees which consists of groups of trees. This algorithm trains several trees on slightly different subsets of data, in which a case is

added to each subset containing random selections from the range of each variable. Each decision tree in this group of trees votes for the classification of each input case.

One of the major advantages of random forest algorithm is its high accuracy among algorithms for classification and it can handle very large sets of data with hundreds / thousands of variables. It also provides an estimate of variables importance.

true label	0	1	
	1576	140	
1	110	749	
predicted label			
		0	1

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FN} + \text{FP}) = 2325 / 2575 = 0.9029$$

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) = 749 / (749 + 110) = 0.8719$$

As shown above in the confusion matrix, the values measured with reference to accuracy and sensitivity can be calculated as 0.9029 and 0.8719, respectively. However, we would like to implement another model with neural networks to see if the sensitivity and accuracy values can be improved.

### **3.8.6 Neural Networks**

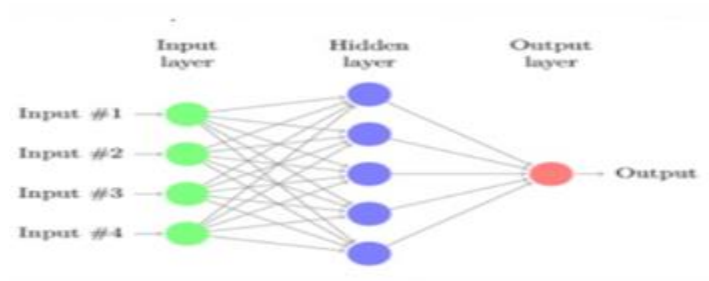
Neural networks used for computation were based on the early understandings of the structure and function of the human brain. The neuron cells receive electrical impulses from neighboring cells and accumulate them until a threshold value is exceeded. Then they fire and impulse to an adjacent

cell . Artificial neural network (ANN) tries to imitate the human nervous system by connecting a series of neurons together to form a large network that processes the information.

This architecture forms a network in which each input variable (called an input node) is connected to one or more output nodes. When the input nodes with summation aggregation function and a logistic activation function are directly connected to an output node, the mathematical processing is analogous to a logistic regression with binary output. The logistic function fits many binary classification problems and can express much of the non-linear effects of the predictors.

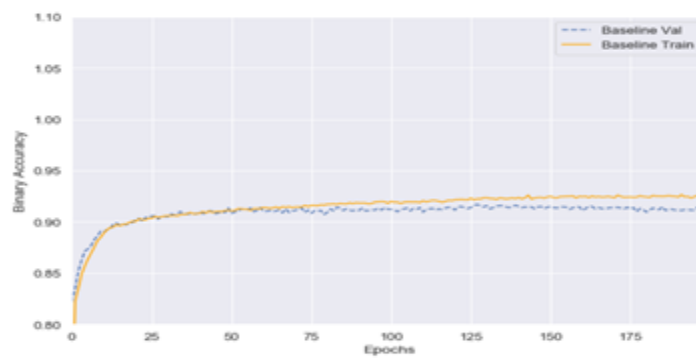
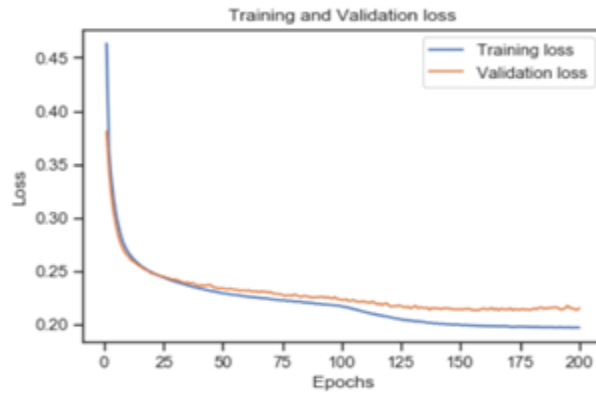
The logistic sigmoid function is a smooth version of the threshold function (sum of dot products). If the sum is greater than 0 the sigmoid function returns a value closer to 1 and if the sum is less than 0 a value is closer to 0, respectively. If the sum is 0 then the sigmoid function returns a value of 0.5.

We implemented a multilayer perceptron i.e. a densely connected neural network which consists of an input layer, at least one or more hidden layers and an output layer.

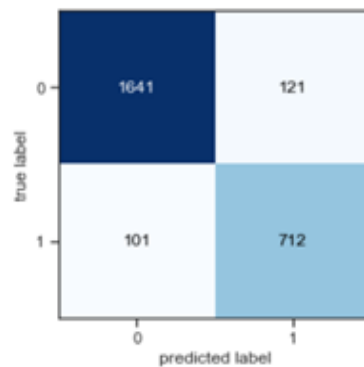


The activation function used in the hidden layers is a rectifier function (ReLU) while the activation function used at the output layer is a sigmoid function.

Now, we generate the loss and accuracy plots over the training and validation data. At the end of each epoch, the model computes its loss and accuracy on the partitioned datasets.



As we can see, the training loss decreases with every epoch and training accuracy with every epoch. However, the validation loss is not really decreasing after 150 epochs and validation accuracy gets improved at every epoch even after 150 epochs. To avoid any overfitting with the training data, we can potentially stop the training after 150 epochs.



