GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

HUMAN DISEASE PREDICTION

A Report for the Evaluation 3 of Project 2

Submitted by

SHOBHIT RANA

(1613101700)

in partial fulfilment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING WITH SPECIALIZATION OF CLOUD COMPUTING AND VIRTUALIZATION

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of

Ms. NILANJANA PRADHAN,

Assistant Professor

APRIL / MAY- 2020

GALGOTIAS
UNIVERSITY

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report **“ HUMAN DISEASE PREDICTION ”** is the bonafide work of “**SHOBHIT RANA (1613101700)”** who carried out the project work under my supervision.

SIGNATURE OF HEAD
Dr. MUNISH SHABARWAL,
PhD (Management), PhD (CS)
Professor & Dean,
School of Computing Science & Engineering

SIGNATURE OF SUPERVISOR
Ms. NILANJANA PRADHAN,
Assistant Professor
School of Computing Science & Engineering

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	**iii**
	LIST OF FIGURES	**iv**
	LIST OF TABLES	**iv**
1.	**INTRODUCTION**	**1**
	1.1 OVERALL DESCRIPTION	**1**
	1.2 PURPOSE	**3**
	1.3 MOTIVATION AND SCOPE	**3**
2.	**EXISTING SYSTEM**	**4**
3.	**PROPOSED SYSTEM**	**4**
4.	**IMPLEMENTATION**	**4**
5.	**OUTPUT**	**11**
6.	**CONCLUSION**	**12**
7.	**REFERENCES**	**12**

ABSTRACT

Many patients go untreated or are not treated accurately by the doctors, proper treatment is necessary for the well-being of the people. Thus, predicting a disease using the patient's symptoms has become an important task these days. To solve this acute shortage of doctors there must be a predicting system for predicting the general diseases which would help in proper utilization of the resources. Data analysis and Machine learning can help in deciding the line of treatment to be followed by extracting knowledge from suitable databases. Healthcare facilities need to be advanced so that better decisions for patient diagnosis and treatment options can be made. In this paper, a model is proposed for predicting the disease suffered by a person by knowing the symptoms. The model uses the Logistic Regression algorithm, which assigns observations to a discrete set of classes and provides a good level of accuracy. It collects the data of a person's symptoms and suggests a suitable disease accordingly. It will help in assisting healthcare practitioners by reducing the pressure on overcrowded clinics. To showcase the accuracy of the proposed model, it has been implemented on a heart disease dataset to predict the occurrence of heart disease in the next 10 years. The implementation will illustrate the effectiveness of the proposed model which can help in the development of an intelligent healthcare system.

Keywords: disease prediction, logistic regression, healthcare

LIST OF FIGURES

- Disease predicting model framework for predicting heart disease
- Number of people with history of heart disease
- independent and dependent variables

LIST OF TABLES

- imported dataset from Kaggle
- Logistic regression results using backward elimination (P value approach)
- Effect in odds of heart disease

1. INTRODUCTION

1.1 OVERALL DESCRIPTION

Many patients go untreated or are not treated accurately by the doctors, proper treatment is necessary for the well-being of the people. Thus, predicting a disease using the patient's symptoms has become an important task these days. There is a lack of doctors in India, there is 1 doctor for every 10,198 doctors in India (WHO recommends the ratio of 1:100). To solve this acute shortage of doctors there must be a predicting system for predicting the general diseases which would help in proper utilization of the resources.

The first step in treating a patient is the correct detection of the wellness of an individual by using the given symptoms. The prediction of the disease has become a vital task lately however the correct prediction of diseases has become too tough for a doctor. The system proposed in this paper is meant to develop a disease prediction system by exploiting machine learning. The classification within the prediction system is done with the help of the logistical regression algorithm. This may facilitate correct prediction of wellness and also facilitate in the correct treatment of disease.

The main focus is on to use machine learning in healthcare to supplement patient care for better results. Machine learning has made easier to identify different diseases and diagnosis correctly. Predictive analysis with the help of efficient multiple machine learning algorithms helps to predict the disease more correctly and help treat patients. The healthcare industry produces large amounts of healthcare data daily that can be used to extract information for predicting disease that can happen to a patient in future while using the treatment history and health data. This hidden information in the healthcare data will be later used for affective decision making for patient's health. Also, this area need improvement by using the informative data in healthcare.

Data volume is an enormous challenge in any industry but particularly in healthcare where data tends to sit idle in databases managed by dated EHR(Electronic Health Record) systems. Many companies build their business on getting large volumes of data from these systems to make them available and actionable as they power predictive analytics, decision support, imaging, operation optimization, and other applications. Other organizations make extensive use of insurance claims data that have recently become available through state governments. However, guidelines for gaining access for commercial purposes are nascent utilizing complex processes, so successful applications have been few and far between.

