Online Disease Detection and Prediction Analysis

A PROJECT REPORT

for

Artificial Intelligence – CSE3013

in

B.TECH- INFORMATION TECHNOLOGY

by

YADHU ANAND K J: 18BIT0373

SIDDHARTH DAS: 18BIT0379

SHRUTI VARSHA VENKATRAMAN: 18BIT0405

HARIDA P K: 18BIT0411

Under the Guidance of

Dr.HARSHITA PATEL

Associate Professor, SITE

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

November, 2020

DECLARATION BY THE CANDIDATE

We hereby declare that the project report entitled "**Online Disease Detection and Prediction Analysis**" submitted by us to Vellore Institute of Technology University, Vellore in partial fulfilment of the requirement for the award of the course **Artificial Intelligence(CSE 3013)** is a record of bonafide project work carried out by us under the guidance of **Dr. Harshita Patel** We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

Place: Vellore

Date: November 2020

Signature

Harida PK

Siddharth Das

Yadhu Anand KJ

Shruti Varsha

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology & Engineering [SITE]

CERTIFICATE

This is to certify that the project report entitled "**Online Disease Detection and Prediction Analysis"** submitted by **Yadhu Anand K J(18BIT0373), Siddharth Das(18BIT0379), Shruti Varsha Venkatraman(18BIT0405)** and **Harida PK(18BIT0411)** to Vellore Institute of Technology University, Vellore in partial fulfilment of the requirement for the award of the course **Artificial Intelligence (CSE3013)** is a record of bonafide work carried out by them under my guidance.

Dr. HARSHITA PATEL

GUIDE

Associate Professor, SITE

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Fall Semester 2020-2021

TABLE OF CONTENTS

Online Disease Detection and Prediction Analysis

ABSTRACT

Health is a vital topic which is connected with our human lives. Healthcare sectors come under this as it disperses into a broader scale. Information and Communication Technology plays an important role in the healthcare sector because modern technology assists to furnish effective and efficient services. The impact of Information Technology in healthcare is immense. Therefore, we propose an online disease detection system that delivers quick guidance in absence of human expertise. Generally, it interprets the user's input by natural language processing but some simpler systems search for keywords within the text and then provide a reply based on the matching keywords or certain patterns. The proposed system will have English as the default language as it would increase the scope for the global community to benefit from it. This system will be embedded into a website to maximize its customer outreach and usage. It will also predict lifestyle-based diseases like heart diseases, diabetes and breast-cancer. The initial model of the project will contain a chatbot which will predict the disease according to the symptoms of the patient and recommend a specialized doctor for respective diseases.

OBJECTIVES

- Virtual Medical Assistant- Analyze symptoms and predict ailments
- Increasing accuracy by asking patients which symptoms bothers you the most.

- Asking Patient about seeing a medical professional for this
  and connecting directly with the specialized doctor.

- Perform descriptive analysis on disease prediction using key factors like Glucose levels, Blood Pressure, Skin Thickness, BMI etc.

- The main Objective is to predict through diagnosis whether a patient has diabetes, breast-cancer or heart disease, based on certain diagnostic measurements included in the dataset.

  Many limitations were faced during the selection of the occurrences from the bigger dataset. The type of dataset and problem is a classic supervised binary classification. The dataset has: 9 = 8 + 1 (Class Attribute) attributes, 768 records describing female patients (of which there were 500 negative instances (65.1%) and 268 positive instances (34.9%)).

- Diagnosis of lifestyle-based diseases like diabetes, breast-cancer and cardiovascular diseases is considered a challenging problem for quantitative research. Some parameters were shown to be ineffective due to some limitations. A single parameter is not very effective to accurately diagnose these diseases and may be misleading in the decision-making process. So, the main objective is to combine different parameters effectively predict these diseases at an early stage. When different parameters were used for prediction of disease, it is predicted with the assistance of significant attributes, and the association of the differing attributes. We will examine the diagnosis using ANN, RF and K-Means Clustering and do a detailed predictive analysis on this subject matter.

INTRODUCTION

Science has made great advancements to enhance the interaction between humans and machines by leaps and bounds. An artificial intelligence software that can simulate a conversation with a user in natural language. Smart interactions save the user's time by helping them to find the right information and address their queries. This technology has made great progress in the healthcare industry. With the help of this algorithm we will be contributing to the betterment of the society. The model is designed to predict the result according to the symptoms with better accuracy. In the initial phase, we are making a chatbot that would diagnose the disease based upon the symptoms exhibited by the patient. The patient would be required to enter the symptoms and the model would accurately predict the disease based upon a dataset that has been prepared carefully from a large number of sources.

On the other hand, there has been a drastic increase in the rate of people suffering from lifestyle-based diseases since a decade. Current human lifestyle is the main reason behind growth of diseases like Diabetes, Breast Cancer and Cardiovascular diseases.

Diabetes is a common, chronic disease. Prediction of diabetes at an early stage can lead to improved treatment. Data mining techniques are widely used for prediction of disease at an early stage. In this research paper, diabetes is predicted using significant attributes, and the relationship of the differing attributes is also characterized. Various tools are used to determine significant attribute selection, and for clustering, prediction, and association rule mining for diabetes.

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society. The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modelling.

Heart disease is the major cause of morbidity and mortality globally: it accounts for more deaths annually than any other cause. Of all heart diseases, coronary heart disease (aka heart attack) is by far the most common and the most fatal. The silver lining is that heart attacks are highly preventable and simple lifestyle modifications (such as reducing alcohol and tobacco use; eating healthily and exercising) coupled with early treatment greatly improves its prognosis. It is, however, difficult to identify high risk patients because of the multifactorial nature of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, et cetera. This is where machine learning and data mining come to the rescue. Doctors and scientists alike have turned to machine learning (ML) techniques to develop screening tools and this is because of their superiority in pattern recognition and classification as compared to other traditional statistical approaches.

In the current medical diagnosis method, there can be three different types of errors.

1-The false-negative type in which a patient in reality is already a diseased patient but test results tell that the person is not having the disease.

2. The false-positive type. In this type, a patient in reality is not a diseased patient but test reports say that he/she is having the disease.

3. The third type is an unclassifiable type in which a system cannot diagnose a given case.

This happens due to insufficient knowledge extraction from past data, a given patient may get predicted in an unclassified type. Such errors in diagnosis may lead to unnecessary treatments or no treatments at all when required. In order to avoid or reduce severity of such impact, there is a need to create a system using machine learning algorithms and data mining techniques which will provide accurate results and reduce human efforts.

LITERATURE SURVEY

[1] Medical ChatBot- "Mrs. Rashmi Dharwadkar1, Dr.Mrs. Neeta A. Deshpande". International journal of computer trends and technology (IJCTT).

According to the author, the paper gives the information regarding products which is useful for consumers to obtain what they want exactly. Question Answering (QA) systems can be identified as information accessing systems which try to answer natural language queries by giving suitable answers making use of attributes available in natural language techniques. The system takes a plain text as input and answering all types of questions output by qualified user is the output. It helps the user to resolve the issue by providing human way interactions using LUIS and cognitive services which is implemented on AWS public cloud. The main of this paper is to provide a generic solution to this problem and build up a system which is useful for medical institutes or hospitals to help the users to freely ask medical dosage related queries by voice.

[2] Early Diagnosis of Human Disease using Artificial Intelligence Anil Kumar and Hemant Kumar Soni

The Most common fuzzy methodology i.e. Mamdani Fuzzy Inference model is used in this proposed work. This methodology includes fuzzification of crisp data, inferring using knowledge base and rule base and defuzzification of fuzzy output to obtain crisp values. Fuzzification involves reading crisp data and mapping or translates it into fuzzy data for the use of inference system. Inferring includes reading fuzzy data and processes them on the basis of fuzzy rules and knowledge. The result obtained from inferring shows fuzziness. Thus in defuzzification, the result which is in fuzzy form is again map or translates into crisp values so that it becomes accurate and fruitful. The existing study reveals the importance of artificial intelligence in the diagnosis of human disease. The above literature review envisages the need for an expert system in different disease diagnosis. The accuracy of predictions based on the sample size of a patient's data poses the challenges in this domain. The use of methodologies under AI which ultimately leads the correct diagnosis motivates us to explore future work in this domain. This study helped us to find the possibilities of the design of Expert System (ES) based on paradigms of AI which would yield better results and accuracy. Although there are different paradigms of AI like Neural Network, Machine Learning, Deep Learning, and Fuzzy Logic the application of Fuzzy logic is versatile and gives the accurate prediction where ever approximation diagnosis is required.

[3] A Tool of Conversation: Chatbot- "M. Dahiya". Automated Medical Chatbot. SSRN Electronic Journal.

A Chatbot is actualized utilizing design looking at, in which the request for the sentence is perceived and a spared reaction design is adjusted to the elite factors of the sentence. Chatbot is relatively new technology. The application of a Chatbot can be seen in various fields in the future. This paper covers the techniques used to design and implement a Chatbot. From the author point of view their Chatbot uses simple pattern matching to represent the input and output whereas other Chatbots use input rules, keyword patterns and output rules to generate a response. Their main aim was to develop a chatbot that is simple, user friendly, must be easily understood and the knowledge base must be compact.

[4]CHATBOT ACCEPTANCE IN HEALTHCARE: EXPLAINING USER ADOPTION OF CONVERSATIONAL AGENTS FOR DISEASE DIAGNOSIS- "Sven Laumer, Christian Maier, Fabian Tobias Gubler". In Proceedings of the 27th European Conference on Information Systems (ECIS), Stockholm & Uppsala, Sweden, June 8-14, 2019.

The author develops a model to introspect whether adoption of conversational agents for disease diagnosis should be promoted or not. The proposed model addressed research gaps in CA research in general, but also in medical health and especially the use of CA in healthcare research in particular. To build the model they used UTAUT2 as a theoretical lens and 35 semi structured interviews with potential users of a CA for disease diagnosis. In their model they revealed that hedonic motivation is not relevant for CA adoption.

[5] Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning-“ Rohit Binu Mathew ; Sandra Varghese ; Sera Elsa Joy ; Swanthana Susan Alex. 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI).

Individuals can associate with the chatbot simply as they do with another human and through a progression of questions; chatbot will distinguish the manifestations of the client and consequently, predicts the symptoms and suggests treatment. This framework can be of extraordinary use to individuals in directing every day registration, makes individuals mindful of their well-being status and urges individuals to make legitimate measures to stay sound. According to this research, such a framework isn't broadly utilized and individuals are less mindful of it. Executing this proposed system can assist individuals with dodging the tedious strategy for visiting emergency clinics by utilizing this liberation from cost application at any place.

[6] Manish Bali , Samahit Mohanty, Subarna Chatterjee, Manash Sarma,Rajesh Puravankara have proposed their work on **Diabot: A Predictive Medical Chatbot using ensemble learning.** They have combined various machine learning algorithms and text pre-processing techniques in a voting ensemble. The advantage of using this is that they have a high performance framework for future disease specific chatbot designs and they combine all the weak classifiers with quantitative performance using meta classifiers. Diagnosis decisions based on the classification result alone will be very weak. Disadvantage of this model is that ensemble learning is very weak in terms of accuracy. As a part of future work we can concentrate on the ingesting patient directly into the diabot server using IOT sensors and build a recommendation engine.

[7] Josip Bozic,Oliver A.Tazl,Franz Wotava have proposed the **Chatbot Testing Using AI Planning**. This is based on AI where each action is considered to be a question to the chatbot.The disadvantage of this model is that the chatbot remains in the mode even if the user responds in an unexpected way. They use a planning based test approach where they create a model which serves as a test case generation. The model of SUT is extracted which is used to plan specifications. Advantage of using this is that when multiple information is being submitted at once it doesn't influence the order. As a future work they can be used for other functional and non-functional testing.

[8] Deepika Joshi, Renu Kant and Sachin Shakya have proposed their work on **Disease prediction using machine learning in a chatbot**. This paper consists of various ML algorithms to predict the disease symptoms along with their diabetics status through the information given by patients. The advantage of this method is that with the help of the modules the bot can predict the disease using a decision tree algorithm as they are proven to give better accuracy.The disadvantage can be mentioned as the computational power of the bot where every time the response time decreases.

[9] Gajendra Prasad K.C, Tathgat Ankit, Satvik Ranjan and Vivek kumar have proposed their work on **A Personalised Medical Assistant Chatbot: Medibot.** The algorithms used here are Sequence to Sequence Model, Apriori which is used in finding the frequent item set in the given dataset. So a chatbot is used to analyse the self-diagnosis. This helps the individual to keep track of the field of medical science for early and faster detection of diseases. The disadvantage is that the lack of correct and accurate medical dataset also in the seq2seq model is very time consuming during the training phase using the hardware which is not capable of handling it.also in the model accurate heart rate,BMI can be given as input to the system and the disease prediction also can be improved.

[10] Flora Amato, Stefano Marrone and Vincenzo Moscato have proposed their work on **Chatbots meet eHealth: automatisation healthcare**. The aim of this work is to showcase the effectiveness of the human intervention paradigms. The biggest challenge is that for modern eHealth applications is to provide intelligent recommender systems which leverage different kinds of available data. The design allows us to adapt it to different clinical scenarios and medical tasks. They implemented various techniques which plan on spreading HOLMeS in different countries for helping the medical knowledge sharing. When interaction requires elaborating machine learning algorithms of accessing the data storage the core system contacts the cluster for accomplishing the specific tasks.

[11]An Advanced Conceptual Diagnostic Healthcare Framework for Diabetes and Cardiovascular Disorders M. Sharma,1 *,G. Singh2 and R. Singh

A hybrid artificial intelligent and smart framework based upon Data Mining, IoT, chatbots, contextual entity search, sentiment analysis and granular computing has been proposed.Therefore, here, a smart data mining and IoT (SMDIoT)based progressed medicinal services framework for capable diabetes and cardiovascular maladies have been proposed. The hybridization of information mining and IoT strategies should give a viable and affordable answer for diabetes and cardiac patients. With the utilization of SMDIoT, fundamental clinical assets and labor will be better used; subsequently, the patient consideration will be additionally upgraded. Notwithstanding social insurance, it will significantly decrease financial misfortunes of diabetes and cardiac patients by limiting different verifiable and express clinical costs.

[12] A PAPER ON CHATBOT FOR MEDICAL DIAGNOSIS ANKIT GARG RAJAT JINDAL SUNNY ASHISH

The suggested Medical Chatbot can interconnect with users and gives them the same virtual realistic experience of visiting a Medical Professional. This Chatbot firstly uses the text classification to detect the intent of the user also known as intent classification and then detects the pattern of the response by the user using AIML (Artificial Intelligence Markup Language) technology. AIML is the markup language which is based on XML. AIML is used to build AI applications. AIML first reduces the message to the few keywords in that message. It retrieves them from the initial messages and then know about the possible health issue/problems that unconfirmed patients might be having, based on the symptoms shown by the user.

[13] Machine Learning and Artificial Intelligence based Diabetes Mellitus Detec- tion and Self-Management: A Systematic Review Jyotismita Chaki, S. Thillai Ganesh, S.K Cidham, S. Ananda Theertan

This paper gave an exhaustive investigation of programmed diabetic location and determination procedures. The fundamental articles are gathered from Scopus and PubMed logical archives. After an exhaustive screening measure, 107 examinations are picked for this investigation. Each exploration is tended to in this investigation from the perspective of four explicit viewpoints, including information bases, ML-based arrangement, and indicative strategies, AI-based wise collaborators for patients with DM, and execution measurements. A few freely open information bases with explicit qualities have been depicted and archived in this investigation. Among these datasets, Pima Indians Diabetes Dataset, DIARETDB1, DIARETDB0, Kaggle, STARE, and Messidor are most generally used for DM identification. Text, shape, and surface element created better results. In ML calculations, the majority of the investigations expressed that Deep Neural Network and Support Vector Machine conveys better order results followed by random forest and Ensemble Classifier. CNN is fundamentally utilized in profound figuring out how to naturally recover and distinguish DM information. Numerous analysts have created diverse astute colleagues like chatbots and robots which can be utilized to upholds the day by day DM the executives cycles of patients like insulin management, diet checking, and so forth. Regarding performance assessment, most of the researchers used the accuracy, specificity, sensitivity, and AUC as indicators.

[14] Symptom Based Health Prediction using Data Mining Vijaya Shetty S Department of Computer Science and Engineering Nitte Meenakshi Institute Of Technology Bengaluru,India Karthik G A Department of Computer Science and Engineering Nitte Meenakshi Institute Of Technology Bengaluru,India M Ashwin Department of Computer Science and Engineering Nitte Meenakshi Institute Of Technology Bengaluru,India

The proposed model utilizes the capability of different Machine learning algorithms combined with text processing to achieve accurate prediction. Text processing has been implemented using Tokenization and, is combined with various algorithms to test the similarities and the outputs. In the health industry, it provides several benefits such as pre-emptive detection of diseases, faster diagnosis, medical history for review of patients, etc. The dataset used had a lot of cleaning and preprocessing needed to be done. The first step was to transpose the datasets into a form with diseases as the target column and each of the symptoms as dummy variables for the diseases, on which models were to be trained. For the study, three different sets of algorithms were used for mapping symptoms into diseases: Decision Trees, Random Forest, and Naïve Bayes. The model trained for predicting diseases gives a very good result for the given dataset. The algorithms used are very efficient and accurate. This model can be integrated with chatbot models such as DialogFlow from Google to add some more interactions.

[15]Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2020). HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System. ***IEEE Access*, *8*, 133034-133050.**

The system proposed system here uses Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to detect and eliminate the outliers, a hybrid Synthetic Minority Over-sampling Technique-Edited Nearest Neighbour (SMOTE-ENN) to balance the training data distribution and XGBoost to predict heart disease. On comparison with other well-known machine learning algorithms, the proposed system outperformed them on the grounds of accuracy, precision and speed..