$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FN} + \text{FP}) = 2325 / 2575 = 0.91$$

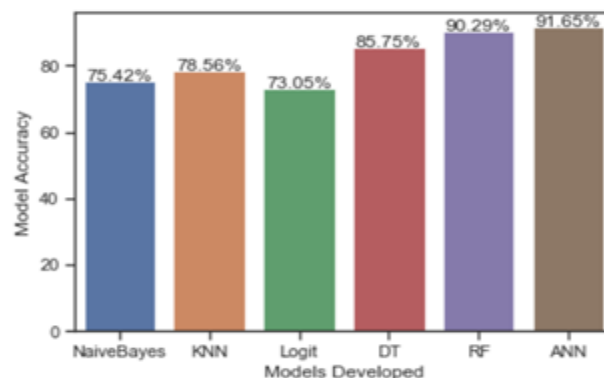
$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) = 712 / (712 + 101) = 0.8757$$

The above plot displays the confusion matrix with an artificial neural network and its accuracy and sensitivity values. These measures are at the highest values that we achieved so far.

### **3.9 Evaluation of the models implemented**

We captured the accuracy measures based on the validation data for each model implemented and compared them to find the best accurate model. Because in our classification problem, it is important to know how well the model performed on the unseen data.

Bar graph



From the above bar graph, we can see various accuracy values for the respective models. The accuracy for all the models is higher when compared with baseline model accuracy and the highest accuracy is obtained for artificial neural network model. The ANN model's sensitivity value (0.875) is also better amongst all the models we developed so far (second best value is 0.871 for random forest model). Hence, we can use this best accurate model to classify the pregnant women's risk of diabetes. In the following section, we are going to use this best accurate model to

implement a Business model for pregnant women / healthcare professionals to predict the risk of diabetes.

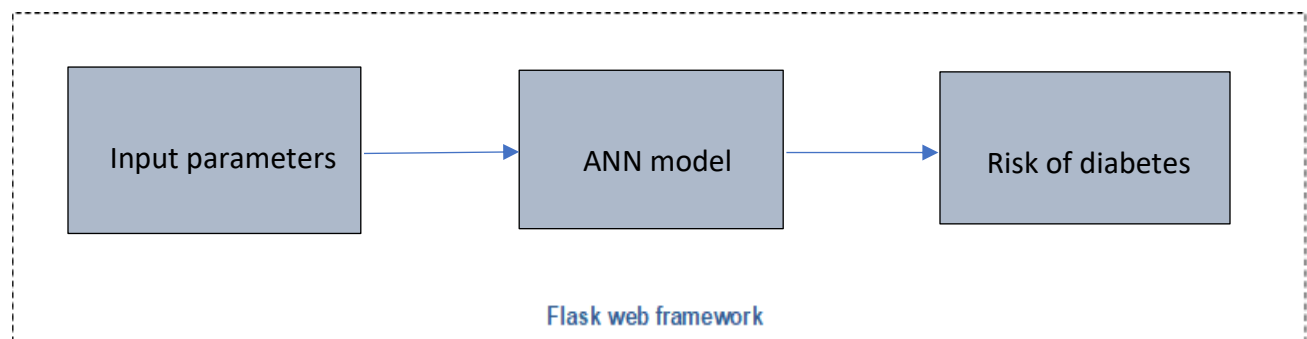
## **CHAPTER 4**

### **Business model for predicting the risk of diabetes in real world**

In this section, we are going to propose a business model which can be used by end users (pregnant women), hospitals or healthcare professionals.

A web application or a mobile application or iWatch application can be developed using a web framework designed to predict the risk. However, the scope of this project is to develop a prototype as a business model using a flask web framework and demonstrate how the prediction using artificial neural network can be used to predict risk of diabetes based on the input parameters provided by the user.

The below mentioned web framework has input parameters that can be passed on to the model we developed in the background and respond back with the outcome. This is achieved using html, models developed, flask web framework interfaces.



The below screenshot is an input screen to allow the user to provide input parameters. In our business use case/ data set we have 8 different parameters to enter.

Diabetes risk prediction in pregnant women

Number of pregnancies?

Glucose level ?

Blood Pressure?

Skin Thickness?

Insulin?

Body Mass Index?

Diabetes Pedigree Function?

Age?

CAL STATE EAST BAY

Sample input parameters provided by an end user to check for her risk to diabetes:

Diabetes risk prediction in pregnant women

Number of pregnancies?

Glucose level ?

Blood Pressure?

Skin Thickness?

Insulin?

Body Mass Index?

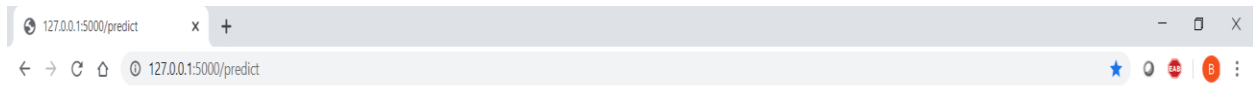
Diabetes Pedigree Function?

Age?