Machine learning in healthcare has recently made headlines. Google has developed a machine learning algorithm to help identify cancerous tumors on mammograms. Stanford is using a deep learning algorithm to identify skin cancer. A recent JAMA article reported the results of a deep machine-learning algorithm that was able to diagnose diabetic retinopathy in retinal images. It's clear that machine learning puts another arrow in the quiver of clinical decision making.

Still, machine learning lends itself to some processes better than others. Algorithms can provide immediate benefit to disciplines with processes that are reproducible or standardized. Also, those with large image datasets, such as radiology, cardiology, and pathology, are strong candidates. Machine learning can be trained to look at images, identify abnormalities, and point to areas that need attention, thus improving the accuracy of all these processes. Long term, machine learning will benefit the family practitioner or internist at the bedside. Machine learning can offer an objective opinion to improve efficiency, reliability, and accuracy.

Human intelligence can hardly be compared to any other phenomenon. Machine learning intelligence in healthcare has a lot of possibilities to improve the smart decisions made by humans. The specific benefits of involving AI into medicine include accurate data can inform specialists about typical patterns, AI can perform as well as a human does and nullifies stress and exhaustion factors, data sets can train machine learning algorithms and models to address key drug production problems which would help in curing more people under lower cost and preserving the personalized approach.

AI uses sophisticated algorithms to extract, learn, predict, and foresee from huge amounts of medical data while also providing professional support and assistance. In regards to diseases, cancer, cardiovascular, and nervous system disorders are the most frequently researched involving ML tools. Self-trained systems can follow supervised and unsupervised learning, facilitating early detection and diagnosis greatly. To perform well, self-trained systems should interact constantly with the clinical studies data, so it's obvious that human activity is interconnected with machine learning.

Healthcare industry has become big business. The healthcare industry produces large amounts of health-care data daily that can be used to extract information for predicting disease that can happen to a patient in future while using the treatment history and health data. This hidden information in the healthcare data will be later used for affective decision making for patient's health. Also, this area need improvement by using the informative data in healthcare. Major challenge is how to extract the information from these data because the amount is very large so some data mining and machine learning techniques can be used. Also, the expected outcome and scope of this project is that if disease can be predicted than early treatment can be given to the patients which can reduce the risk of life and save life of patients and cost to get treatment of diseases can be reduced up to some extent by early recognition. The rapid adoption of electronic health records has created a wealth of new data about patients, which is a goldmine for improving the understanding of human health.

When machine learning is employed in aid to supplement taking care of patients, high results are achieved. It has made it easier to spot various types of diseases and perform diagnoses accurately. Performing predictive analysis with the assistance of multiple efficient machine learning algorithms may facilitate predicting any disease with great accuracy and help to treat patients. The huge amount of medical information containing treatment history and health data are often used to extract data for predicting diseases that may happen to a patient within the future. The hidden data within the medical information are often later used for an effective decision-making process for the patient's health.

One of the foremost vital applications of machine learning is within the field of healthcare. The healthcare facilities have to be compelled to be advanced so that better decisions for patient treatment are often made. Once machine learning is employed in healthcare, it helps individuals to process vast and complex disease datasets to analyze them into helpful clinical insights. Then this will be further employed by medical practitioners to provide accurate treatment to patients. Hence, machine learning, once enforced in healthcare will result in high patient satisfaction. In this paper, the logistic regression algorithm will be used to predict diseases using the patient's treatment history and health data.

Logistic Regression algorithm

The logistical regression is also referred to as the sigmoid function that helps easy representation of graphs. It additionally provides high accuracy. In this algorithm, the data should be first imported and then it should be trained. It is a type of regression analysis algorithm, which is used for prediction of the outcome of a categorical dependent variable from a set of independent or predictor variables.

The key representation in logistic regression are the coefficients, just like linear regression. The coefficients in logistic regression are estimated using a process called maximum-likelihood estimation. In logistical regression, the variable quantity is usually binary. It is principally used for prediction and also calculating the probability. Logistic regression is used because it is an efficient regression predictive analysis algorithm that doesn’t discard any data and uses all data efficiently.

1.2 PURPOSE

Disease prediction using patient treatment history and health data by applying data mining and machine learning techniques is ongoing struggle for the past decades. Many works have been applied data mining techniques to pathological data or medical profiles for prediction of specific diseases. These approaches tried to predict the reoccurrence of disease. Also, some approaches try to do prediction on control and progression of disease. The recent success of deep learning in disparate areas of machine learning has driven a shift towards machine learning models that can learn rich, hierarchical representations of raw data with little pre-processing and produce more accurate results. Numbers of papers have been published on several data mining techniques for diagnosis of heart disease such as Decision Tree, Naive Bayes, neural network, kernel density, automatically defined groups, bagging algorithm and support vector machine showing different levels of accuracies in diseases prediction.