[16]Patil, P. B., Shastry, P. M., & Ashok Kumar, P. S. (2020). MACHINE LEARNING BASED ALGORITHM FOR RISK PREDICTION OF CARDIOVASCULAR DISEASE (CVD). ***Journal of Critical Reviews*, *7*(9), 836-844.**

The proposed system uses Naive' Bayesian classifier, decision tree and k- nearest neighbour algorithm. After the prediction if the accuracy value is above 95%, the disease condition is satisfied. Analysis showed that the proposed neural network model fared better than other deep learning techniques

[17]Manogaran, G., Varatharajan, R., & Priyan, M. K. (2018). Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system. *Multimedia tools and applications*, *77*(4), 4379-4399.

This paper put forth a heart disease diagnosis using Multiple Kernel Learning with Adaptive Neuro-Fuzzy Inference System (MKL with ANFIS) based deep learning method. The results from the proposed MKL with ANFIS method has produced high sensitivity (98%), high specificity (99%) and less Mean Square Error (0.01) for the for the KEGG Metabolic Reaction Network dataset.

[18]Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists

Yun Liu, PhD; Timo Kohlberger, PhD; Mohammad Norouzi, PhD; George E. Dahl, PhD; Jenny L. Smith, DO:

The authors have proposed a work to evaluate the application and clinical implementation of a state-of-the-art deep learning–based artificial intelligence algorithm (LYmph Node Assistant or LYNA) for detection of metastatic breast cancer in sentinel lymph node biopsies.The main limitation in this proposed method is that LYNA lacks context about the anatomic position of the current field of view and will be unable to automatically make position-dependent determinations such as extranodal extension and lymph-vascular invasion.LYNA achieved a slide-level area under the receiver operating characteristic (AUC) of 99% and a tumor-level sensitivity of 91% at 1 false positive per patient on the Camelyon16 evaluation dataset.

[19]Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System Alejandro Rodríguez-Ruiz, Elizabeth Krupinski, Jan-Jurre Mordang, Kathy Schilling, Sylvia H. Heywang-Köbrunner, Ioannis Sechopoulos, Ritse M. Mann

The main of this research is to compare breast cancer detection performance of radiologists reading mammographic examinations unaided versus supported by an artificial intelligence (AI) system.The AI system is trained, validated, and tested by using a database containing more than 9000 mammograms with cancer.Radiologists improved their detection performance when using AI support, with the average AUC increasing from 0.87 to 0.89 (difference, 0.02; P = .002)On average, the AUC was higher with AI support than with unaided reading (0.89 vs 0.87, respectively; P = .002). Sensitivity increased with AI support (86% [86 of 100] vs 83% [83 of 100]; P = .046).

[20]Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists
Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Mireille Broeders, Gisella Gennaro, Paola Clauser, Thomas H Helbich,JNCI: Journal of the National Cancer Institute, Volume 111, Issue 9, September 2019.

The main aim of the authors in this paper is to compare the stand-alone performance of an AI system to that of radiologists in detecting breast cancer in DM.The performance of the AI system was statistically non inferior to that of the average of the 101 radiologists. The AI system had a 0.840 (95% confidence interval [CI] = 0.820 to 0.860) area under the ROC curve and the average of the radiologists was 0.814 (95% CI = 0.787 to 0.841) (difference 95% CI = −0.003 to 0.055). The AI system had an AUC higher than 61.4% of the radiologists.As a result, the evaluated AI system achieved a cancer detection accuracy comparable to an average breast radiologist in this retrospective setting.

CASE STUDY:

Objectives:

Perceptions on the use of chatbots in healthcare

We are aiming at achieving a better expert system using WebApp as Chatbots not always gives a better accurate disease prediction from users end. This study cites the perceptions regarding the effectiveness of chatbots in the medical sector from Dr. K.Duraisamy , Cardiologist — Universal Hospital (Erode, Tamil Nadu) along with exploring various chatbots and its uses by us stating the limitations of using chatbots.

Problems and Outcomes :

- **When I (Harida) contacted Dr.K.Duraisamy over phone to get "His perception regarding the use of chatbot in the medical sector as a physician and if he encourages the use of them which almost replaces the doctors to some extent**" he added his views upon the use of chatbots from his experience that they are very primary level of detecting the disease and that they are not still widely used in a larger part of Tamil Nadu and yet as a physician he states that patients lack the empathy as they cannot connect easily with the chatbots in terms of emotions and they always not directs to the accurate prediction which might leave the patients in a huge dilemma regarding the disease if its a major one and mentioned that as a physician they do encourage the use of chatbots as patients need not always visit a hospital for very small medical reasons as this atmosphere influence patients mental state by creating an illusion regarding their disease if they are very serious. But he has also added that since majority of people have less or no knowledge related to the disease or if its something related to women's in terms of pregnancy or maternity related issues chatbots can no where connect them emotionally and can never help them as they can't really protect the personal information so patients are hesitant to share their information.

- There are a lot of false predictive values in the Chatbot which concerns the majority of the patients or users from using them.
- When we were researching by using various chatbots online we could find that at times the chatbot is not very responsive and as mentioned by the doctor it doesn’t predict accurate disease by giving a common symptom as in case of fever when we enter the symptom as headache and cold it couldn’t differentiate if we have got cold or fever as the symptom is same for both.

----> **ADA:** It is a wonderful app which asks users to register when they start to use it, ask if we are 16 or above and if we agree to provide our responses for research. And then, asks for basic

information like their name, gender, DOB and questions like if they are pregnant (in case we chose gender as female), if they have diabetes, high blood pressure. AFer this, they provide the op?on to start symptom analysis. They ask if we are asking for ourselves or for another person.

- The app then goes on to ask what is the most prominent symptom. AFer entering that, we are asked ques?ons about the symptoms to which we could choose between the op?ons: yes, no, I don’t know. An option of providing feedback during each question is there, and for some questions, a picture is added to make the user know what they mean. Adding to it, we can go back to change the option then we choose one by one, backwards.

- The number of questions vary. The average amount of time taken to finish answering these questions is 5-7 minutes.

- AFer the questions, we can view the report which is made by the app with a foreword telling that the report includes only possible illness and possible cause for symptoms and that it is not a diagnosis. When I (Shruti) tried entering the main symptom as headache, I got questions varying and I answered. The report that was generated in the end showed that I might have acute mountain sickness, along with a paragraph of how people would get this and how we can avoid getting sick during hiking. Further, it shows likely causes and symptoms which were not listed as the main ones for this illness. Although I answered the questions and kept in mind about sinusias the illness in mind, the possible illness was shown as acute mountain sickness.

- The user interface for this app is very pleasing and very efficient. Even though the possibility was away from the illness, the performance of this app along with all the information was collected over the past years and has made it learn more.

- Similarly when examining **MediCareBot** is another such app, whose name indicates as an app similar to the one above. But, in the starting, while launching the application for the first me, the login page is displayed. No create account or sign in op?on is visible. Thus going further into the app itself is not possible.

- Also majority of the exis?ng chatbots doesn't have doctor's recommenda?ons in them which is a major drawback as users couldn't connect for further in depth prediction for other major diseases.

Recommended Solution:

Dr. Duraisamy feels that use of chatbots should be improved by adding a physicians record along with the patients so that patients can feel the virtual connect with the doctor. He also says that there should be more user friendly framework which is feasible to every patients even if they are illiterate as to improve their efficiency. From our research after studying various research works and thesis we propose the use of online disease predictor system along with analysis using a webApp as in this era almost everyone uses mobile phone with internet connection. So they can be efficiently used even with basic knowledge and this will improve the accuracy of the disease prediction as well by connecting the user to their respective doctors for further advanced analysis of disease.

Result:

So from analysing from the above problem along with our own perceptions we have come up with a WebApp interface which for now aims at predicting Diabetics, Cardio and breast cancer related disease along with providing doctors recommendation to the website so that once a person enters the system and predicts the disease they will be redirected to the list of physicians where they can choose the doctors according to their convenience from any part of India which would be more encouragable than the chatbots.

Reference:

1) Dr. K.Duraisamy , Cardiologist — Universal Hospital (Erode, Tamil Nadu)

2) Chatbots and various portals referred to in the literature survey [page 8-14].

PROPOSED METHODOLOGY

- **DISEASE DETECTION CHATBOT:**

First the data in the dataset is tokenized to word using NLTK module.

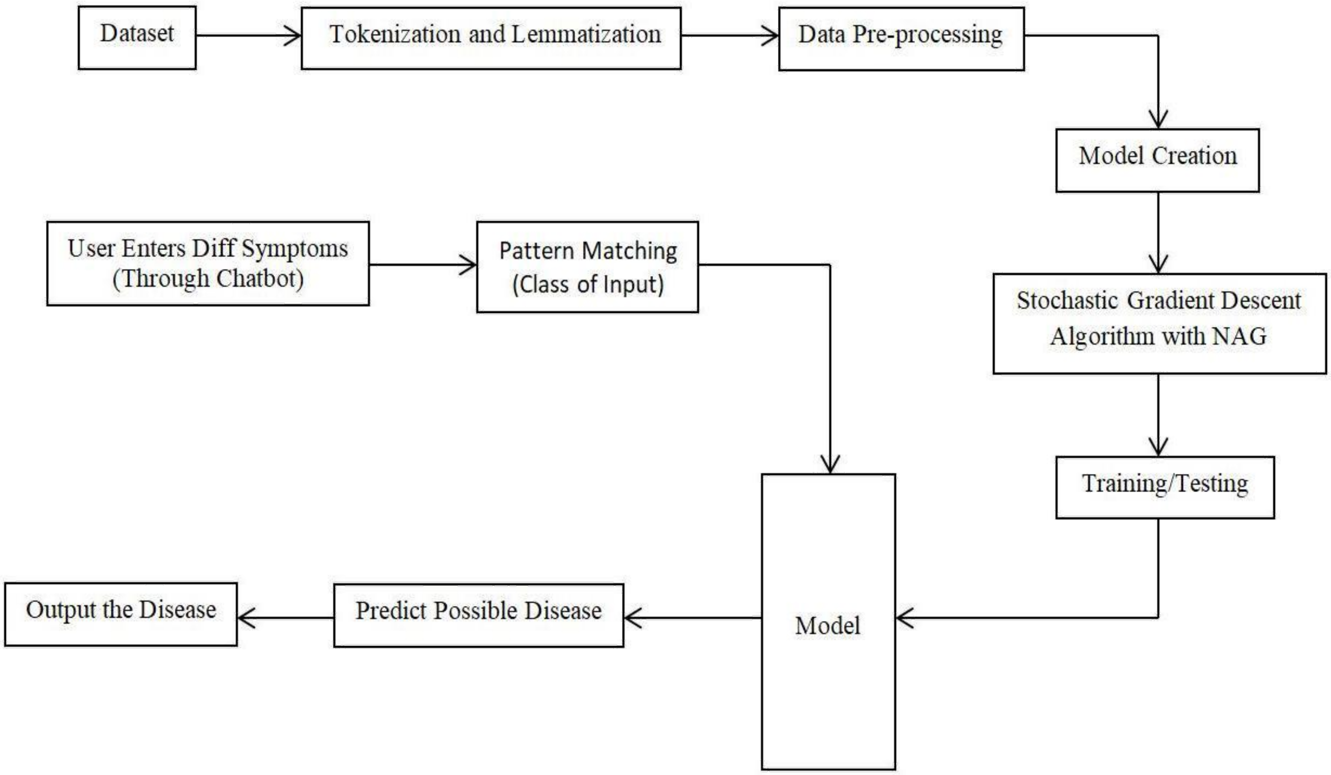

Dataset
Tokenization and Lemmatization
Data Pre-processing
Model Creation
User Enters Diff Symptoms (Through Chatbot)
Pattern Matching (Class of Input)
Stochastic Gradient Descent Algorithm with NAG
Training/Testing
Model
Output the Disease
Predict Possible Disease

Figure 1 Flow of the Chatbot

Each word is appended to the list of words. After this each word is lemmatized and duplicate words are removed from the list of words. These words are then stored using a pickle file which will be used to predict the data. The input of the training data uses the pattern and the output is the class to which the input pattern belongs to. The data is converted into numbers to be fed to the neural network

The model contains 3 layers. First layers contain 128 neurons followed by the second layer containing 64 neurons only. The 3rd layer, that is the output layer, has a number of neurons equal to the number of intents to predict output intent using softmax. After the model is compiled, Stochastic Gradient Descent Algorithm is used with Nesterov Accelerated Gradient to train the data model. Then the model is fitted and saved.

To display the predicted output, the trained model is loaded and a graphical user interface is used. The model just predicts which class the input belongs to. This class is used to retrieve a random response from the list of responses with the help of various functions like bag of words. Tkinter library is used to display the predicted output on the GUI.

PREDICTION ANALYSIS:

- Perform descriptive analysis on heart disease prediction, breast cancer prediction and diabetes prediction using key factors like Glucose levels, Blood Pressure, Skin Thickness, BMI etc.

- Visually explore these variables, you may need to look for the distribution of these variables using histograms. Treat the missing values accordingly.

- We observe integers as well as float data-type of variables in this dataset. Create a count (frequency) plot describing the data types and the count of variables.

- Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of actions.

- Create scatter charts between the pair of variables to understand the relationships. Describe your findings.

- Perform correlation analysis. Visually explore it using a heat map.

- Devise strategies for model building. It is important to decide the right validation framework. Express your thought process. Would Cross validation be useful in this scenario?

- Apply an appropriate classification algorithm to build a model. Compare various models with the results from KNN.

- Create a classification report by analyzing sensitivity, specificity, AUC (ROC curve) etc. Please try to be as descriptive as possible to explain what values of these parameters you settled for? any why?

- Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:

  a) Pie chart to describe the diabetic/non-diabetic population
  b) Scatter charts between relevant variables to analyze the relationships
  c) Histogram/frequency charts to analyse the distribution of the data
  d) Heatmap of correlation analysis among the relevant variables
  e) Create bins of Age values and analyse different variables for these age brackets using a bubble chart.

- DIABETES PREDICTION ANALYSIS:

□ There are 768 rows and 9 columns in this dataset

□ The dataset has nine attributes in which there are eight independent variables (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age) and one dependent variable (Outcome).

□ BMI and DiabetesPedigreeFunction are a float data type and rest of the variables are integer data type.

- The Variables have a lot of zero values which can be represented as missing values

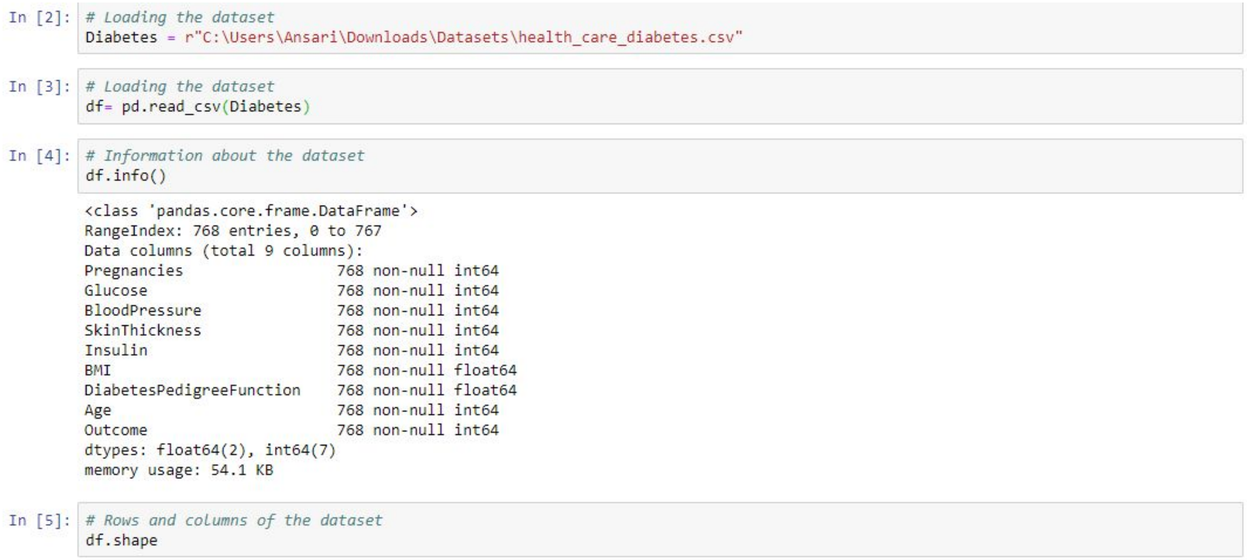
- The missing values '0' is replaced by mean to explore the dataset.

- The Outcome variable shows that there are 500 non-diabetic people and 268 diabetic people.
- It means that 65.1% people are diabetic and 34.9% people are non-diabetic.

- The parameters Glucose, Blood Pressure, BMI are normally distributed.
- Pregnancies, Insulin, Age, DiabetesPedigreeFunction are rightly skewed.