CAL STATE EAST BAY

The output screen displayed once user clicks on predict icon. In this case, the patient is diabetic, and we advised the user to contact the primary care physician to discuss further.





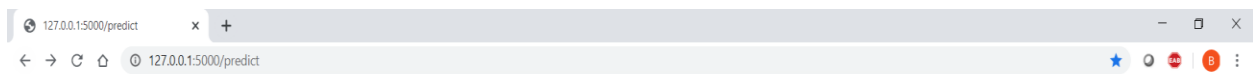
Diabetes risk prediction in pregnant women



## The risk prediction is Diabetic

**\*\*Note:** The outcome is based on hypothetical model. Please follow the instructions from your primary care physician to discuss further to stay fit and healthy!

Another business use case, with different user parameters.

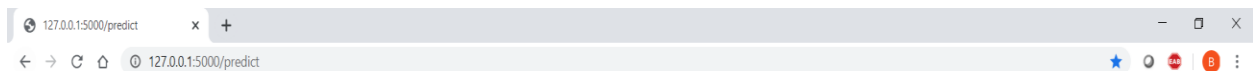


Diabetes risk prediction in pregnant women



Number of pregnancies?	<input type="text" value="0"/>
Glucose level ?	<input type="text" value="170"/>
Blood Pressure?	<input type="text" value="80"/>
Skin Thickness?	<input type="text" value="23"/>
Insulin?	<input type="text" value="56"/>
Body Mass Index?	<input type="text" value="23"/>
Diabetes Pedigree Function?	<input type="text" value="0.627"/>
Age?	<input type="text" value="28"/>

In this case, the patient is non-diabetic, and we advised the user to contact the primary care physician to discuss further.



Diabetes risk prediction in pregnant women



## The risk prediction is Non-Diabetic

**\*\*Note:** The outcome is based on hypothetical model. Please follow the instructions from your primary care physician to discuss further to stay fit and healthy!



## **CHAPTER 5**

### **Study of heart disease classification**

#### **5.1 Data Acquisition and Data extraction**

Coronary Heart Disease (CHD) is the most common type of heart disease killing over 370,000 people annually. The data is taken from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The appropriate interpretation of this kind of data related to heart disease can be very helpful in taking precautions as required. Machine learning helps us in this process making decisions and predictions from the large quantity of data produced by the healthcare industry.

#### **5.2 Tools used to analyze the dataset**

We used the open source programming tool Python 3. We also made use of the inbuilt libraries present in python to do all the required analysis. Jupyter notebook is used to do all the required coding.

#### **5.3 Data description**

In this section we try to analyze the data set and the goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD) or not. The [link](#) for the dataset can be found in Kaggle.

The dataset provides the information related to patients and it includes over 4238 records and 16 different attributes. It is important to first understand these attributes for better understanding.

## **5.4 Attributes**

Each attribute is a potential risk factor in this study: demographic, behavioral and medical risk factors are all included.

### **Demographic:**

sex: male or female

age: age of the patient

### **Behavioral**

currentSmoker: The patient is a current smoker

cigsPerDay: The number of cigarettes that the person smoked on average in one day

### **Medical( history):**

BPMeds: whether the patient was on blood pressure medication

prevalentStroke: whether patient had previously had a stroke

prevalentHyp: whether the patient was hypertensive

diabetes: whether the patient had diabetes

### **Medical(current):**

totChol: total cholesterol level

sysBP: systolic blood pressure

diaBP: diastolic blood pressure

BMI: Body Mass Index

heartRate: heart rate

glucose: glucose level

Predict variable: 10-year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

**Below screenshot represents few records from the Heart Diseases dataset:**

```
HeartDatasets.head(5)
```

Sex	Age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCI
1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	
0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	
1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	
0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	
0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	

## **5.5 Statistical metrics**

Analysis and evaluation of these metrics will help us to prepare the data for modeling.