1.3 MOTIVATIONS AND SCOPE

At present in order to remain healthy, regular body diagnosis is necessary. Today, there are multiple sources available as individual prediction or recommendation system but the need of the hour is to have an integrated model comprising both. Also, it would be more appropriate and convenient if people could get basic diagnosis online 24x7 rather than visiting hospitals & clinics frequently. Thus, reducing cost and saving time. If certain anomalies found in the diagnosis then recommendation of nearby specialist and hospitals according to user’s preference would facilitate in quick and appropriate treatment. Healthcare being a domain evolving continuously and generating a huge amount of data develops a need to use the data for useful knowledge which attracts large organizations to invest heavily in this field.

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. The application permits user to share their heart connected problems. It then processes user specific details to ascertain for varied illness that might be related to it. Here we tend to use some intelligent data mining techniques to guess the foremost correct illness that might be related to patient’s details. Based on result, system automatically shows the result specific doctors for more treatment. The system permits user to look at doctor’s details and can also be used in case of emergency.

3. EXISTING SYSTEM

In the existing system, practical use of various collected data is time consuming, machine can predict diseases but cannot predict the sub types of the diseases caused by occurrence of one disease. It fails to predict all possible conditions of the people. Existing system handles only structured data. A machine can detect a disease but cannot expect the sub types of the diseases and diseases caused by the existence of one bug. The predictions of diseases have been non-specific and indefinite. For occurrence, if a group of people are foreseen with Diabetes, doubtless some of them might have complex risk for Heart diseases due to the actuality of Diabetes. Diagnosis of the condition solely depends upon the doctor's intuition and patient's records. Detection is not possible at an earlier stage that might later potentially harm the patient.

4. PROPOSED SYSTEM

The proposed system has been developed to classify people, who are stricken by disease and healthy people. The performance of the predictive model with selected features is tested to predict the probabilities of suffering from heart disease. Feature selection algorithm was used to select important features, and on these selected features, the performance of the classifiers was tested. The Framingham heart condition dataset is taken from Kaggle and has been employed in our study. The popular machine learning classifier logistic regression is employed within the system. The model's validation and performance evaluation metrics are computed. It is flexible and can be widely used for various diseases with high rates of success. The methodology of the proposed system is structured into five stages which include:

(1) pre-processing of a dataset, (2) feature selection, (3) cross-validation method, (4) machine learning classifiers, and (5) classifiers' performance evaluation methods.