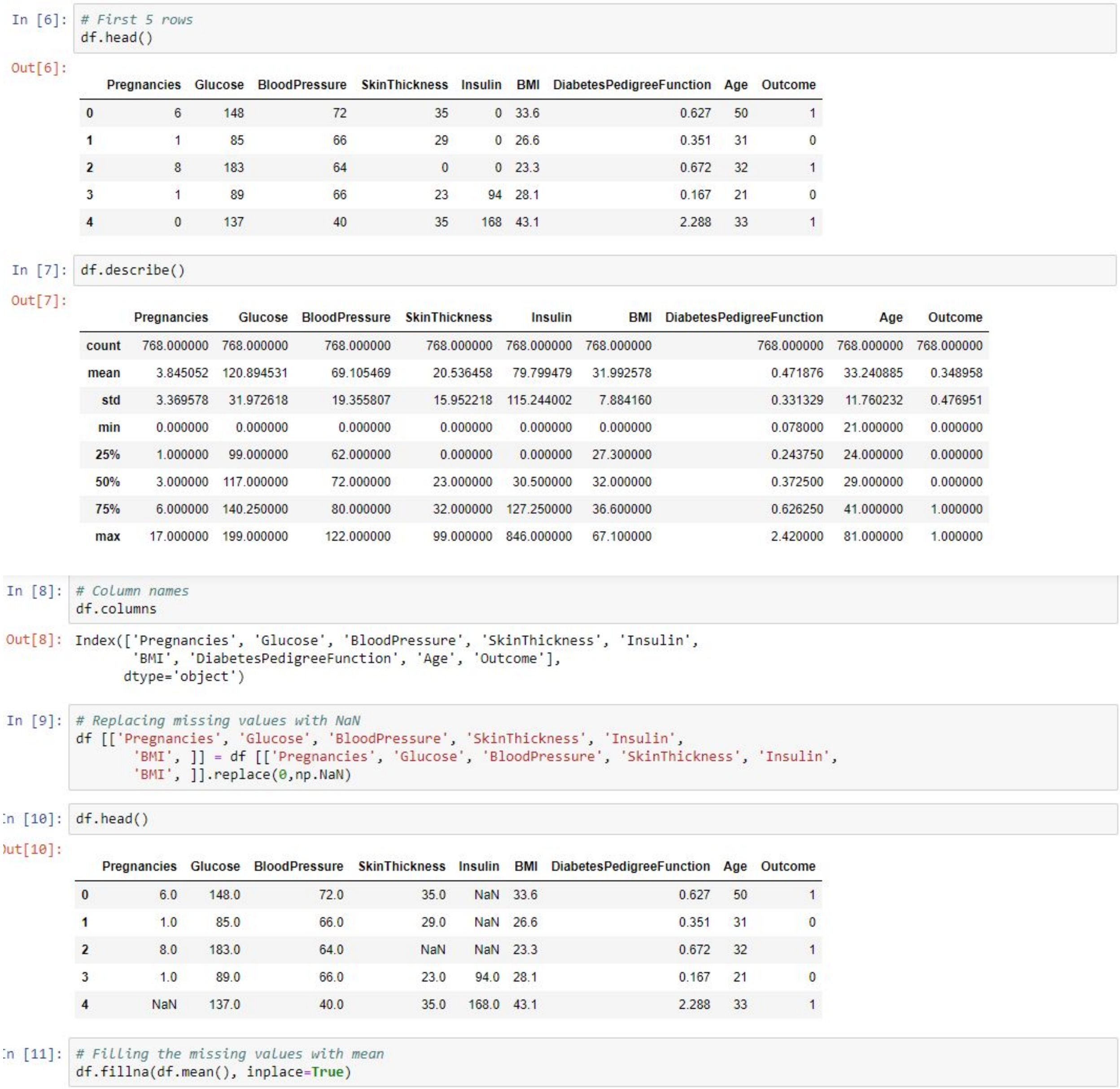
- Blood Pressure, Skin Thickness, Insulin, BMI have outliers.

The code for the results is here –



```
In [2]: # Loading the dataset
        Diabetes = r"C:\Users\Ansari\Downloads\Datasets\health_care_diabetes.csv"

In [3]: # Loading the dataset
        df= pd.read_csv(Diabetes)

In [4]: # Information about the dataset
        df.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 768 entries, 0 to 767
        Data columns (total 9 columns):
        Pregnancies                 768 non-null int64
        Glucose                     768 non-null int64
        BloodPressure               768 non-null int64
        SkinThickness               768 non-null int64
        Insulin                     768 non-null int64
        BMI                         768 non-null float64
        DiabetesPedigreeFunction    768 non-null float64
        Age                         768 non-null int64
        Outcome                     768 non-null int64
        dtypes: float64(2), int64(7)
        memory usage: 54.1 KB

In [5]: # Rows and columns of the dataset
        df.shape
```

Figure 2 Code snippet for diabetes prediction with its attributes



```
In [6]: # First 5 rows
        df.head()
```

Out[6]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

```
In [7]: df.describe()
```

Out[7]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

```
In [8]: # Column names
        df.columns

Out[8]: Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
               'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
              dtype='object')
```

```
In [9]: # Replacing missing values with NaN
        df [['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
               'BMI', ]] = df [['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
               'BMI', ]].replace(0,np.NaN)
```

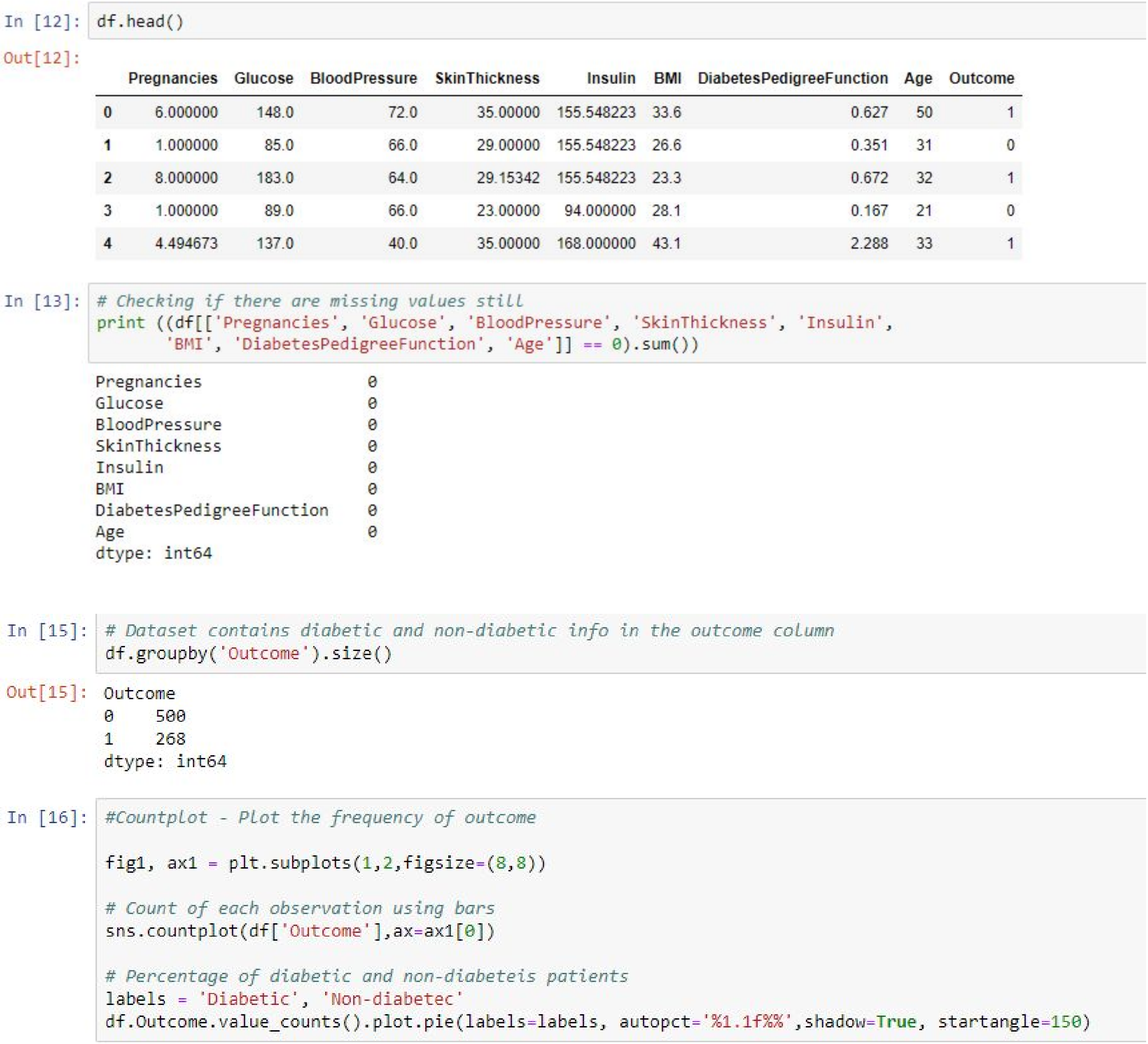
```
In [10]: df.head()
```

Out[10]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6.0	148.0	72.0	35.0	NaN	33.6	0.627	50	1
1	1.0	85.0	66.0	29.0	NaN	26.6	0.351	31	0
2	8.0	183.0	64.0	NaN	NaN	23.3	0.672	32	1
3	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	NaN	137.0	40.0	35.0	168.0	43.1	2.288	33	1

```
In [11]: # Filling the missing values with mean
         df.fillna(df.mean(), inplace=True)
```

Figure 3 Diabetes prediction dataset and code



```
In [12]: df.head()
```

Out[12]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6.000000	148.0	72.0	35.00000	155.548223	33.6	0.627	50	1
1	1.000000	85.0	66.0	29.00000	155.548223	26.6	0.351	31	0
2	8.000000	183.0	64.0	29.15342	155.548223	23.3	0.672	32	1
3	1.000000	89.0	66.0	23.00000	94.000000	28.1	0.167	21	0
4	4.494673	137.0	40.0	35.00000	168.000000	43.1	2.288	33	1

```
In [13]: # Checking if there are missing values still
         print ((df[['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
                'BMI', 'DiabetesPedigreeFunction', 'Age']] == 0).sum())

         Pregnancies                 0
         Glucose                     0
         BloodPressure               0
         SkinThickness               0
         Insulin                     0
         BMI                         0
         DiabetesPedigreeFunction    0
         Age                         0
         dtype: int64
```

```
In [15]: # Dataset contains diabetic and non-diabetic info in the outcome column
         df.groupby('Outcome').size()

Out[15]: Outcome
         0    500
         1    268
         dtype: int64
```

```
In [16]: #Countplot - Plot the frequency of outcome

         fig1, ax1 = plt.subplots(1,2,figsize=(8,8))

         # Count of each observation using bars
         sns.countplot(df['Outcome'],ax=ax1[0])

         # Percentage of diabetic and non-diabeteis patients
         labels = 'Diabetic', 'Non-diabetec'
         df.Outcome.value_counts().plot.pie(labels=labels, autopct='%1.1f%%',shadow=True, startangle=150)
```

Figure 4 Grouping outcome code snippet for diabetes prediction

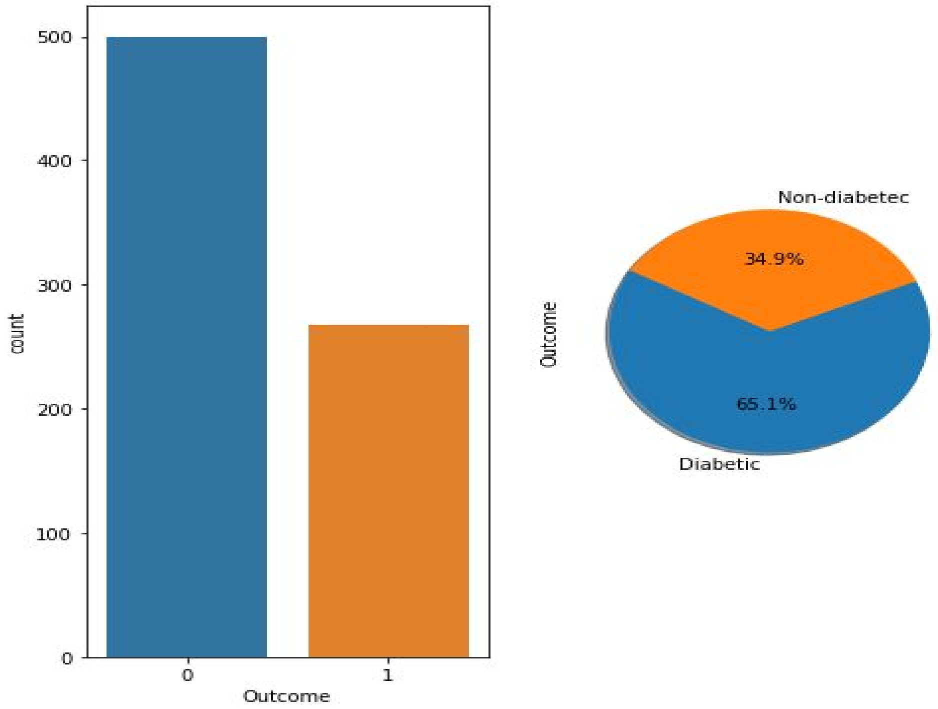
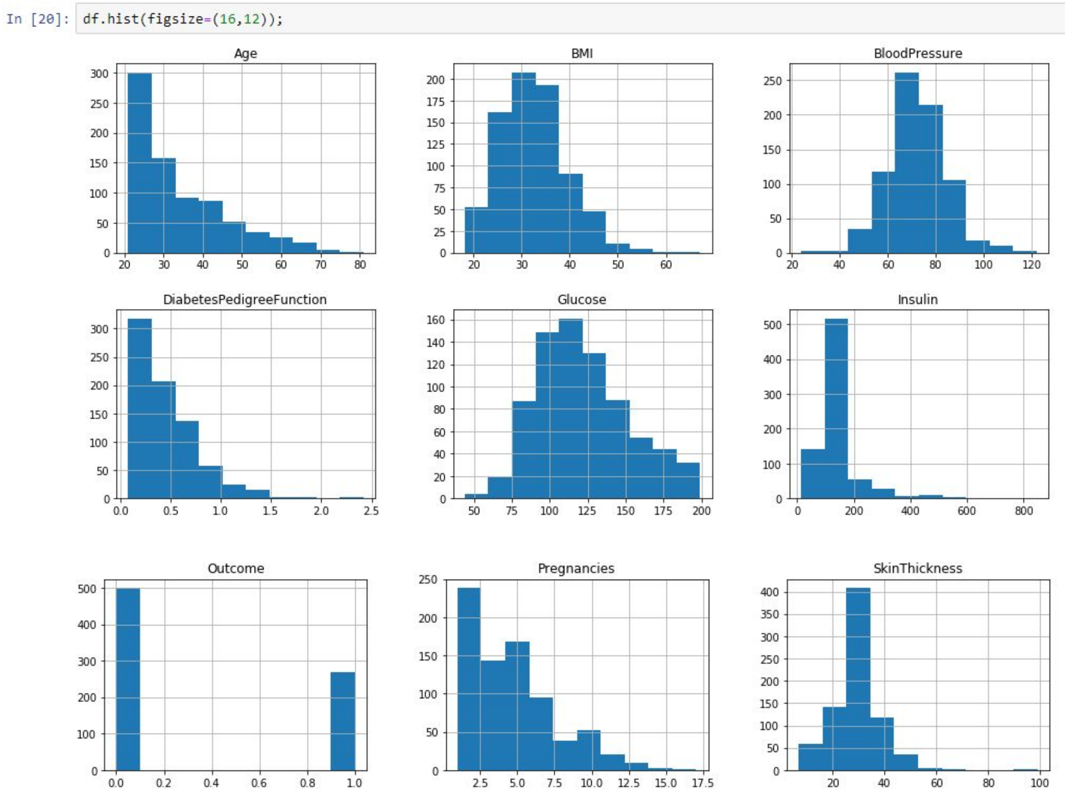
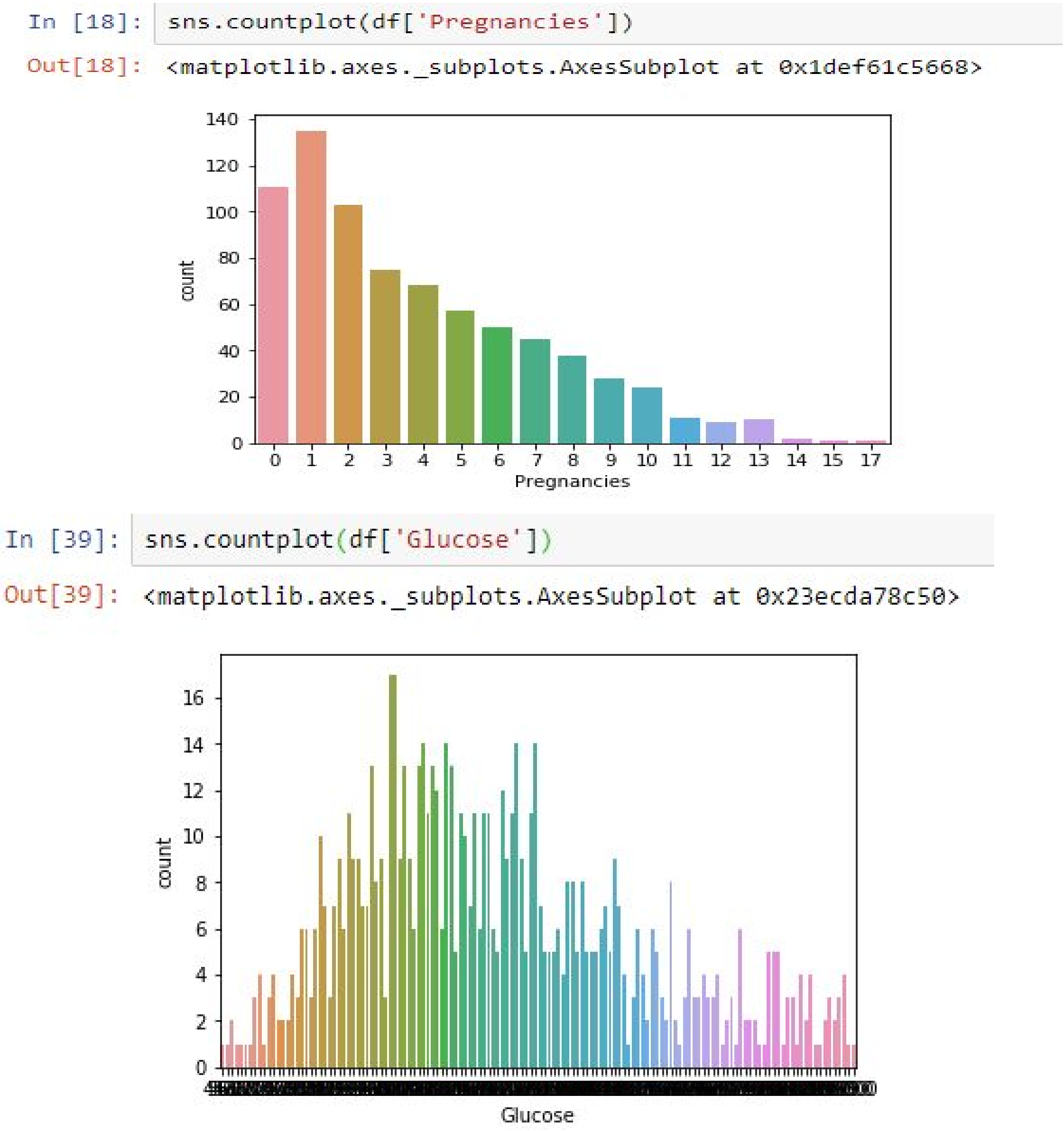