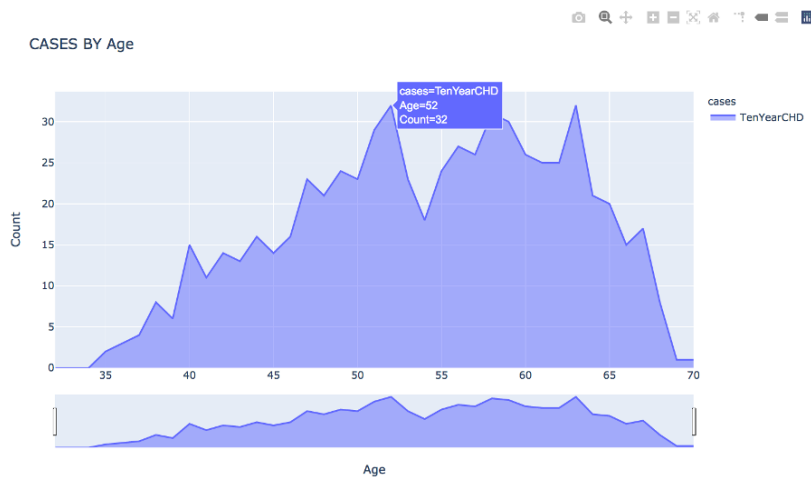
```
HeartDatasets.describe().T
```

	count	mean	std	min	25%	50%	75%	max
male	4238.0	0.429212	0.495022	0.00	0.00	0.0	1.000	1.0
age	4238.0	49.584946	8.572160	32.00	42.00	49.0	56.000	70.0
education	4133.0	1.978950	1.019791	1.00	1.00	2.0	3.000	4.0
currentSmoker	4238.0	0.494101	0.500024	0.00	0.00	0.0	1.000	1.0
cigsPerDay	4209.0	9.003089	11.920094	0.00	0.00	0.0	20.000	70.0
BPMeds	4185.0	0.029630	0.169584	0.00	0.00	0.0	0.000	1.0
prevalentStroke	4238.0	0.005899	0.076587	0.00	0.00	0.0	0.000	1.0
prevalentHyp	4238.0	0.310524	0.462763	0.00	0.00	0.0	1.000	1.0
diabetes	4238.0	0.025720	0.158316	0.00	0.00	0.0	0.000	1.0
totChol	4188.0	236.721585	44.590334	107.00	206.00	234.0	263.000	696.0
sysBP	4238.0	132.352407	22.038097	83.50	117.00	128.0	144.000	295.0
diaBP	4238.0	82.893464	11.910850	48.00	75.00	82.0	89.875	142.5
BMI	4219.0	25.802008	4.080111	15.54	23.07	25.4	28.040	56.8
heartRate	4237.0	75.878924	12.026596	44.00	68.00	75.0	83.000	143.0
glucose	3850.0	81.966753	23.959998	40.00	71.00	78.0	87.000	394.0
TenYearCHD	4238.0	0.151958	0.359023	0.00	0.00	0.0	0.000	1.0

## **5.6 Data profiling**

### **5.6.1 Distribution of Outcome Variable**

Let us look at the people's age who are suffering from the disease or not. Here TenYearCHD=1 implies that the Patient is suffering from heart disease and TenYearCHD = 0 implies the Patient is not suffering.



We see that most of the patients who are suffering are of the age of 52, followed by 63. Majorly, people belonging to the age group 50+ are suffering from the disease.

### **5.6.2 Distribution of independent Variables**

There are 3579 patients with no heart disease and 562 patients with risk of heart disease. This indicates there are more non heart disease records when compared with diabetic records in the data set. The people with the highest risk of developing CHD are between the ages of 51 and 63 and the number of sick people generally increases with age. We can also observe that there are slightly more males suffering from CHD than females. The percentage of people who have CHD is almost equal between smokers and non-smokers.

CHD is higher among the diabetic, and those with prevalent hypertension as compared to those who do not have similar issues. A larger percentage of the people who have CHD are on blood pressure medication and the Diabetes Blood Pressure is aged between 35 and 65.

## **5.7 Data Cleansing / Data filtering**

Data cleaning is an important step in any of the data related problems. As all the machine learning models learn from data it is important that the data, we feed them is specifically preprocessed and refined as per the problem in hand. The steps include data cleaning, preprocessing, feature engineering etc.

We installed all the required libraries that are useful for our study.

```
import os
import pandas as pd
import numpy as np
import datetime
import matplotlib.pyplot as plt
import seaborn as sns
import calmap
from datetime import date, timedelta, datetime

from plotly.offline import plot, iplot, init_notebook_mode
init_notebook_mode(connected=True)

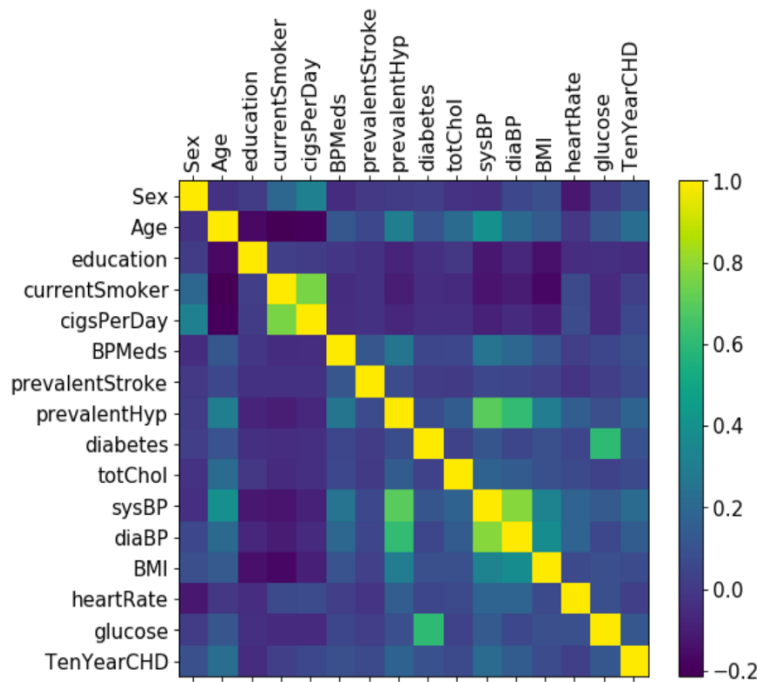
import plotly.express as px
import plotly.graph_objs as go
import plotly.figure_factory as ff
from plotly.subplots import make_subplots
from biokit.viz import corrplot

from pywaffle import Waffle
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, VotingClassifier
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import f1_score, accuracy_score, recall_score, precision_score
from sklearn.metrics import classification_report, confusion_matrix, roc_auc_score, roc_curve
from xgboost import XGBClassifier
```