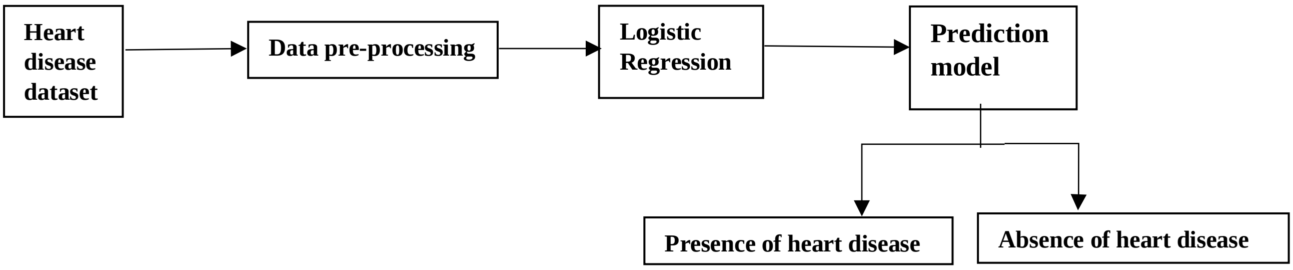
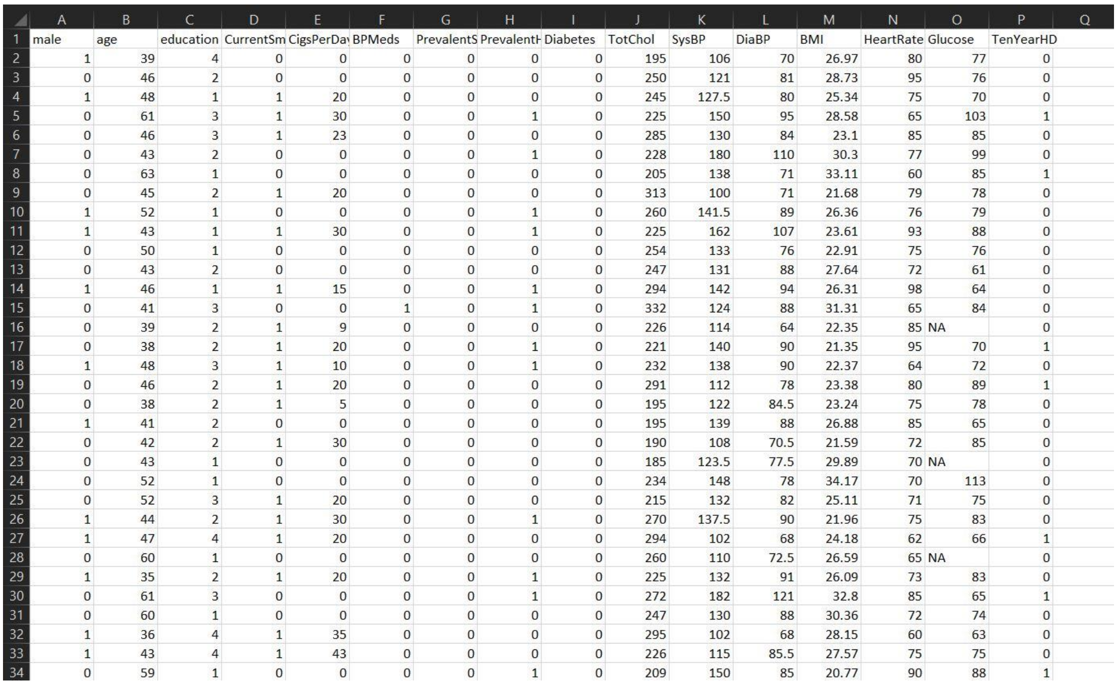


Figure 1: Disease predicting model framework for predicting heart disease

5. IMPLEMENTATION

The proposed system is implemented on a heart disease dataset, which is available on the Kaggle website and it contains a cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a risk of future heart disease(HD) in 10 years or not. The dataset provides the patients' information like demographic, behavioral and medical risk factors. The dataset contains over 4,000 records and about 15 attributes.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	male	age	education	CurrentSm	CigsPerDay	BPMeds	PrevalentS	PrevalentH	Diabetes	TotChol	SysBP	DiaBP	BMI	HeartRate	Glucose	TenYearHD	
2	1	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77	0	
3	0	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76	0	
4	1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0	
5	0	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103	1	
6	0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0	
7	0	43	2	0	0	0	0	1	0	228	180	110	30.3	77	99	0	
8	0	63	1	0	0	0	0	0	0	205	138	71	33.11	60	85	1	
9	0	45	2	1	20	0	0	0	0	313	100	71	21.68	79	78	0	
10	1	52	1	0	0	0	0	1	0	260	141.5	89	26.36	76	79	0	
11	1	43	1	1	30	0	0	1	0	225	162	107	23.61	93	88	0	
12	0	50	1	0	0	0	0	0	0	254	133	76	22.91	75	76	0	
13	0	43	2	0	0	0	0	0	0	247	131	88	27.64	72	61	0	
14	1	46	1	1	15	0	0	1	0	294	142	94	26.31	98	64	0	
15	0	41	3	0	0	1	0	1	0	332	124	88	31.31	65	84	0	
16	0	39	2	1	9	0	0	0	0	226	114	64	22.35	85	NA	0	
17	0	38	2	1	20	0	0	1	0	221	140	90	21.35	95	70	1	
18	1	48	3	1	10	0	0	1	0	232	138	90	22.37	64	72	0	
19	0	46	2	1	20	0	0	0	0	291	112	78	23.38	80	89	1	
20	0	38	2	1	5	0	0	0	0	195	122	84.5	23.24	75	78	0	
21	1	41	2	0	0	0	0	0	0	195	139	88	26.88	85	65	0	
22	0	42	2	1	30	0	0	0	0	190	108	70.5	21.59	72	85	0	
23	0	43	1	0	0	0	0	0	0	185	123.5	77.5	29.89	70	NA	0	
24	0	52	1	0	0	0	0	0	0	234	148	78	34.17	70	113	0	
25	0	52	3	1	20	0	0	0	0	215	132	82	25.11	71	75	0	
26	1	44	2	1	30	0	0	1	0	270	137.5	90	21.96	75	83	0	
27	1	47	4	1	20	0	0	0	0	294	102	68	24.18	62	66	1	
28	0	60	1	0	0	0	0	0	0	260	110	72.5	26.59	65	NA	0	
29	1	35	2	1	20	0	0	1	0	225	132	91	26.09	73	83	0	
30	0	61	3	0	0	0	0	1	0	272	182	121	32.8	85	65	1	
31	0	60	1	0	0	0	0	0	0	247	130	88	30.36	72	74	0	
32	1	36	4	1	35	0	0	0	0	295	102	68	28.15	60	63	0	
33	1	43	4	1	43	0	0	0	0	226	115	85.5	27.57	75	75	0	
34	0	59	1	0	0	0	0	1	0	209	150	85	20.77	90	88	1	