Figure 5 Outcome of grouping in diabetes prediction into diabetic and non-diabetic



```
In [20]: df.hist(figsize=(16,12));
```

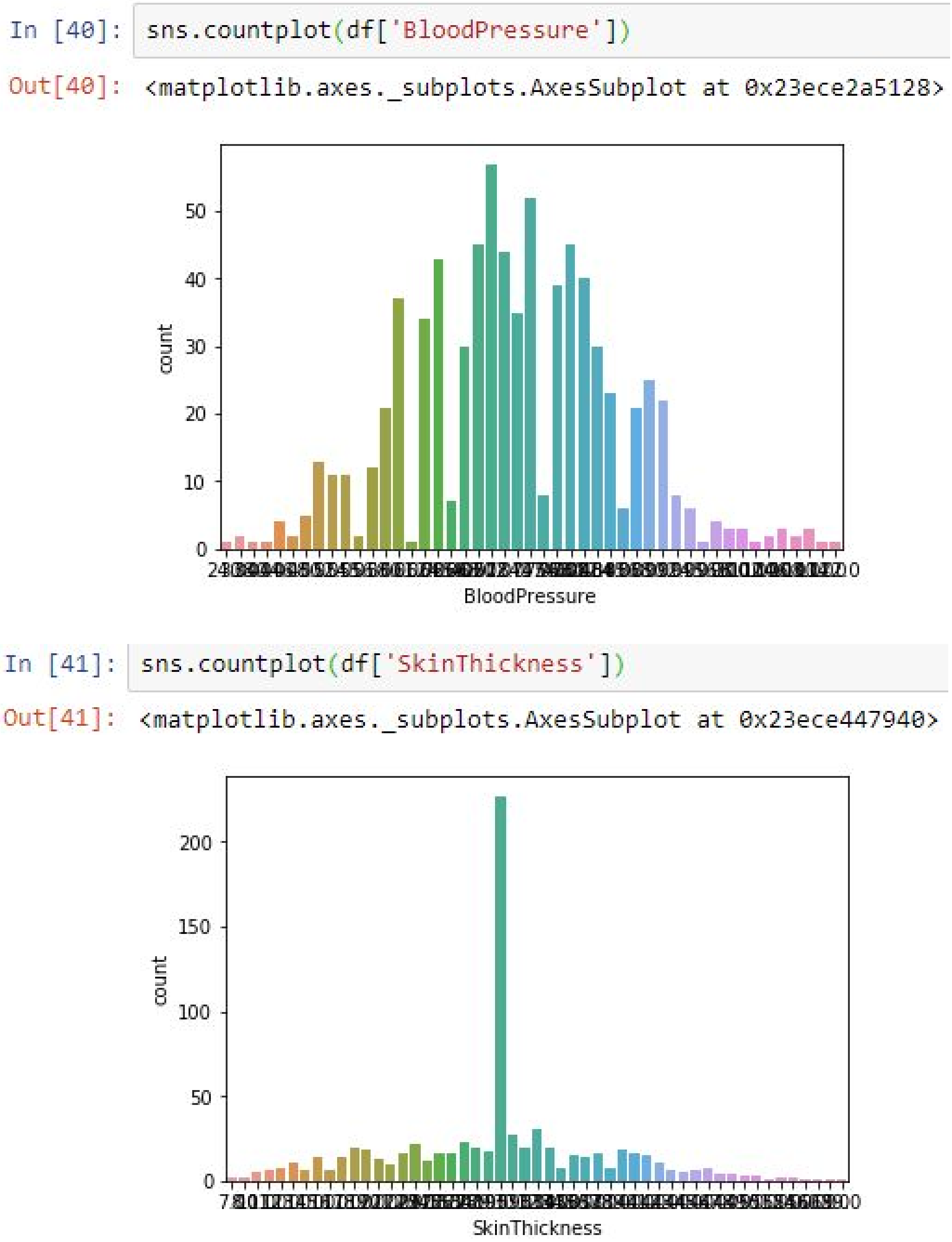
Figure 6 Plots for each attribute for diabetes prediction



```
In [18]: sns.countplot(df['Pregnancies'])
Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x1def61c5668>
```

```
In [39]: sns.countplot(df['Glucose'])
Out[39]: <matplotlib.axes._subplots.AxesSubplot at 0x23ecda78c50>
```

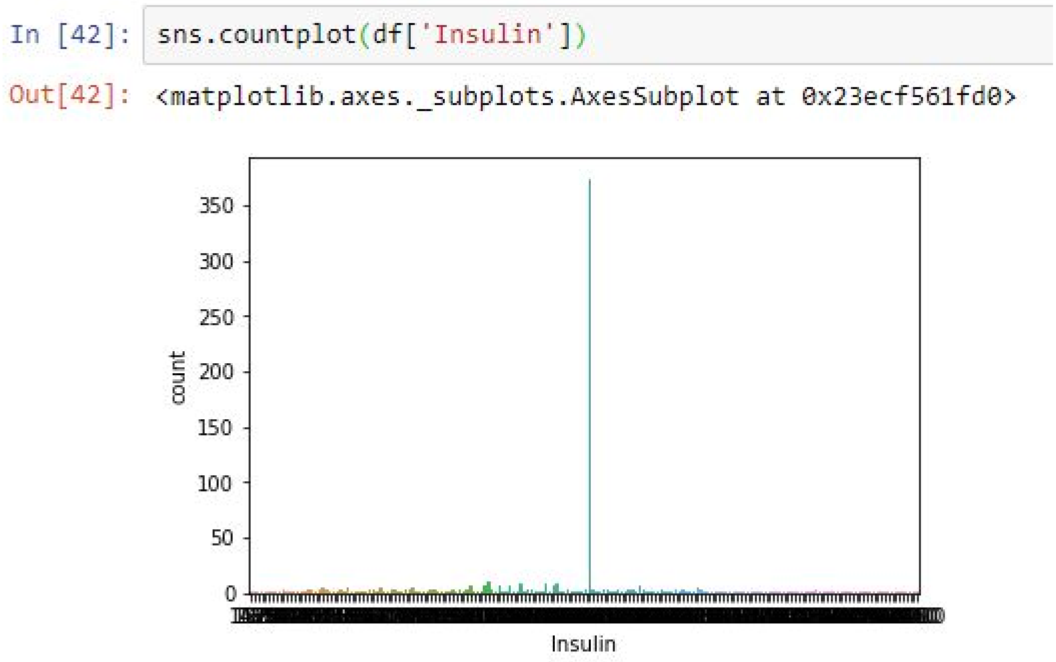
Figure 7 Plots for Pregnancies against count and Glucose against count in diabetes prediction



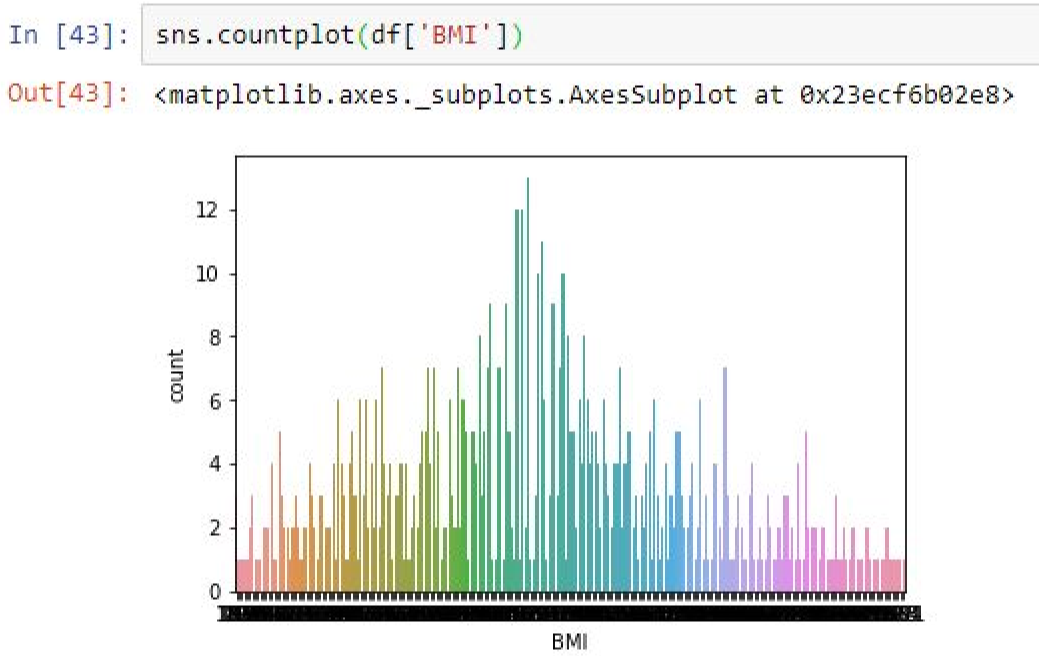
```
In [40]: sns.countplot(df['BloodPressure'])

Out[40]: <matplotlib.axes._subplots.AxesSubplot at 0x23ece2a5128>
```

```
In [41]: sns.countplot(df['SkinThickness'])

Out[41]: <matplotlib.axes._subplots.AxesSubplot at 0x23ece447940>
```

Figure 8 Plots for Blood pressure against count and Skin Thickness against count in diabetes prediction

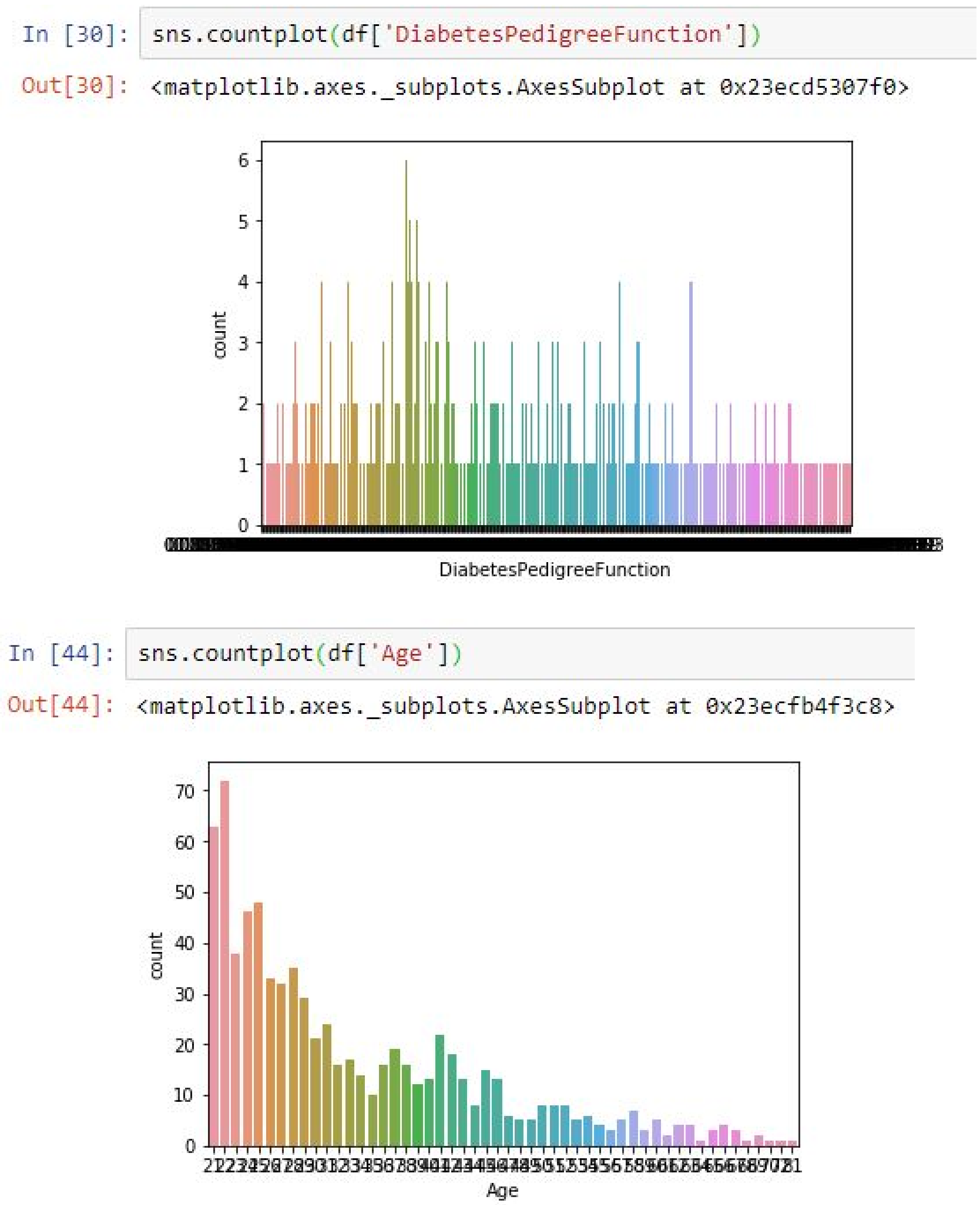


```
In [42]: sns.countplot(df['Insulin'])

Out[42]: <matplotlib.axes._subplots.AxesSubplot at 0x23ecf561fd0>
```



```
In [43]: sns.countplot(df['BMI'])

Out[43]: <matplotlib.axes._subplots.AxesSubplot at 0x23ecf6b02e8>
```

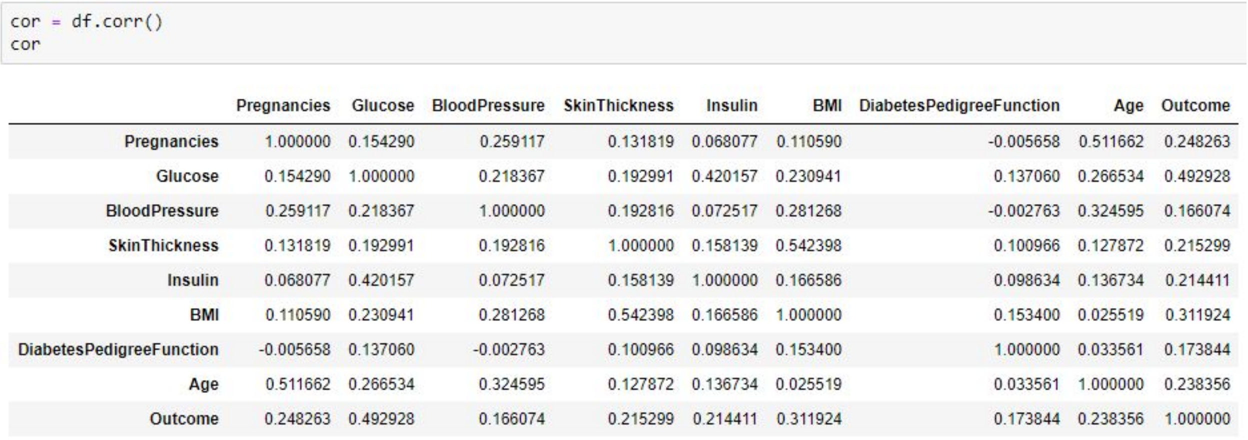
Figure 9 Plots for Insulin against count and BMI against count in diabetes prediction



```
In [30]: sns.countplot(df['DiabetesPedigreeFunction'])

Out[30]: <matplotlib.axes._subplots.AxesSubplot at 0x23ecd5307f0>
```

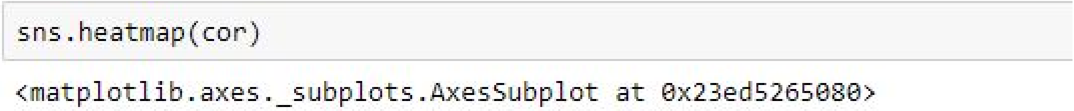
```
In [44]: sns.countplot(df['Age'])

Out[44]: <matplotlib.axes._subplots.AxesSubplot at 0x23ecfb4f3c8>
```

Figure 10 Plots for Diabetes Pedigree function against count and Age against count in diabetes prediction



```
cor = df.corr()
cor
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.154290	0.259117	0.131819	0.068077	0.110590	-0.005658	0.511662	0.248263
Glucose	0.154290	1.000000	0.218367	0.192991	0.420157	0.230941	0.137060	0.266534	0.492928
BloodPressure	0.259117	0.218367	1.000000	0.192816	0.072517	0.281268	-0.002763	0.324595	0.166074
SkinThickness	0.131819	0.192991	0.192816	1.000000	0.158139	0.542398	0.100966	0.127872	0.215299
Insulin	0.068077	0.420157	0.072517	0.158139	1.000000	0.166586	0.098634	0.136734	0.214411
BMI	0.110590	0.230941	0.281268	0.542398	0.166586	1.000000	0.153400	0.025519	0.311924
DiabetesPedigreeFunction	-0.005658	0.137060	-0.002763	0.100966	0.098634	0.153400	1.000000	0.033561	0.173844
Age	0.511662	0.266534	0.324595	0.127872	0.136734	0.025519	0.033561	1.000000	0.238356
Outcome	0.248263	0.492928	0.166074	0.215299	0.214411	0.311924	0.173844	0.238356	1.000000