- The heart disease data set contains null values.
- Total number of rows with missing values is 645. Since, it is only 15 percent of the entire dataset we will fill the missing values based on Median and Mean values, respectively.
- Correlation coefficients: The corrplot package is a graphical display of a correlation matrix, confidence interval.

currentSmoker	1	.76	.20	.02	-.10	-.13	-.11	-.21	-.17	-.05	-.05	.02	-.04	-.05	-.03	.06
cigsPerDay	.76	1	.32	.01	-.07	-.09	-.06	-.19	-.09	-.04	-.04	.06	-.04	-.06	-.03	.07
Sex	.20	.32	1	.01	.01	-.04	.06	-.03	.08	-.03	-.05	.09	.02	.01	-	-.12
education	.02	.01	.01	1	-.07	-.12	-.06	-.16	-.14	-.01	-.01	-.05	-.04	-.03	-.03	-.04
prevalentHyp	-.10	-.07	.01	-.07	1	.70	.62	.31	.30	.14	.26	.18	.08	.08	.07	.15
sysBP	-.13	-.09	-.04	-.12	.70	1	.78	.39	.33	.17	.25	.22	.11	.13	.06	.18
diaBP	-.11	-.06	.06	-.06	.62	.78	1	.21	.38	.14	.19	.15	.05	.06	.05	.18
Age	-.21	-.19	-.03	-.16	.31	.39	.21	1	.14	.21	.12	.23	.10	.12	.06	-.01
BMI	-.17	-.09	.08	-.14	.30	.33	.38	.14	1	.09	.10	.07	.09	.08	.02	.07
totChol	-.05	-.04	-.03	-.01	.14	.17	.14	.21	.09	1	.06	.07	.04	.03		.07
BPMeds	-.05	-.04	-.05	-.01	.26	.25	.19	.12	.10	.06	1	.09	.05	.05	.12	.02
TenYearCHD	.02	.06	.09	-.05	.18	.22	.15	.23	.07	.07	.09	1	.10	.12	.06	.02
diabetes	-.04	-.04	.02	-.04	.08	.11	.05	.10	.09	.04	.05	.10	1	.61	.01	.05
glucose	-.05	-.06	.01	-.03	.08	.13	.06	.12	.08	.03	.05	.12	.61	1	.02	.09
prevalentStroke	-.03	-.03	-	-.03	.07	.06	.05	.06	.02		.12	.06	.01	.02	1	-.02
heartRate	.06	.07	-.12	-.04	.15	.18	.18	-.01	.07	.07	.02	.02	.05	.09	-.02	1

- Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients.
- There are no features with more than 0.5 correlation with the Ten-year risk of developing CHD. However, the features with the highest correlations are age, prevalent hypertension and systolic blood pressure.
- Also, there are a couple of features that are highly correlated with one another and we can use either of one in building a machine learning model.
- These include: Blood glucose and diabetes, systolic and diastolic blood pressures, cigarette smoking and the number of cigarettes smoked per day. Here, we carried out feature selection to pick the best features.
- The correlation matrix can be reordered according to the correlation coefficient. This is important to identify the hidden structure and pattern in the matrix.



## **5.8 Data Partitioning**

The dataset has been divided into a training set and a test set and individual classifiers are trained using the training dataset. The efficiency of the classifiers is tested with the test data. For our study we divided the data in the ratio of 80: 20, respectively. The training data size is 80% and testing data size is 20% of the whole data. Evaluation metrics like confusion matrix, ROC curve, Lift charts are used.

## **5.9 Modeling**

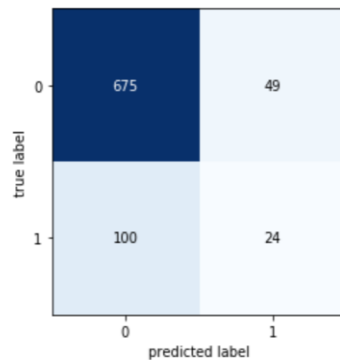
Classification is a supervised learning procedure that is used for predicting the outcome from existing data. Hence, we use different classification algorithms for our study and finally evaluate the model with any of the available evaluation metrics.



### **5.9.1 Baseline model**

The Bayesian network is a graphical prediction model based on probability theory. Bayesian networks are built from probabilistic distributions, and they utilize the laws of probability for prediction and diagnosis. The Naïve Bayes classifier or simply, the Bayesian classifier, is based on the Bayes theorem.