Figure 2: an imported dataset from Kaggle

Then we make a choice for choosing the variable quantity and variable quantity, where each attribute is taken into account as a possible risk factor. There are several demographic, behavioral and medical risk factors involved.

Demographic:

sex: male or female(Nominal)

age: age of the patient(Continuous)

Behavioral:

CurrentSmoker- whether or not the patient may be a current smoker (Nominal)

CigsPerDay: the number of cigarettes that the person smoked on the average in one day(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

Medical(history):

BPMeds: whether or not the patient was on vital sign medication (Nominal)

PrevalentStroke: whether or not the patient had a stroke before (Nominal)

PrevalentHyp: whether the patient was hypertensive or not (Nominal)

Diabetes: whether the patient had diabetes or not (Nominal)

Medical(current):

TotChol: Total Cholesterol level (Continuous)

SysBP: Systolic Blood Pressure (Continuous)

DiaBP: Diastolic Blood Pressure (Continuous)

BMI: Body Mass Index (Continuous)

HeartRate: pulse rate (Continuous - In medical research, variables like pulse rate though after all discrete, yet are considered continuous thanks to an outsized number of possible values.)

Glucose: Glucose level (Continuous)

Predict variable (desired target): risk of future heart disease(HD) in 10 years (binary: “1” means “Yes” and “0” means “No”)

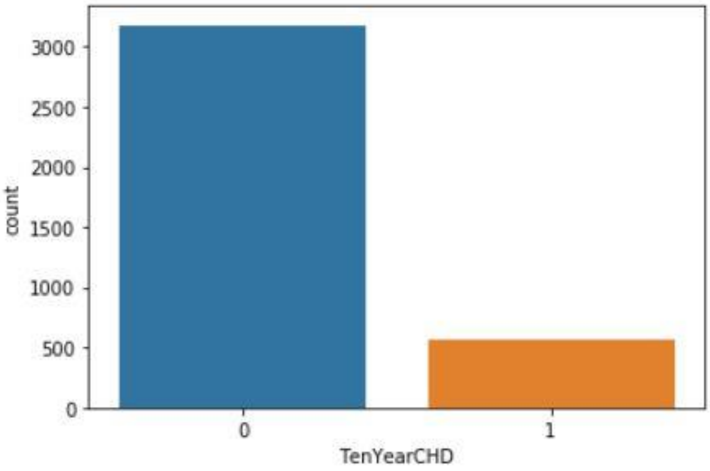


Figure 3. Number of people with history of heart disease

The above figure shows the medical record of 4,000 people out of which about 3,500 people have not suffered from cardiovascular disease in the past whereas 5,000 people have suffered from cardiovascular disease.