```
sns.heatmap(cor)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x23ed5265080>
```

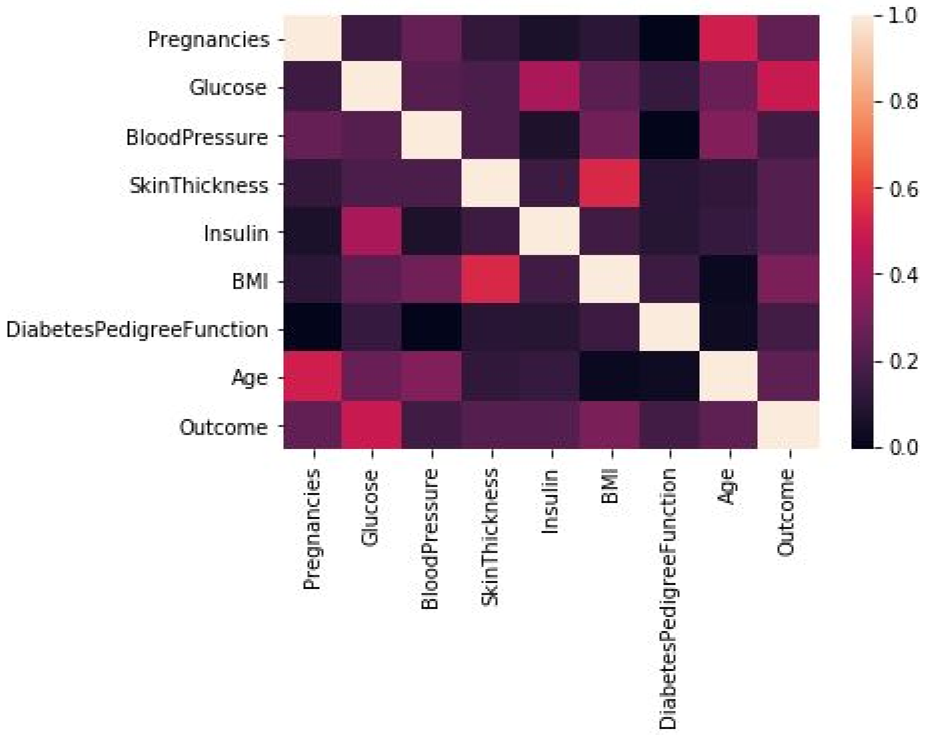


Figure 11 Diabetes prediction attributes against each other and its Heatmap plot

The correlation plot shows the relation between the parameters.

Glucose, Age, BMI and Pregnancies are the most correlated parameters with the Outcome

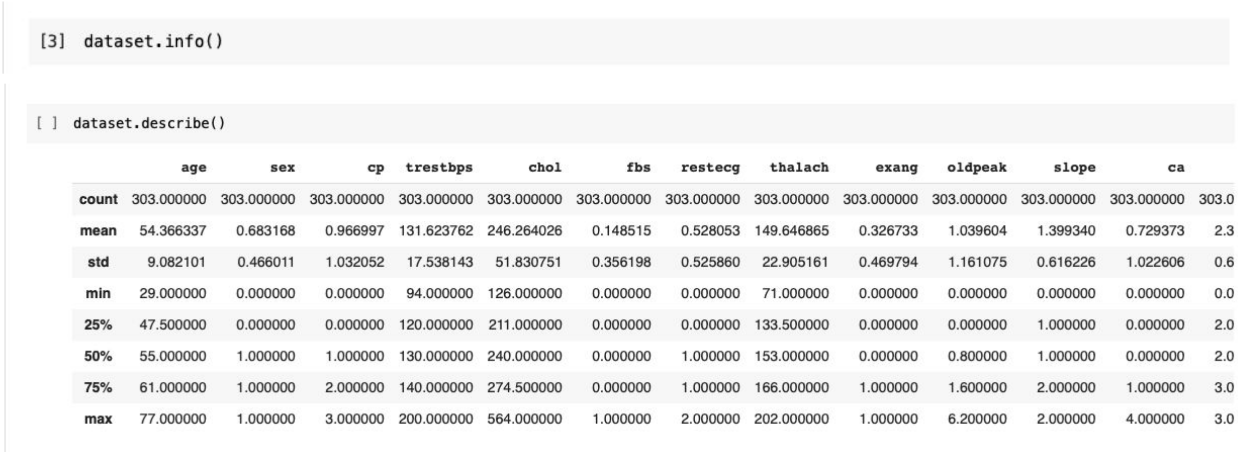
Insulin and DiabetesPedigreeFunction have little correlation with the outcome.

Blood Pressure and Skin Thickness have tiny correlation with the outcome.

There is a little correlation between Age and Pregnancies, Insulin and Skin Thickness, BMI and Skin Thickness, Insulin and Glucose.

HEART DISEASE PREDICTION ANALYSIS:

- There are 768 rows and 9 columns in this dataset
- The dataset has various attributes like Sex, age, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope.
- While the maximum for age reaches 77, the maximum of chol (serum cholesterol) is 564.

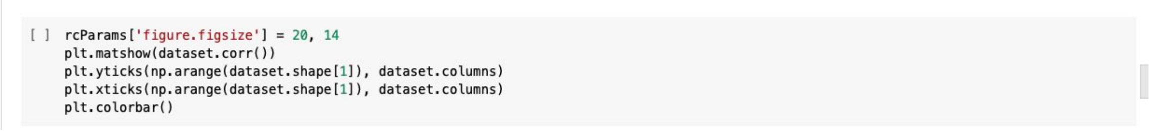


```
[3]  dataset.info()
```

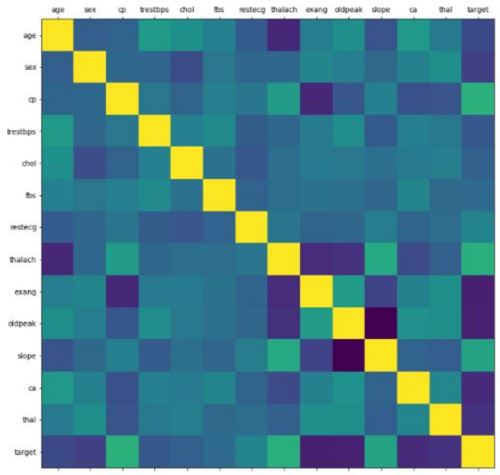
```
[ ]  dataset.describe()
```

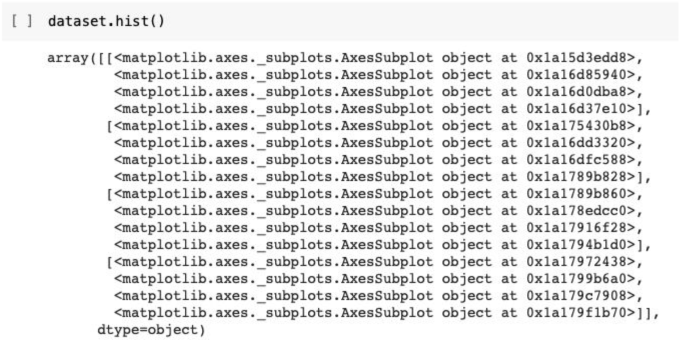
```
              age        sex         cp   trestbps       chol        fbs    restecg    thalach      exang    oldpeak      slope         ca
count  303.000000 303.000000 303.000000 303.000000 303.000000 303.000000 303.000000 303.000000 303.000000 303.000000 303.000000 303.000000 303.0
mean    54.366337   0.683168   0.966997 131.623762 246.264026   0.148515   0.528053 149.646865   0.326733   1.039604   1.399340   0.729373   2.3
std      9.082101   0.466011   1.032052  17.538143  51.830751   0.356198   0.525860  22.905161   0.469794   1.161075   0.616226   1.022606   0.6
min     29.000000   0.000000   0.000000  94.000000 126.000000   0.000000   0.000000  71.000000   0.000000   0.000000   0.000000   0.000000   0.0
25%     47.500000   0.000000   0.000000 120.000000 211.000000   0.000000   0.000000 133.500000   0.000000   0.000000   1.000000   0.000000   2.0
50%     55.000000   1.000000   1.000000 130.000000 240.000000   0.000000   1.000000 153.000000   0.000000   0.800000   1.000000   0.000000   2.0
75%     61.000000   1.000000   2.000000 140.000000 274.500000   0.000000   1.000000 166.000000   1.000000   1.600000   2.000000   1.000000   3.0
max     77.000000   1.000000   3.000000 200.000000 564.000000   1.000000   2.000000 202.000000   1.000000   6.200000   2.000000   4.000000   3.0
```

Figure 12 Heart disease prediction dataset (first 8 lines)



```
[ ]  rcParams['figure.figsize'] = 20, 14
     plt.matshow(dataset.corr())
     plt.yticks(np.arange(dataset.shape[1]), dataset.columns)
     plt.xticks(np.arange(dataset.shape[1]), dataset.columns)
     plt.colorbar()
```





```
[ ]  dataset.hist()

     array([[<matplotlib.axes._subplots.AxesSubplot object at 0x1a15d3edd8>,
             <matplotlib.axes._subplots.AxesSubplot object at 0x1a16d85940>,
             <matplotlib.axes._subplots.AxesSubplot object at 0x1a16d0dba8>,
             <matplotlib.axes._subplots.AxesSubplot object at 0x1a16d37e10>],
            [<matplotlib.axes._subplots.AxesSubplot object at 0x1a175430b8>,
             <matplotlib.axes._subplots.AxesSubplot object at 0x1a16dd3320>,
             <matplotlib.axes._subplots.AxesSubplot object at 0x1a16dfc588>,
             <matplotlib.axes._subplots.AxesSubplot object at 0x1a1789b828>],
            [<matplotlib.axes._subplots.AxesSubplot object at 0x1a1789b860>,
             <matplotlib.axes._subplots.AxesSubplot object at 0x1a178edcc0>,
             <matplotlib.axes._subplots.AxesSubplot object at 0x1a17916f28>,
             <matplotlib.axes._subplots.AxesSubplot object at 0x1a1794b1d0>],
            [<matplotlib.axes._subplots.AxesSubplot object at 0x1a17972438>,
             <matplotlib.axes._subplots.AxesSubplot object at 0x1a1799b6a0>,
             <matplotlib.axes._subplots.AxesSubplot object at 0x1a179c7908>,
             <matplotlib.axes._subplots.AxesSubplot object at 0x1a179f1b70>]],
           dtype=object)
```

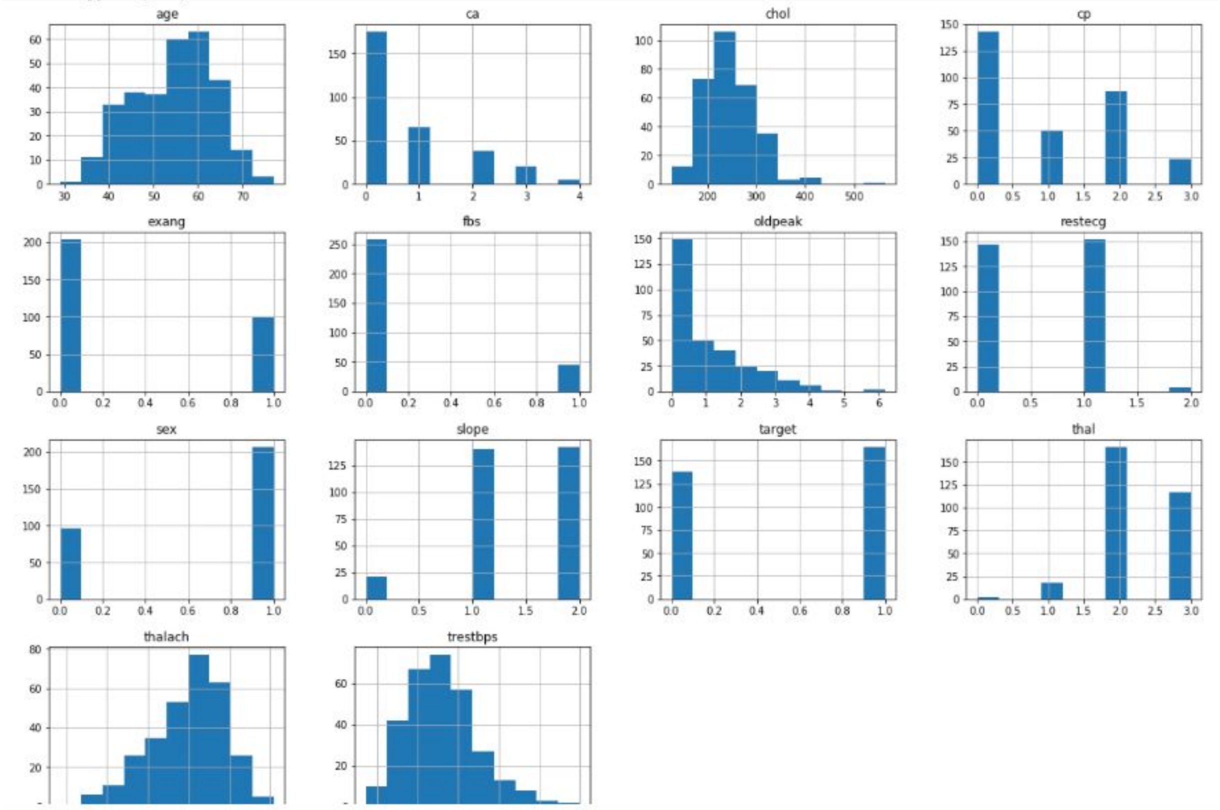
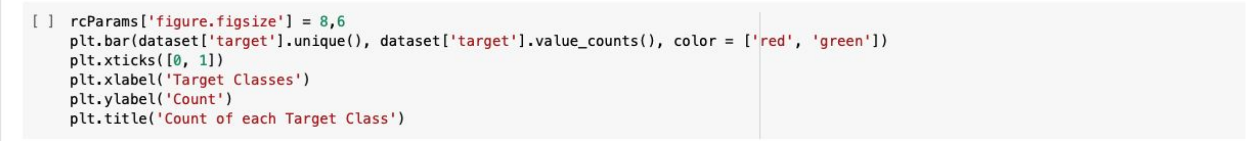


Figure 13 Visualization of the heart disease prediction dataset



```
[ ]  rcParams['figure.figsize'] = 8,6
     plt.bar(dataset['target'].unique(), dataset['target'].value_counts(), color = ['red', 'green'])
     plt.xticks([0, 1])
     plt.xlabel('Target Classes')
     plt.ylabel('Count')
     plt.title('Count of each Target Class')
```

```
Text(0.5, 1.0, 'Count of each Target Class')
```

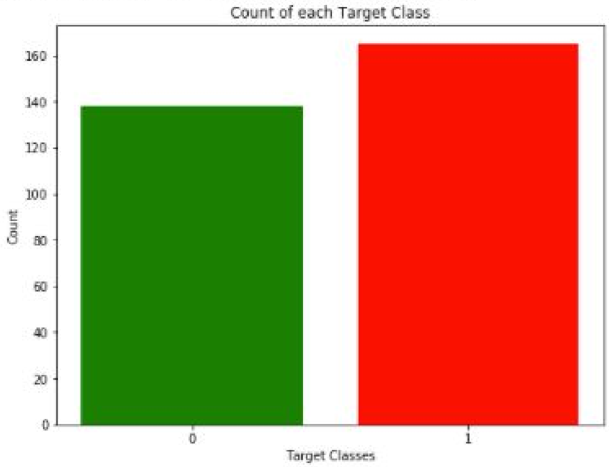
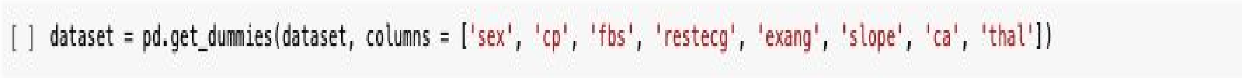
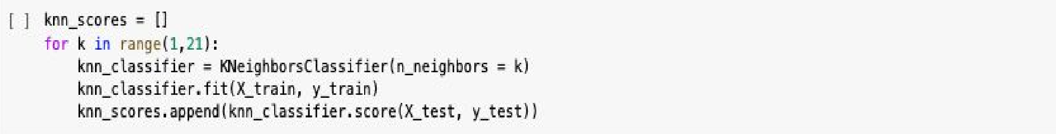


__Figure 14__ Checking for target classes and showing the classes are different.

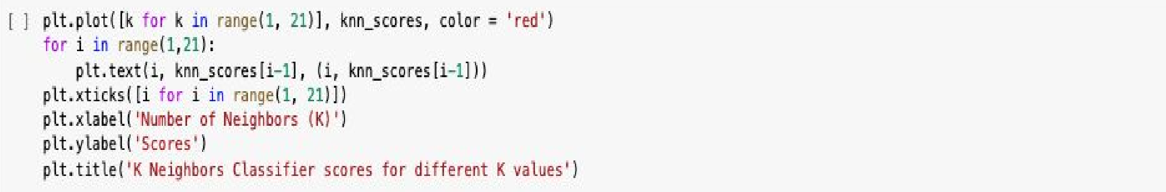


```
[ ]  dataset = pd.get_dummies(dataset, columns = ['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal'])
```

__Figure 15__ Code snippet creating dummy columns for categorical variables



```
[ ] knn_scores = []
    for k in range(1,21):
        knn_classifier = KNeighborsClassifier(n_neighbors = k)
        knn_classifier.fit(X_train, y_train)
        knn_scores.append(knn_classifier.score(X_test, y_test))
```