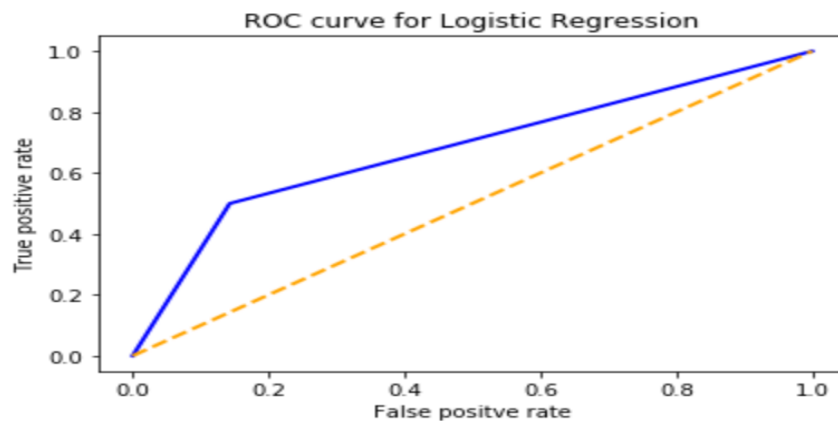
Accuracy of our naive bayes model on test data is: 0.8243  
Confusion Matrix



### **5.9.2 Logistic regression**

Logistic regression is a type of regression analysis in statistics used for prediction of outcome of a categorical dependent variable from a set of predictor or independent variables. In logistic regression the dependent variable is always binary. Logistic regression is mainly used for prediction and calculating the probability of success.

Accuracy of our LOGISTIC REGRESSION model: 0.8538



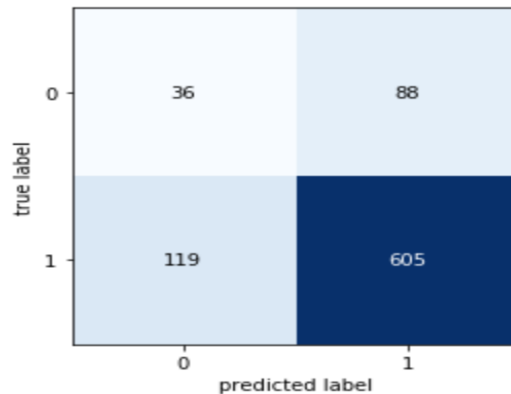
ROC curve for Logistic Regression is 0.6785714285714286

### 5.9.3 Decision Tree

A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question, edges represent the answers to the question and the leaves represent the actual output or class label. They are used in non-linear decision making with simple linear decision surfaces.

Decision trees classify the examples by sorting them down the tree from the root to some leaf node, with the leaf node providing the classification to the example. Each node in the tree acts as a test case for some attribute, and each edge descending from that node corresponds to one of the possible answers to the test case. This process is recursive in nature and is repeated for every subtree rooted at the new nodes.

Accuracy of our Decision tree model: 84.1981  
Confusion Matrix for Decision Tree Model



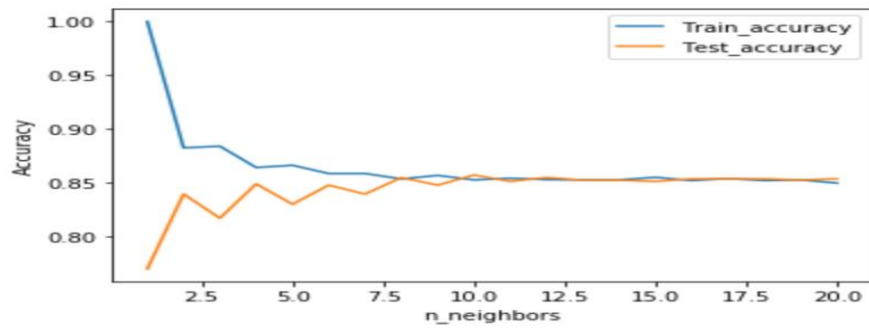
#### **5.9.4 k-Nearest Neighbors (k- NN)**

The k-nearest-neighbors is a data classification algorithm that attempts to determine what group a data point is by looking at the data points around it. The k-nearest-neighbor is an example of a "lazy learner" algorithm because it does not generate a model of the data set beforehand. The only calculations it makes are when it is asked to poll the data point's neighbors. This makes k-NN very easy to implement for data mining.

This classifier looks for the classes of k nearest neighbors of a given data point and based on the majority class, it assigns a class to this data point. However, the number of neighbors can be varied.

We varied them from 1 to 20 neighbors in our study and calculated the test score in each case.

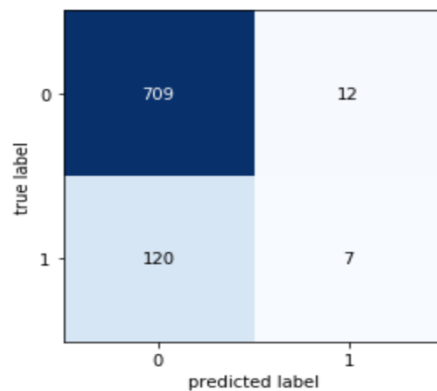
Accuracy of our K-NN classifie model: 0.8538



### 5.9.5 Neural Network

The proposed system implements the concept of multilayered neural networks are proved to be effective for practical applications. This system is processed in two phases: in the first phase 14 attributes are fed as input and then the network is trained with training data by back propagation learning algorithm.