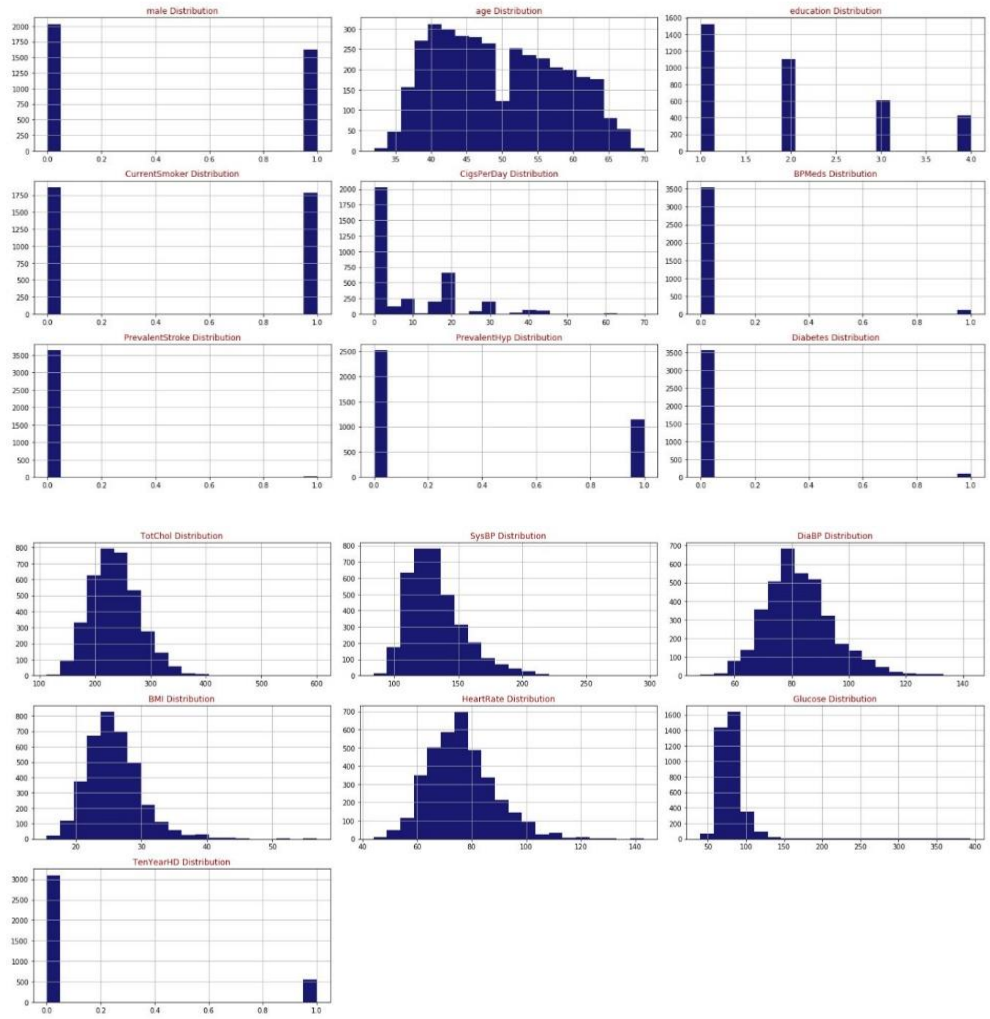
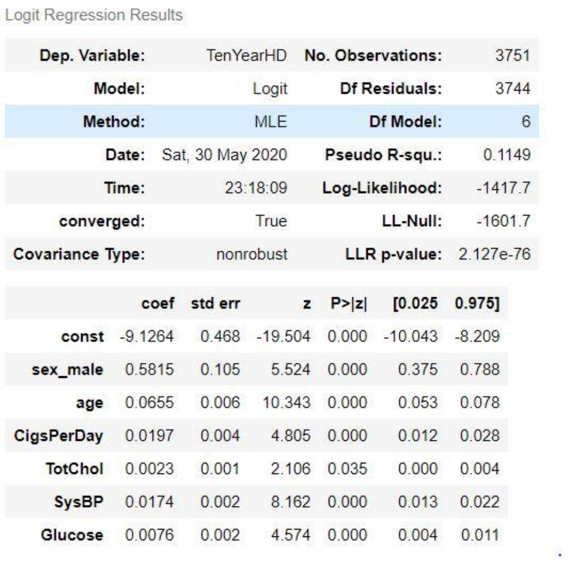


Figure 4: independent and dependent variables

The heart disease dataset is then split into two subsets i.e. training data and testing data and that we fit our model on train data to form predictions on the test data After that two things can end up happening, we might overfit our model or we might underfit our model. Any of those things happening would affect the predictability of our model, so we might find ourselves employing a model with lower accuracy. For the heart disease dataset, the training set is taken as 80% of the actual data and test set as 20% of the data.



Logit Regression Results

Dep. Variable:	TenYearHD	**No. Observations:**	3751
Model:	Logit	**Df Residuals:**	3744
Method:	MLE	**Df Model:**	6
Date:	Sat, 30 May 2020	**Pseudo R-squ.:**	0.1149
Time:	23:18:09	**Log-Likelihood:**	-1417.7
converged:	True	**LL-Null:**	-1601.7
Covariance Type:	nonrobust	**LLR p-value:**	2.127e-76

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-9.1264	0.468	-19.504	0.000	-10.043	-8.209
sex_male	0.5815	0.105	5.524	0.000	0.375	0.788
age	0.0655	0.006	10.343	0.000	0.053	0.078
CigsPerDay	0.0197	0.004	4.805	0.000	0.012	0.028
TotChol	0.0023	0.001	2.106	0.035	0.000	0.004
SysBP	0.0174	0.002	8.162	0.000	0.013	0.022
Glucose	0.0076	0.002	4.574	0.000	0.004	0.011

Figure 5: Logistic regression results using backward elimination (P value approach)

The results above show some of the attributes with P value higher than the preferred alpha (5%) and thereby showing low statistically significant relationship with the probability of heart disease. We use backward elimination approach to remove those attributes with highest P value one at a time, then the regression is run repeatedly until all attributes have P Values less than 0.05. An attribute having P value less than 0.05 shows that the change in its value will cause change in the odds of having a heart disease.