```
[ ] plt.plot([k for k in range(1, 21)], knn_scores, color = 'red')
    for i in range(1,21):
        plt.text(i, knn_scores[i-1], (i, knn_scores[i-1]))
    plt.xticks([i for i in range(1, 21)])
    plt.xlabel('Number of Neighbors (K)')
    plt.ylabel('Scores')
    plt.title('K Neighbors Classifier scores for different K values')
```

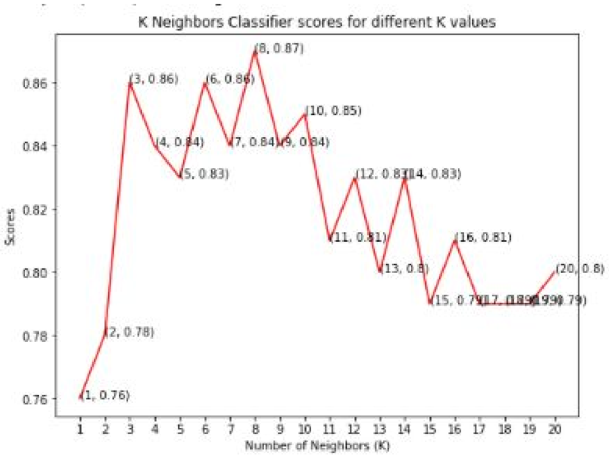


Figure 16 Code snippet for getting K-score to achieve best score and plot showing the maximum score is available at 0.87 for the 8 neighbors

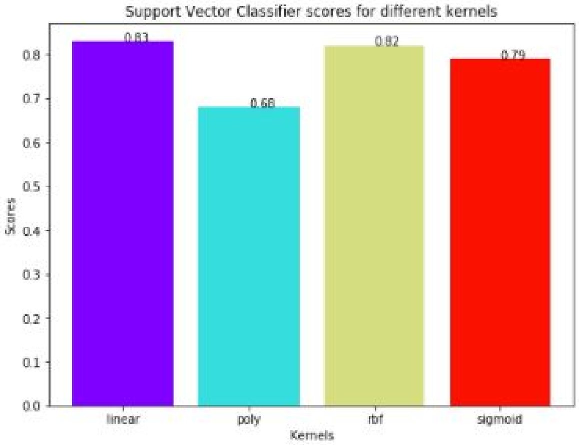
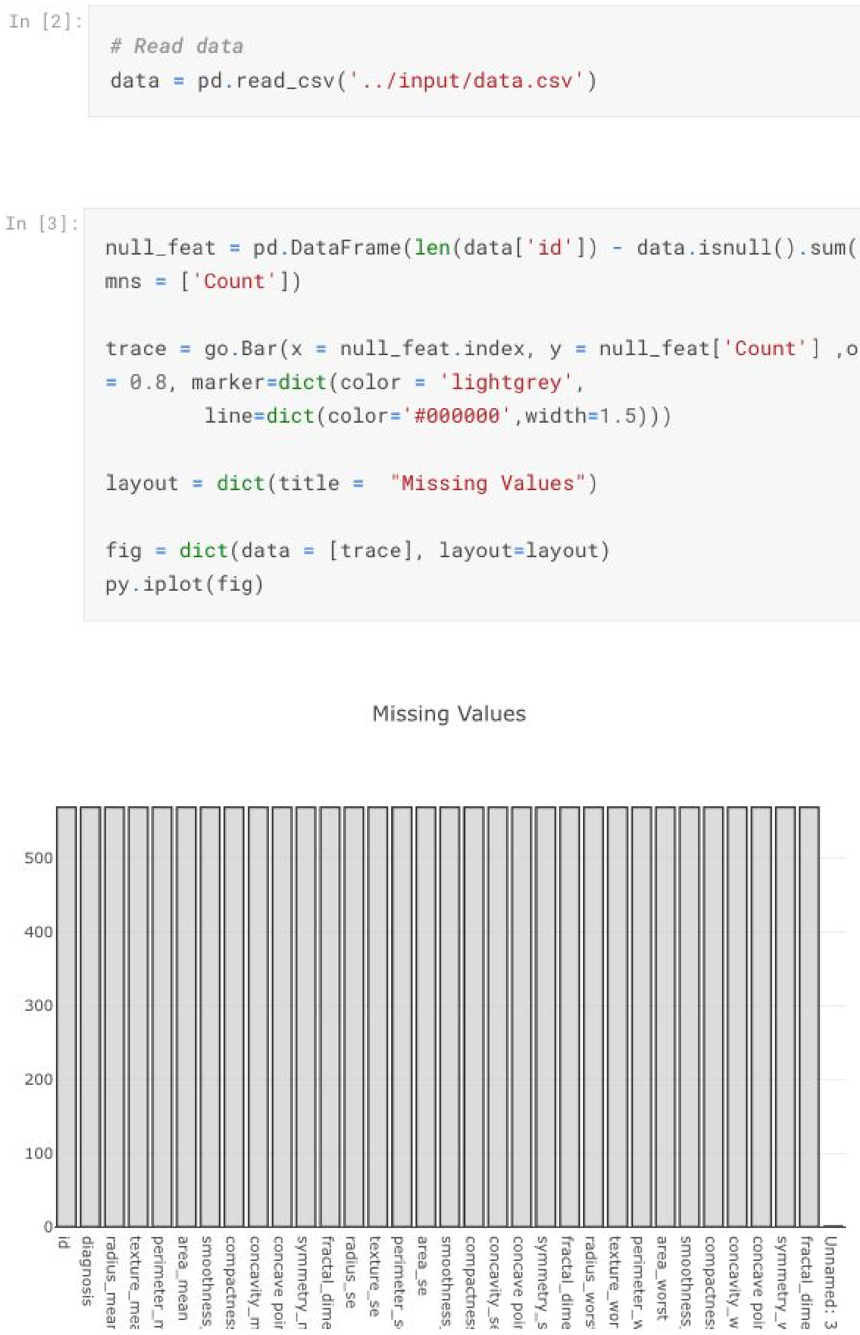


Figure 17 The graph shows the best classified scores based on machine learning algorithms

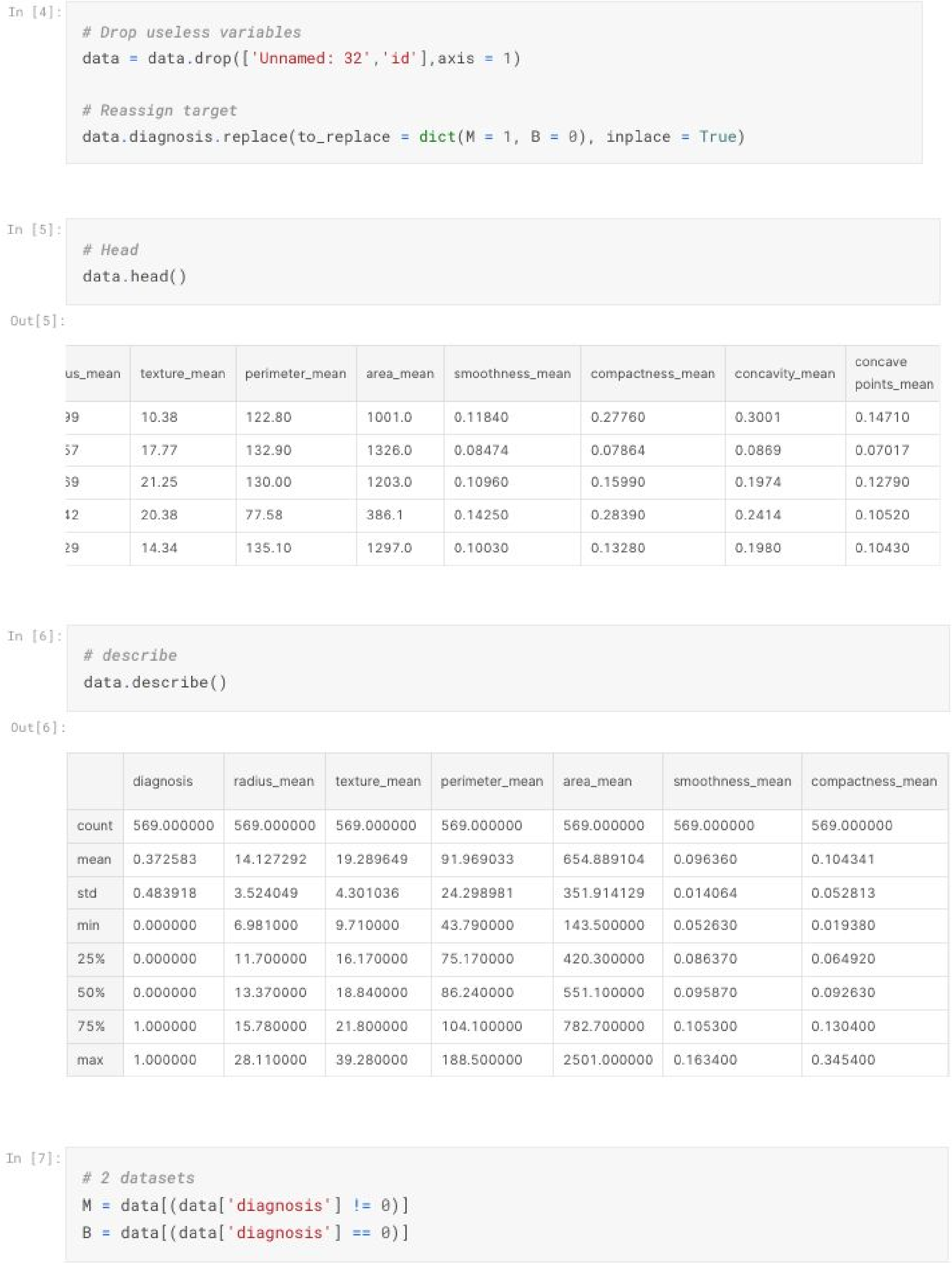
- There are totally 7 different attributes in the dataset : radius_mean, texture_mean ,Perimeter_mean, area_mean, Smoothness_mean, compactness_mean.



```
In [2]:
# Read data
data = pd.read_csv('../input/data.csv')
```

```
In [3]:
null_feat = pd.DataFrame(len(data['id']) - data.isnull().sum(), colu
mns = ['Count'])

trace = go.Bar(x = null_feat.index, y = null_feat['Count'] ,opacity
= 0.8, marker=dict(color = 'lightgrey',
        line=dict(color='#000000',width=1.5)))

layout = dict(title =  "Missing Values")

fig = dict(data = [trace], layout=layout)
py.iplot(fig)
```

Figure 18 Code snippet to find the missing values and its plot



In [4]:

```
# Drop useless variables
data = data.drop(['Unnamed: 32','id'],axis = 1)

# Reassign target
data.diagnosis.replace(to_replace = dict(M = 1, B = 0), inplace = True)
```

In [5]:

```
# Head
data.head()
```

Out[5]:

us_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710
57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017
69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790
42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520
29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430

In [6]:

```
# describe
data.describe()
```

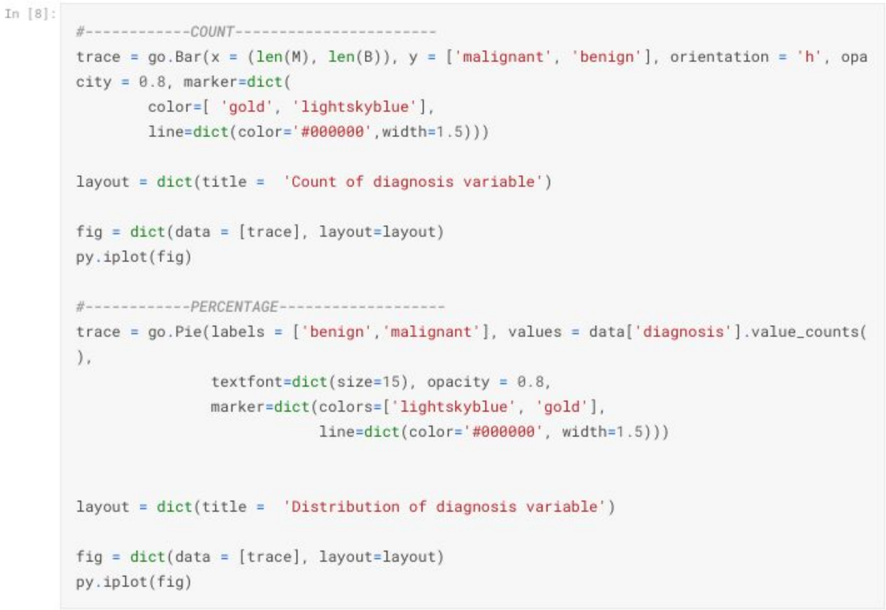
Out[6]:

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	0.372583	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341
std	0.483918	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813
min	0.000000	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380
25%	0.000000	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920
50%	0.000000	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630
75%	1.000000	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400
max	1.000000	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400

In [7]:

```
# 2 datasets
M = data[(data['diagnosis'] != 0)]
B = data[(data['diagnosis'] == 0)]
```

Figure 19 Code snippets to refine and filter the dataset



```
In [8]:
#------------COUNT-----------------------
trace = go.Bar(x = (len(M), len(B)), y = ['malignant', 'benign'], orientation = 'h', opa
city = 0.8, marker=dict(
        color=[ 'gold', 'lightskyblue'],
        line=dict(color='#000000',width=1.5)))

layout = dict(title =  'Count of diagnosis variable')

fig = dict(data = [trace], layout=layout)
py.iplot(fig)

#------------PERCENTAGE-------------------
trace = go.Pie(labels = ['benign','malignant'], values = data['diagnosis'].value_counts(
),
               textfont=dict(size=15), opacity = 0.8,
               marker=dict(colors=['lightskyblue', 'gold'],
                           line=dict(color='#000000', width=1.5)))


layout = dict(title =  'Distribution of diagnosis variable')

fig = dict(data = [trace], layout=layout)
py.iplot(fig)
```

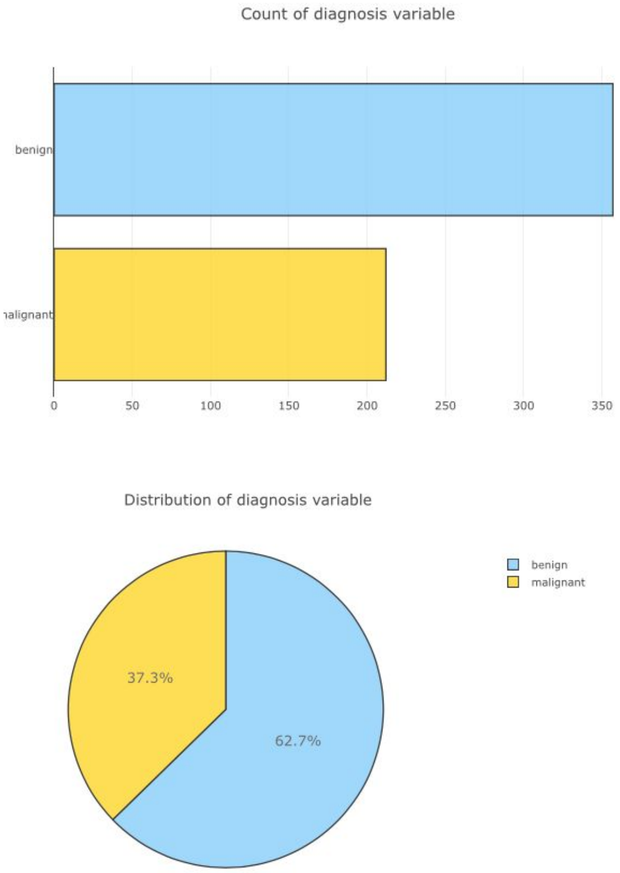
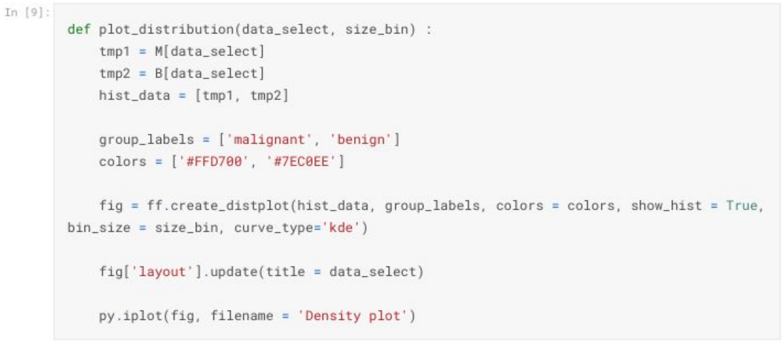
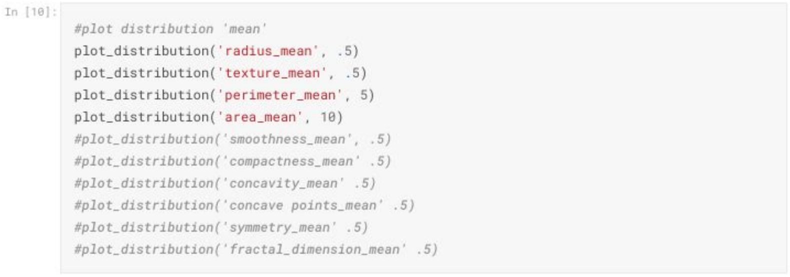


Figure 20 Code snippets for count and distribution of diagnosis variable, and its plots



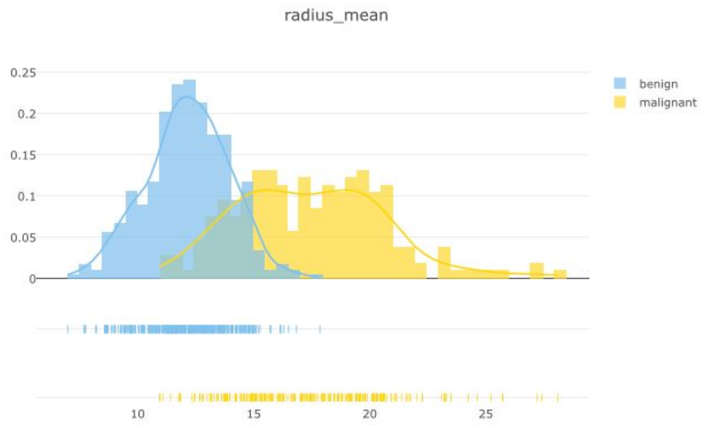
In [9]:

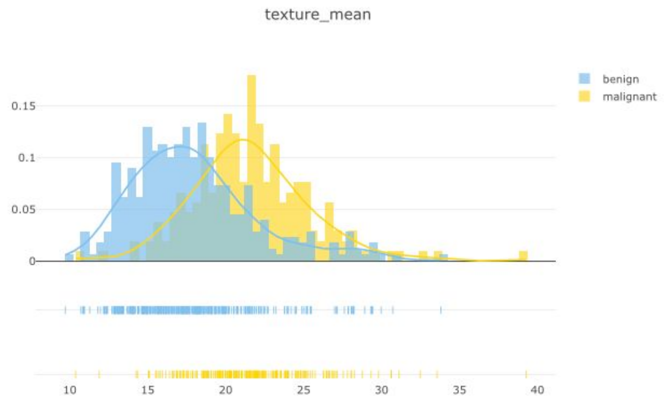
```
def plot_distribution(data_select, size_bin) :
    tmp1 = M[data_select]
    tmp2 = B[data_select]
    hist_data = [tmp1, tmp2]

    group_labels = ['malignant', 'benign']
    colors = ['#FFD700', '#7EC0EE']

    fig = ff.create_distplot(hist_data, group_labels, colors = colors, show_hist = True,
bin_size = size_bin, curve_type='kde')

    fig['layout'].update(title = data_select)

    py.iplot(fig, filename = 'Density plot')
```



In [10]:

```
#plot distribution 'mean'
plot_distribution('radius_mean', .5)
plot_distribution('texture_mean', .5)
plot_distribution('perimeter_mean', 5)
plot_distribution('area_mean', 10)
#plot_distribution('smoothness_mean', .5)
#plot_distribution('compactness_mean' .5)
#plot_distribution('concavity_mean' .5)
#plot_distribution('concave points_mean' .5)
#plot_distribution('symmetry_mean' .5)
#plot_distribution('fractal_dimension_mean' .5)
```





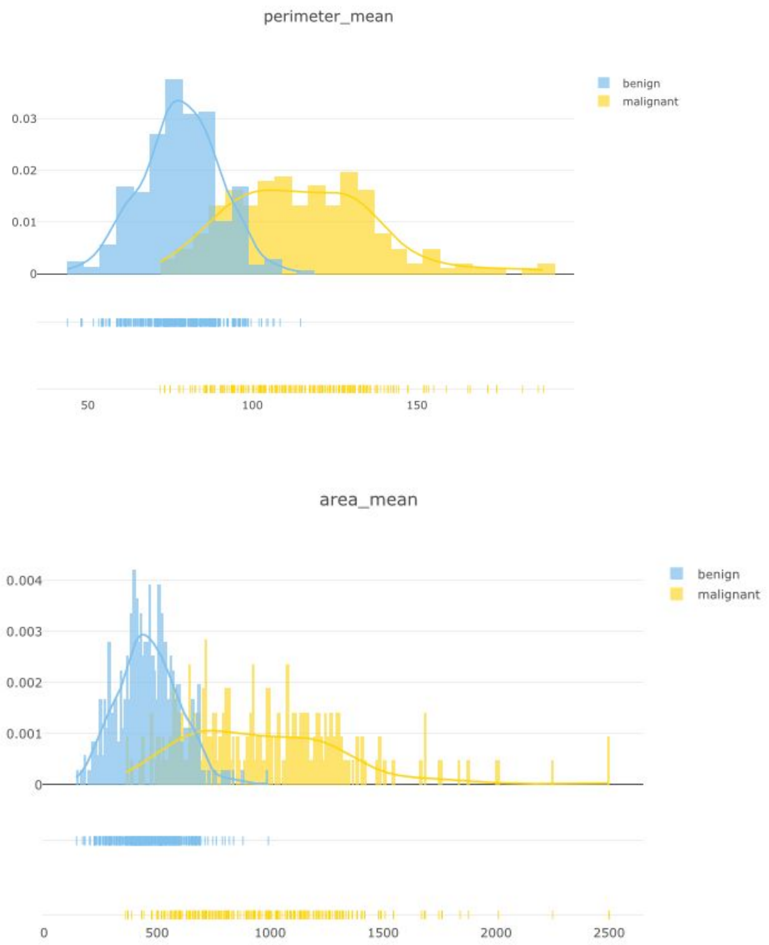
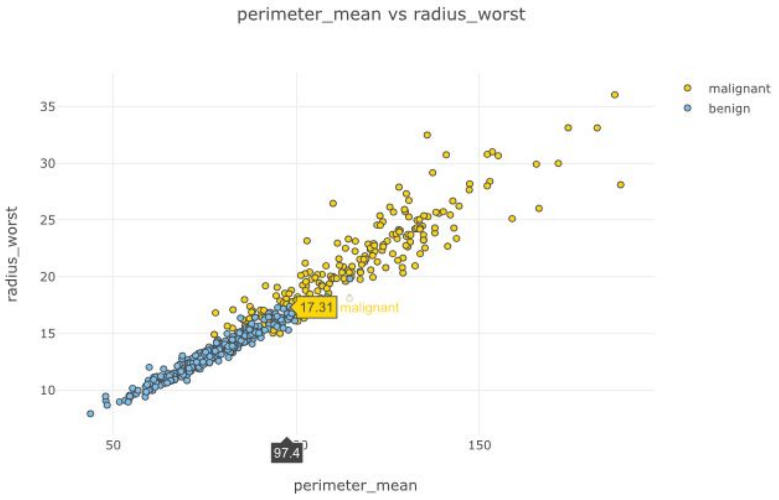


Figure 21 Code snippet for four parameters and its plots



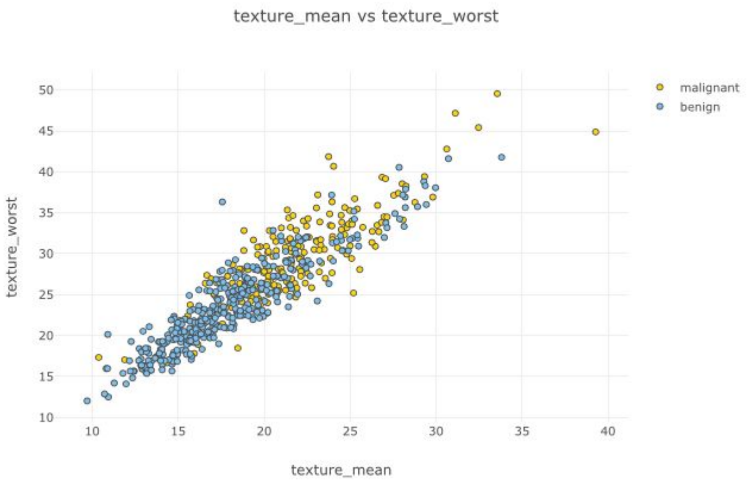


Figure 22 Plot distributions for the four parameters

IMPLEMENTATION

Through a series of questions about symptoms, it diagnoses the health condition of patient.

- Language: Python.
- Modules used: scikit-learn, pandas, numpy
- Model: Decision Tree

SAMPLE CODE (CHATBOT)

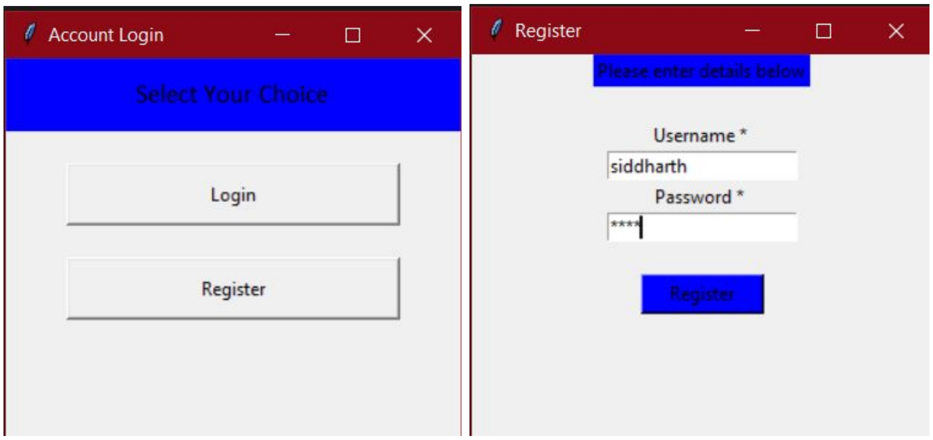
```
import pandas as pd

from sklearn import preprocessing

from sklearn.tree import DecisionTreeClassifier,_tree

import numpy as np

from sklearn.model_selection import train_test_split
from sklearn import model_selection
from sklearn.tree import export_graphviz
import warnings
warnings.filterwarnings("ignore", category=DeprecationWarning)
training = pd.read_csv('Training.csv')
testing  = pd.read_csv('Testing.csv')
cols     = training.columns
cols     = cols[:-1]
x        = training[cols]
y        = training['prognosis']
y1       = y
reduced_data = training.groupby(training['prognosis']).max()
#mapping strings to numbers
le = preprocessing.LabelEncoder()
le.fit(y)
y = le.transform(y)
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.33, random_state=42)
testx    = testing[cols]
testy    = testing['prognosis']
testy    = le.transform(testy)
clf1  = DecisionTreeClassifier()
clf = clf1.fit(x_train,y_train)
#print(clf.score(x_train,y_train))
#print ("cross result========")
#scores = cross_validation.cross_val_score(clf, x_test, y_test, cv=3)
#print (scores)
#print (scores.mean())
```

```
#print(clf.score(testx,testy))
importances = clf.feature_importances_
indices = np.argsort(importances)[::-1]
features = cols
#feature_importances
#for f in range(10):
#       print("%d. feature %d - %s (%f)" % (f + 1, indices[f], features[indices[f]]
,importances[indices[f]]))
print("Please reply Yes or No for the following symptoms")
def print_disease(node):
    #print(node)
    node = node[0]
    #print(len(node))
    val  = node.nonzero()
    #print(val)
    disease = le.inverse_transform(val[0])
    return disease
def tree_to_code(tree, feature_names):
    tree_ = tree.tree_
    #print(tree_)
    feature_name = [
        feature_names[i] if i != _tree.TREE_UNDEFINED else "undefined!"
        for i in tree_.feature
    ]
    #print("def tree({}):".format(", ".join(feature_names)))
    symptoms_present = []
    def recurse(node, depth):
```

```
    indent = "  " * depth
    if tree_.feature[node] != _tree.TREE_UNDEFINED:
        name = feature_name[node]
        threshold = tree_.threshold[node]
        print(name + " ?")
        ans = input()
        ans = ans.lower()
        if ans == 'yes':
            val = 1
        else:
            val = 0
        if  val <= threshold:
            recurse(tree_.children_left[node], depth + 1)
        else:
            symptoms_present.append(name)
            recurse(tree_.children_right[node], depth + 1)
    else:
        present_disease = print_disease(tree_.value[node])
        print( "You may have " +  present_disease )
        red_cols = reduced_data.columns
        symptoms_given = red_cols[reduced_data.loc[present_disease].values[0].nonzero()]
        print("symptoms present  " + str(list(symptoms_present)))
        print("symptoms given "  +  str(list(symptoms_given)) )
        confidence_level = (1.0*len(symptoms_present))/len(symptoms_given)
        print("confidence level is " + str(confidence_level))
recurse(0, 1)
```

OUTPUTS:

- Registration:



Please enter details below

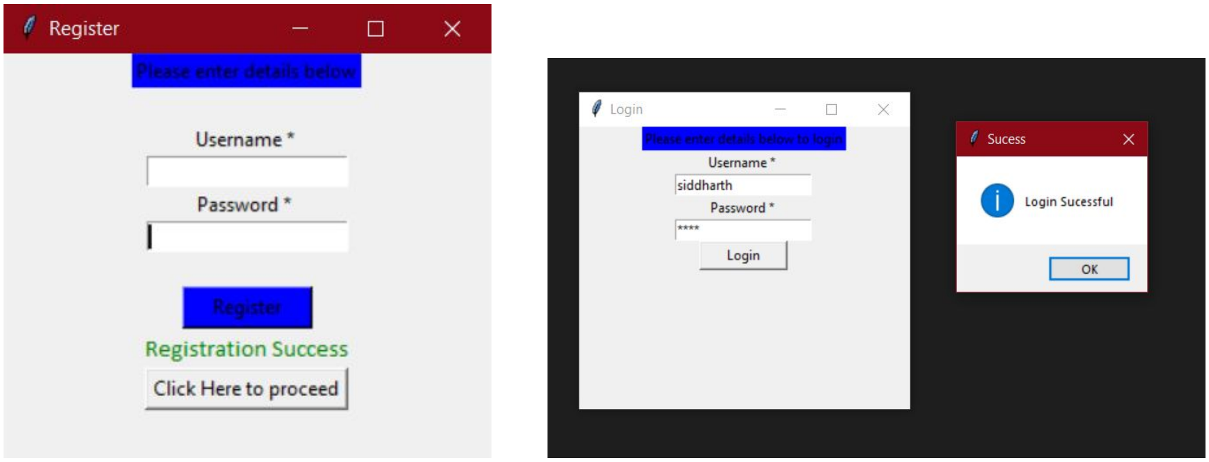
Username * siddharth

Password * ****

Figure 23 User interface of the chatbot and the login dialogue box

- Login

The first interface the user will see is the account login box. The user can login using their registered username and password.



Please enter details below

Username *

Password *

Registration Success

Please enter details below to login

Username * siddharth

Password * ****

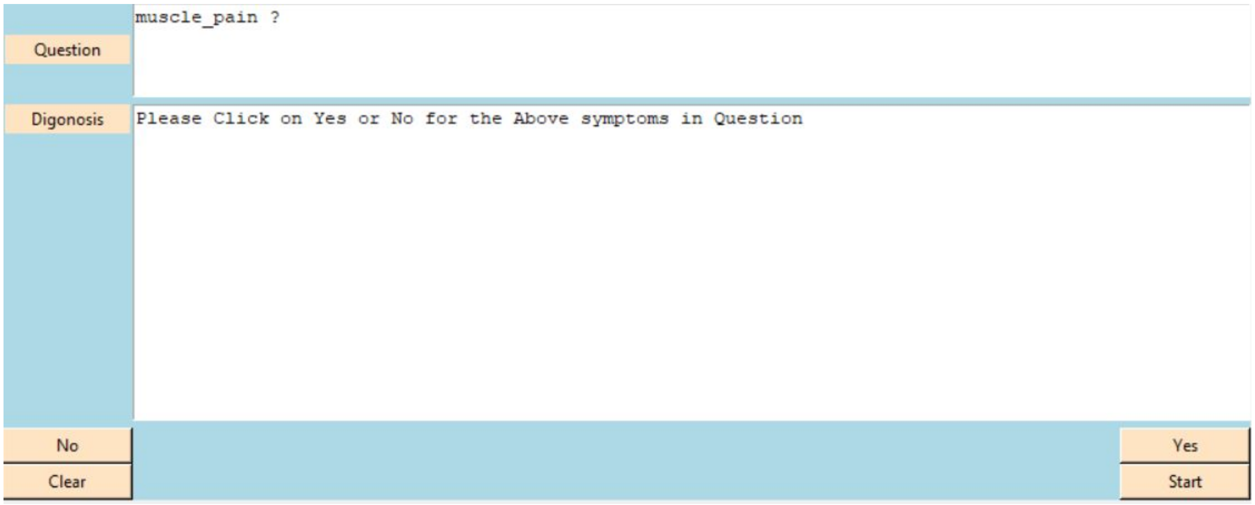
Login Sucessful

Figure 24 User interface for registration, login after registration and successful login message.

If the user has not yet registered, they can do so by entering their choice of username and password. They can login it and if it is verified, the successful login message will be displayed.

- Q&A Console:

- **Disease is predicted according to specific symptoms by asking Yes/No questions:**

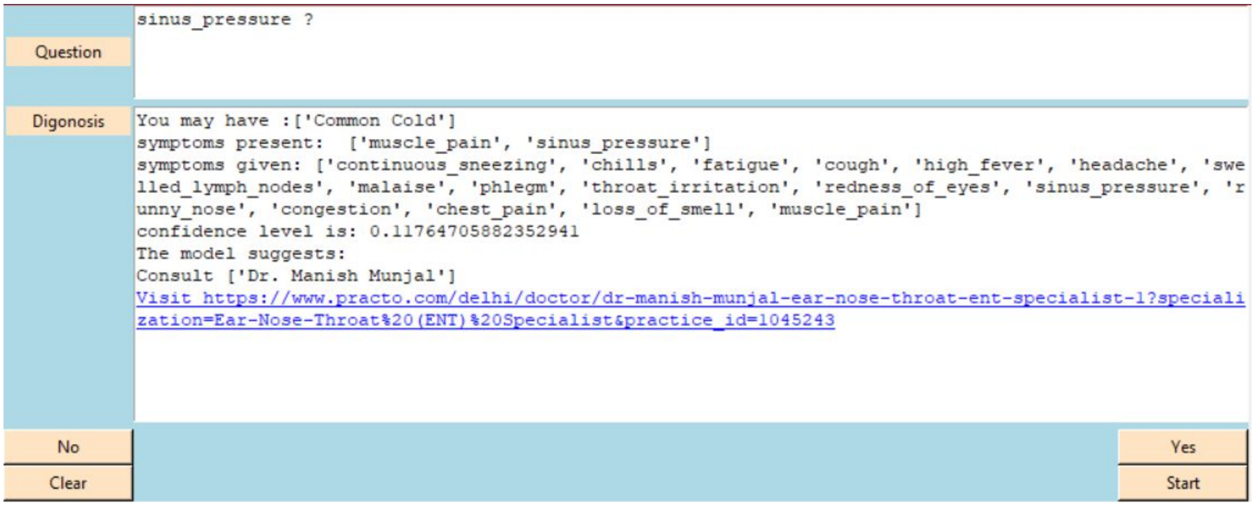