Confusion Matrix

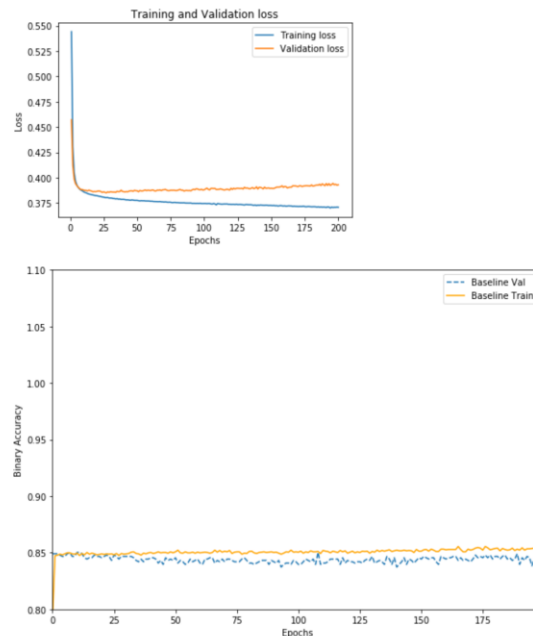


sensitivity of ANN= 0.05511811023622047  
Accuracy for Neural Network 85.02358198165894

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FN} + \text{FP}) = 721 / 848 = 0.85$$

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) = 709 / (709 + 120) = 0.8552$$

The above plot displays the confusion matrix with an artificial neural network its accuracy and sensitivity values. These measures are at the highest values that we achieved so far.

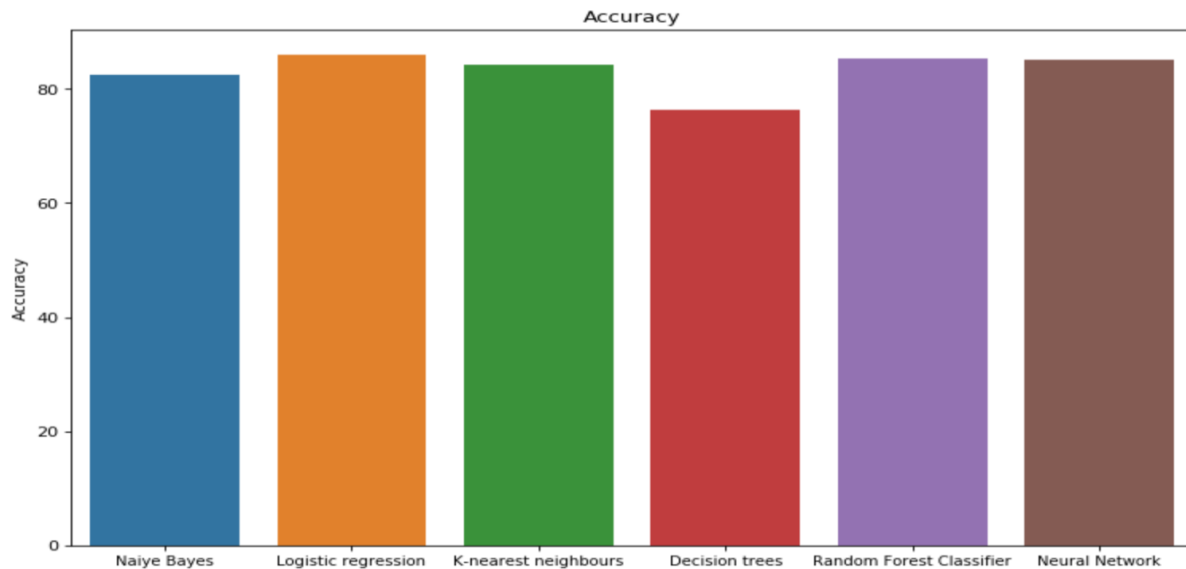


As we can see, the training loss decreases with every epoch and training accuracy with every epoch. However, the validation loss is not really decreasing after 120 epochs and validation accuracy gets improved at every epoch even after 120 epochs. To avoid any overfitting with the training data, we can potentially stop the training after 120 epochs.

### **5.10 Comparison of models**

In this dataset we captured the accuracy of models measured based on the training and testing data for each model implemented and we also compared them to find the best accurate model. Because in our classification problem, it is important to know how well the model performed on the unseen data.

	Accuracy
Naiye Bayes	82.429245
Logistic regression	85.966981
K-nearest neighbours	84.316038
Decision trees	76.297170
Random Forest Classifier	85.377358
Neural Network	85.023582



### **5.11 Best model for heart disease is LOGISTIC REGRESSION**

From the above bar graph, we can see various accuracy values for the respective models. The accuracy for all the models is higher when compared with baseline model accuracy and the highest accuracy is obtained for Logistic Regression model (85.9666). Logistic Regression, Random Forest and Neural Network model all three have accuracy 85% and Most of all the model accuracy is greater than 82%. Hence, we can use this best accurate model to classify the patient Heart Diseases. In the following section, we are going to use this best accurate model to implement a Business model for predicting the ten-year risk of developing Heart Diseases.

### **5.12 Cross validation score of the best model**

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model. It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split. The general procedure is as follows: Shuffle the dataset randomly, Split the dataset into k groups. For each unique group: Take the group as a hold out or test data set, Take the remaining groups as a training data set, fit a model on the training set and evaluate it on the test set. Retain the evaluation score and discard the model. Summarize the skill of the model using the sample of model evaluation scores.