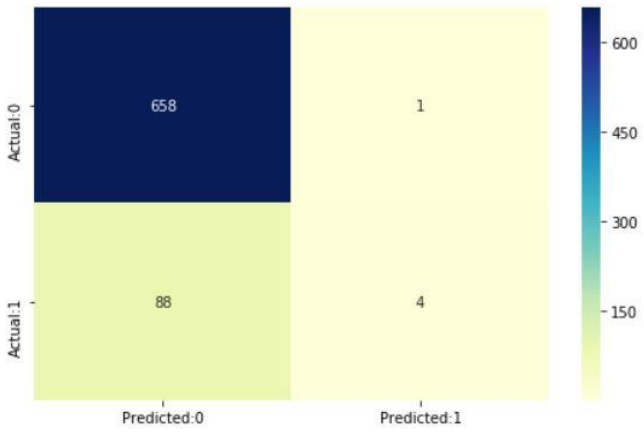


Figure 6: Confusion matrix for model evaluation

The confusion matrix shows that the model made 662 correct predictions and 89 incorrect ones. From the above statistics it can be made clear that the model is highly specific than sensitive. The negative values are predicted more accurately than the positives. Also, the model achieves an accuracy rate of 88%.

# CODE

# importing libraries

import pandas as pd

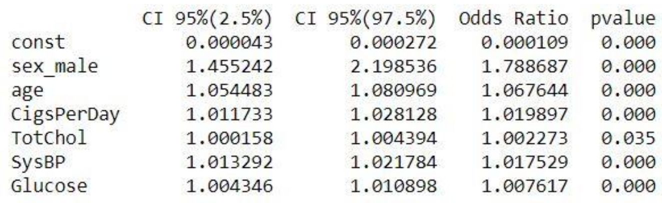
```
import numpy as np
import statsmodels.api as sm
import scipy.stats as st
import matplotlib.pyplot as plt
import seaborn as sn
from sklearn.metrics import confusion_matrix
import matplotlib.mlab as mlab
%matplotlib inline
# importing dataset
heart_df=pd.read_csv("C:/Users/Shobit/disease prediction/framingham.csv")
# removing the row/columns which are irrelevant
heart_df.drop(['education'],axis=1,inplace=True)
# removing all null values from given data
count=0
for i in heart_df.isnull().sum(axis=1):
    if i>0:
        count=count+1
print('Total number of rows with missing values is ', count)
print('since it is only',round((count/len(heart_df.index))*100), 'percent of the entire dataset the rows with missing values are excluded.')
heart_df.dropna(axis=0,inplace=True)
# Data exploratory analysis of cleaned data
def draw_histograms(dataframe, features, rows, cols):
    fig=plt.figure(figsize=(20,20))
    for i, feature in enumerate(features):
        ax=fig.add_subplot(rows,cols,i+1)
        dataframe[feature].hist(bins=20,ax=ax,facecolor='midnightblue')
        ax.set_title(feature+" Distribution",color='DarkRed')
    fig.tight_layout()
    plt.show()
draw_histograms(heart_df,heart_df.columns,6,3)
# TenYearHD is made dependent variable and others independent variable
```

```
heart_df.TenYearHD.value_counts()
sn.countplot(x='TenYearHD',data=heart_df)
# Logistic regression
from statsmodels.tools import add_constant as add_constant
heart_df_constant = add_constant(heart_df)
heart_df_constant.head()
st.chisqprob = lambda chisq, df: st.chi2.sf(chisq, df)
cols=heart_df_constant.columns[:-1]
model=sm.Logit(heart_df.TenYearHD,heart_df_constant[cols])
result=model.fit()
result.summary()
# Feature selection: backward elimination (P value approach)
def back_feature_elem (data_frame,dep_var,col_list):
    """ Takes in the dataframe, the dependent variable and a list of column names, runs the
regression repeatedly eleminating feature with the highest
    P-value above alpha one at a time and returns the regression summary with all p-values
below alpha"""

    while len(col_list)>0 :
        model=sm.Logit(dep_var,data_frame[col_list])
        result=model.fit(disp=0)
        largest_pvalue=round(result.pvalues,3).nlargest(1)
        if largest_pvalue[0]<(0.05):
            return result
            break
        else:
            col_list=col_list.drop(largest_pvalue.index)
result=back_feature_elem(heart_df_constant,heart_df.TenYearHD,cols)
result.summary()
# Regression result
params = np.exp(result.params)
conf = np.exp(result.conf_int())
conf['OR'] = params
```

```
pvalue=round(result.pvalues,3)
conf['pvalue']=pvalue
conf.columns = ['CI 95%(2.5%)', 'CI 95%(97.5%)', 'Odds Ratio','pvalue']
print ((conf))
# Data is splitted with 80:20 and then the model is trained through logistic regression
import sklearn
x=heart_df.iloc[:,:-1]
y=heart_df.iloc[:,-1]
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=.20,random_state=5)
from sklearn.linear_model import LogisticRegression
logreg=LogisticRegression()
logreg.fit(x_train,y_train)
y_pred=logreg.predict(x_test)
# Model evaluation
sklearn.metrics.accuracy_score(y_test,y_pred)
# Confusion Matrix for total no. of correct and incorrect prediction
from sklearn import metrics
cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix
from sklearn.metrics import confusion_matrix
cm=confusion_matrix(y_test,y_pred)
conf_matrix=pd.DataFrame(data=cm,columns=['Predicted:0','Predicted:1'],index=['Actual:0','
Actual:1'])
plt.figure(figsize = (8,5))
sn.heatmap(conf_matrix, annot=True,fmt='d',cmap="YlGnBu")
```