Cross-validated scores [0.85141509 0.85023585 0.8490566 0.85123967 0.84651712]

The Cross-Validation accuracy is: 84.97%

Best model accuracy is: 85.96%



## **CHAPTER 6**

### **Applications and Findings**

The business model proposed to classify a diabetic disease can be used by end users or healthcare professionals. The input parameters might require preliminary tests to determine exact values and can be passed on to best accurate model to predict the outcome.

The model predicted with 0.85 accuracy. The model is more specific than sensitive. Accuracy for those who have 10-year risk of coronary heart disease with threshold 0.1 is 85.36% Now as we have good accuracy for predicting people with chances of heart disease in the next 10 years, we can treat them and guide them well in advance.

However, our proposal is not limited to classification of the outcome alone. It can also be used by healthcare management professionals to simulate the input parameters, educate the patients about the risk of any chronic conditions in the next stages of their life and health conditions (diabetes classification, 10-year risk of heart disease etc.) and advise them preventive measures.

In medical research models, we should be more concerned about sensitivity (True Positive) as a key parameter in addition to accuracy to find the best model. This parameter plays a major role in diagnosis applications because false detection of non-disease (False Negative) will worsen the patient's condition than the false detection (False Positive) of disease.

It is always better to execute the models designed based on the data available of a chronic disease to improve the models and their accuracy. Hence, we have implemented all the models for diabetes and heart disease data set to find the best accurate model, respectively. We cannot generalize the best accurate model from one data set to another data set with different input parameters although the outcome is to classify the presence or absence of disease.





## **CHAPTER 7**

### **Conclusion**

The diabetes and heart disease data sets have been studied individually for business and data understanding followed by data preprocessing and modeling and their evaluation for deployment.

These models have been validated against the unseen data i.e. test / validation data (mutually exclusive from training data). Once we have the best accurate model based on factors ( accuracy, sensitivity), a prototype has been built to classify / predict the risk of diabetes in pregnant women using artificial neural networks as a classification model in the background.

The prototype extracts the outcome from the ANN model and its hidden knowledge from the data set on which it is trained on. This prototype of diabetes risk prediction business model can be further enhanced to include symptoms, precautionary measures, overview of the disease etc. and also it can be rolled out as an end user or business application like mobile application or web application UI.

Men seem to be more susceptible to heart disease than women. Increase in age, number of cigarettes smoked per day and systolic Blood Pressure also show increasing odds of having heart disease. Total cholesterol shows no significant change in the odds of CHD. This could be due to the presence of 'good cholesterol (HDL) in the total cholesterol reading. Medical research studies prove that diabetes can damage the nervous system which can result in another chronic condition like heart disease. We would like to suggest a subsequent mandatory step to use these two different case studies implemented on diabetes and heart disease prediction models together. Because, the presence or absence of diabetes can help in predicting the risk of heart disease.

## **References**

- We would like to thank professor Dr. Jiming Wu for his valuable inputs and guidance throughout this study of the project.
- Link for diabetic dataset  
<https://www.kaggle.com/fmendes/diabetes-from-dat263x-lab01>
- Link for heart disease dataset  
<https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset>
- Data-Mining-for-Bioinformatics-Applications by By He Zengyou
- Handbook-of-Statistical-Analysis-and-Data-Mining-Applications,by RobertNisbet, John Elder and Gary Miner
- Phases of the CRISP- M process ( Chapman et al., 2000 )
- Data mining for business analytics - Galit Shmueli
- Introduction to Machine Learning with Python: A Guide for Data Scientists by Andreas C. MullerSarah Guido September 26, 2016, “O’Reilly Media, Inc.”
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC193602/>
- <https://www.medicalnewstoday.com/articles/150999>
- <https://www.ncoa.org/blog/chronic-versus-acute-disease/>
- <https://www.cdc.gov/diabetes/basics/diabetes.html>
- <https://www.cdc.gov/pregnancy/diabetes.html>
- <https://www.cdc.gov/pregnancy/diabetes-gestational.html>
- <https://www.toppr.com/guides/biology/why-do-we-fall-ill/health-and-its-significance/>

- <https://www.statisticshowto.com/probability-and-statistics/skewed-distribution/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5001642/>
- <https://machinelearningmastery.com/how-to-get-baseline-results-and-why-they-matter/>
- <https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>
- <https://www.datacamp.com/community/tutorials/decision-tree-classification-python>
- <https://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/understanding-medical-tests-sensitivity-specificity-and-positive-predictive-value/>
- <https://www.lexjansen.com/nesug/nesug10/hl/hl07.pdf>
- [https://www.who.int/healthinfo/global\\_burden\\_disease/GlobalHealthRisks\\_report\\_full.pdf](https://www.who.int/healthinfo/global_burden_disease/GlobalHealthRisks_report_full.pdf)
- <https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/heart-disease-stroke>
- <https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machine-learning-library/>
- <https://towardsdatascience.com/end-to-end-python-framework-for-predictive-modeling-b8052bb96a78>