6. OUTPUT



```
            CI 95%(2.5%)  CI 95%(97.5%)  Odds Ratio  pvalue
const           0.000043       0.000272    0.000109   0.000
sex_male        1.455242       2.198536    1.788687   0.000
age             1.054483       1.080969    1.067644   0.000
CigsPerDay      1.011733       1.028128    1.019897   0.000
TotChol         1.000158       1.004394    1.002273   0.035
SysBP           1.013292       1.021784    1.017529   0.000
Glucose         1.004346       1.010898    1.007617   0.000
```

Figure 7: Effect in odds of heart disease

The fitted model shows that, holding all other features constant, the odds of suffering from cardiovascular disease for males are 78.8% higher than the odds for females. The coefficient for age says that, holding all others constant, we will see 6.76% increase in the odds of suffering from cardiovascular disease with one year increase in age. Similarly, with every extra cigarette one smokes there is a 2% increase in the odds of getting cardiovascular disease. For Total cholesterol level and glucose level there is no significant change. Also, there is a 1.7% increase in odds for every unit increase in systolic Blood Pressure.

7. CONCLUSION

It has been concluded that men are more prone to heart disease than women. An increase in age, along with the number of cigarettes smoked per day and systolic blood pressure also show increased odds of getting a cardiovascular disease. Total cholesterol shows no significant change in the odds of Heart Disease (HD), this could be due to the presence of good cholesterol in the cholesterol reading. Similarly, glucose too causes a very negligible change in odds (0.2%).

The future enhancement of the proposed system will result in prediction diseases by using advanced techniques and algorithms in less time complexity. An intelligent system may be developed using the proposed model that can lead to the selection of proper treatment methods. Data analysis and Machine learning can be of very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable databases.

8. REFERENCES

[1] https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset

[2] Akash C. Jamgade and Prof. S. D. Zade, “Disease Prediction Using Machine Learning”, International Research Journal of Engineering and Technology, Vol. 5, Issue 6, May 2019.

[3] Vinitha S, Sweetlin S, Vinusha H and Sajini S, “Disease prediction using machine learning over Big Data”, Computer Science & Engineering: An International Journal, Vol. 8, No. 1, February 2018.

[4] S. Patel and H. Patel, “Survey of data mining techniques used in healthcare domain”, Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March 2016.

[5] M. Abinaya, M. Marimuthu, K.S. Hariesh, K. Madhankumar and V. Pavithra, “A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach”, International Journal of Computer Applications, Vol. 181, No. 18, September 2018.

[6] Min Chen, Yixue Hao, Kai Hwang, Lu Wang and Lin Wang, “Disease Prediction by Machine Learning Over Big Data From Healthcare Communities”, IEEE, Vol. 5, April 2017.

[7] Tarigoppula V.S Sriram, M. Venkateswara Rao, G V Satya Narayana, DSVGK Kaladhar and T Pandu Ranga Vital, “Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms”, International Journal of Engineering and Innovative Technology, Vol. 3, Issue 3, September 2013.

[8] https://www.machinelearningplus.com/statistics/p-value/

[9] Reddy Prasad, Pidaparthi Anjali, S. Adil, N. Deepa, “Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning”, International Journal of Engineering and Advanced Technology, Vol. 8, Issue 3S, February 2019.