```
muscle_pain ?
```

```
Please Click on Yes or No for the Above symptoms in Question
```

Figure 25 Chatbot Q&A console where the user is asked for symptoms, with the user choice boxes:’yes’,’no’,’clear’ and ‘start’

- **After the series of questions, the disease is predicted and the model suggests to consult a specialized doctor in the respective disease.**



```
sinus_pressure ?
```

```
You may have :['Common Cold']
symptoms present:  ['muscle_pain', 'sinus_pressure']
symptoms given: ['continuous_sneezing', 'chills', 'fatigue', 'cough', 'high_fever', 'headache', 'swe
lled_lymph_nodes', 'malaise', 'phlegm', 'throat_irritation', 'redness_of_eyes', 'sinus_pressure', 'r
unny_nose', 'congestion', 'chest_pain', 'loss_of_smell', 'muscle_pain']
confidence level is: 0.11764705882352941
The model suggests:
Consult ['Dr. Manish Munjal']
Visit https://www.practo.com/delhi/doctor/dr-manish-munjal-ear-nose-throat-ent-specialist-1?speciali
zation=Ear-Nose-Throat%20(ENT)%20Specialist&practice_id=1045243
```

Figure 26 Chatbot Q&A console predicting an illness, with its symptoms and a doctor suggestion

SAMPLE CODE (Online Disease Detection and Prediction Analysis WebApp)

- Home.html

```
<!DOCTYPE html>
<html lang="en">
 <head>
  <!-- Required meta tags -->
  <meta charset="utf-8" />
  <meta
   name="viewport"
   content="width=device-width, initial-scale=1, shrink-to-fit=no"
```

```
    />

    <!-- Bootstrap CSS -->
    <link
      rel="stylesheet"
      href="https://stackpath.bootstrapcdn.com/bootstrap/4.3.1/css/bootstrap.min.css"

integrity="sha384-ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/j6cY/iJTQUOhcWr7x9JvoRxT2MZw1T"
      crossorigin="anonymous"
    />
    <title>Online Disease Predictor</title>
    <style>
      section.pricing {
        background: lightgreen;
        background: linear-gradient(to right, red, yellow);
      }
      .pricing .card {
        border: none;
        border-radius: 1rem;
        transition: all 0.2s;
        box-shadow: 0 0.5rem 1rem 0 rgba(0, 0, 0, 0.1);
      }
      .pricing hr {
        margin: 1.5rem 0;
      }
      .pricing .card-title {
        margin: 0.5rem 0;
```

```
  font-size: 0.9rem;
  letter-spacing: 0.1rem;
  font-weight: bold;
}
.pricing .card-price {
  font-size: 3rem;
  margin: 0;
}
.pricing .card-price .period {
  font-size: 0.8rem;
}
.pricing ul li {
  margin-bottom: 1rem;
}
.pricing .text-muted {
  opacity: 0.7;
}
.pricing .btn {
  font-size: 90%;
  border-radius: 5rem;
  letter-spacing: 0.1rem;
  font-weight: bold;
  padding: 1rem;
  opacity: 0.7;
  transition: all 0.2s;
}
```

```
    /* Hover Effects on Card */

    @media (min-width: 992px) {
      .pricing .card:hover {
        margin-top: -0.25rem;
        margin-bottom: 0.25rem;
        box-shadow: 0 0.5rem 1rem 0 rgba(0, 0, 0, 0.3);
      }
      .pricing .card:hover .btn {
        opacity: 1;
      }
    }
  </style>
</head>

<body>
  <nav class="navbar navbar-expand-lg navbar-light bg-light">
    <a class="navbar-brand" href="{% url 'home' %}"
      ><b>Online Disease Detection and Prediction Analysis</b></a
    >
    <button
      class="navbar-toggler"
      type="button"
      data-toggle="collapse"
      data-target="#navbarNavAltMarkup"
      aria-controls="navbarNavAltMarkup"
      aria-expanded="false"
    aria-label="Toggle navigation"
  >
    <span class="navbar-toggler-icon"></span>
  </button>
  <div class="collapse navbar-collapse" id="navbarNavAltMarkup">
    <div class="navbar-nav ml-auto">
      <a class="nav-item nav-link mr-3" href="{% url 'heart' %}"
        >Heart Disease Prediction
      </a>
      <a class="nav-item nav-link mr-3" href="{% url 'diabetes' %}"
        >Diabetes Prediction
      </a>
      <a class="nav-item nav-link mr-3" href="{% url 'breast' %}"
        >Breast Cancer Prediction</a
      >
      <a class="nav-item nav-link mr-3" href="{% url 'bmi' %}"
        >Body Mass Index(Health Status)
      </a>
    </div>
  </div>
</nav>
<div
  class="jumbotron"
  style="
    background-image: url('/static/zora.jpg');
    background-size: 100%auto;
  "
```

```
    >
            <h1 class="display-5 text-center"><b></b></h1>
      <br /><br /><br />
      <p class="text-right text-light">
        Developed by
        <strong>
          Siddharth Das</strong>
      </p>
      <p class="lead text-right">
        <a
          class="btn btn-dark btn-lg"
          href="https://github.com/iamsiddharthdas"
          role="button"
          >View on GitHub</a
        >
      </p>
    </div>
    <h1 class="text-center"></h1>
    <p class="lead text-center" style="font-size: 20px">
      <strong><b>
            A specialized Web App for prediction of Heart Disease,Breast Cancer and Diabetics
Prediction along with BMI Status
      </strong></b>
    </p>
    <section class="pricing py-5">
      <div class="container">
        <div class="row">
```

```
<!-- Free Tier -->
<div class="col-lg-4">
  <div class="card mb-5 mb-lg-0">
    <div class="card-body">
      <h5 class="card-title text-muted text-uppercase text-center">
        HEART DISEASE PREDICTION
      </h5>
      <hr />
      <p class="">
            Heart disease is one of the most critical human diseases in the world which affects humans liefstyle very badly.
          Do you face shortness of Breath? Do you feel the numbness or weakness in your legs or amrs?
            Do you feel the pain in your chest or any sort of discomforts? Don't worry we are here to help you :)
      Avail our service at free of cost anytime anywhere...


      </p>
      <a
        href="{% url 'heart' %}"
        class="btn btn-block btn-danger text-uppercase"
        >KNOW YOUR HEART STATUS</a
      >
    </div>
  </div>
</div>
<!-- Plus Tier -->
<div class="col-lg-4">
```

```
  <div class="card mb-5 mb-lg-0">
    <div class="card-body">
      <h5 class="card-title text-muted text-uppercase text-center">
        DIABETES PREDICTION
      </h5>

      <hr />
      <p class="">
           Diabetics a serious raising concern among humans which has no age barriers.Can be controlled through proper diets and workouts.
             We are here at your service.If you have blurred vision,frequent urination,increased thirst,hunger,fatigue dont take a chance..
        Please click the below button to know your diabetics status.

      </p>
      <a
        href="{% url 'diabetes' %}"
        class="btn btn-block btn-dark text-uppercase"
        >KNOW YOUR DIABETES STATUS</a
      >
    </div>
  </div>
</div>
<!-- Pro Tier -->
<div class="col-lg-4">
  <div class="card">
    <div class="card-body">
      <h5 class="card-title text-muted text-uppercase text-center">
```

```
                BREAST CANCER PREDICTION
              </h5>

              <hr />
              <p class="">
                Breast cancer is one of the most common cancers among women
                worldwide, representing the majority of new cancer cases and
                cancer-related deaths according to global statistics, making
                  it a significant public health problem in today’s society.Common symptoms includes from Lump in breast,bloody discharge from niple,
                      change in the shape or texture in the nipple or breast..Don't be scared with these symptoms they are curable with proper medications.

              </p>
              <a
                href="breast"
                class="btn btn-block btn-primary text-uppercase"
                >KNOW BREAST CANCER STATUS</a
              >
            </div>
          </div>
        </div>
      </div>
    </div>
  </section>

  <!-- Optional JavaScript -->
  <!-- jQuery first, then Popper.js, then Bootstrap JS -->
```

```
    <script
      src="https://code.jquery.com/jquery-3.3.1.slim.min.js"

integrity="sha384-q8i/X+965DzO0rT7abK41JStQIAqVgRVzpbzo5smXKp4YfRvH+8abtTE1Pi6jizo"
      crossorigin="anonymous"
    ></script>
    <script
      src="https://cdnjs.cloudflare.com/ajax/libs/popper.js/1.14.7/umd/popper.min.js"

integrity="sha384-UO2eT0CpHqdSJQ6hJty5KVphtPhzWj9WO1clHTMGa3JDZwrnQq4sF86dIHNDz0W1"
      crossorigin="anonymous"
    ></script>
    <script
      src="https://stackpath.bootstrapcdn.com/bootstrap/4.3.1/js/bootstrap.min.js"

integrity="sha384-JjSmVgyd0p3pXB1rRibZUAYoIIy6OrQ6VrjIEaFf/nJGzIxFDsf4x0xIM+B07jRM"
      crossorigin="anonymous"
    ></script>
  </body>
</html>
```

OUTPUT

This is a web app mainly for prediction of Heart Disease, Diabetes prediction and Breast Cancer prediction trained using machine learning with Kaggle datasets. And it also includes a BMI calculator for regular health status checkup.

Homepage:

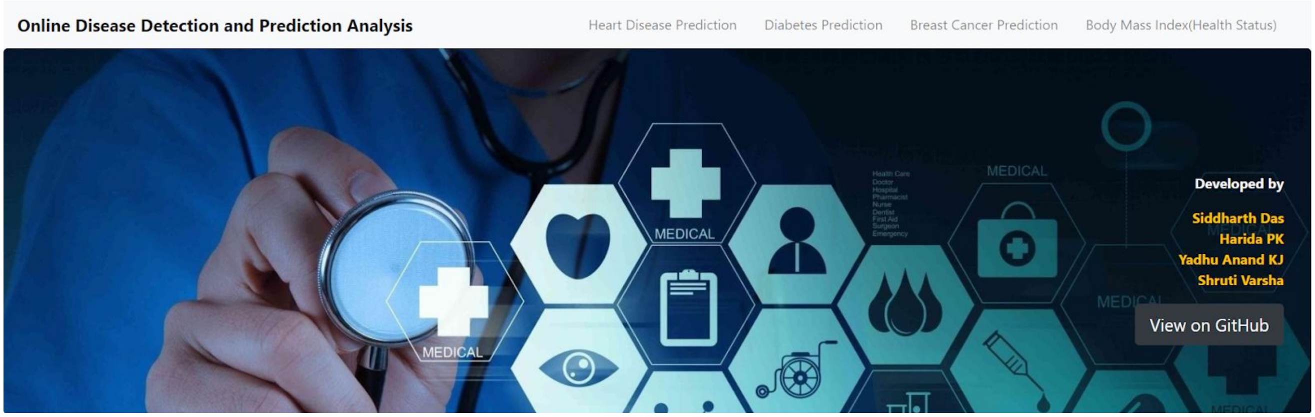
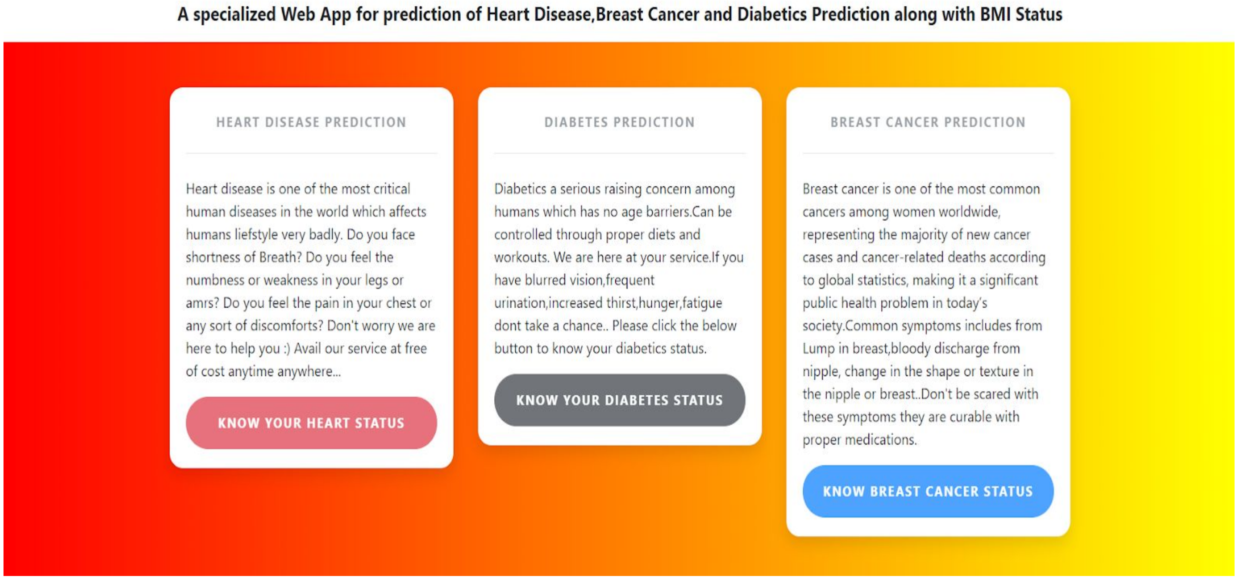


Figure 27 Homepage of the web app

Menu:


A specialized Web App for prediction of Heart Disease,Breast Cancer and Diabetics Prediction along with BMI Status

HEART DISEASE PREDICTION

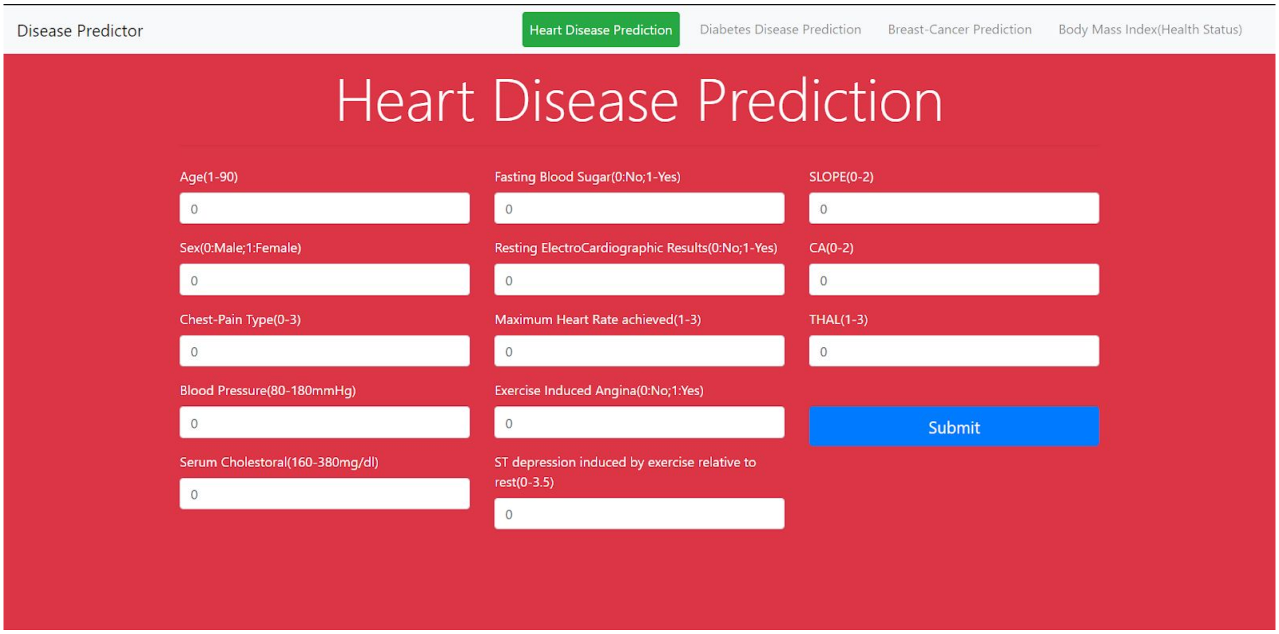
Heart disease is one of the most critical human diseases in the world which affects humans liefstyle very badly. Do you face shortness of Breath? Do you feel the numbness or weakness in your legs or amrs? Do you feel the pain in your chest or any sort of discomforts? Don't worry we are here to help you :) Avail our service at free of cost anytime anywhere...

DIABETES PREDICTION

Diabetics a serious raising concern among humans which has no age barriers.Can be controlled through proper diets and workouts. We are here at your service.If you have blurred vision,frequent urination,increased thirst,hunger,fatigue dont take a chance.. Please click the below button to know your diabetics status.

BREAST CANCER PREDICTION

Breast cancer is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society.Common symptoms includes from Lump in breast,bloody discharge from nipple, change in the shape or texture in the nipple or breast..Don't be scared with these symptoms they are curable with proper medications.



Heart Disease Prediction

Age(1-90)
0

Sex(0:Male;1:Female)
0

Chest-Pain Type(0-3)
0

Blood Pressure(80-180mmHg)
0

Serum Cholestoral(160-380mg/dl)
0

Fasting Blood Sugar(0:No;1-Yes)
0

Resting ElectroCardiographic Results(0:No;1-Yes)
0

Maximum Heart Rate achieved(1-3)
0

Exercise Induced Angina(0:No;1:Yes)
0

ST depression induced by exercise relative to rest(0-3.5)
0

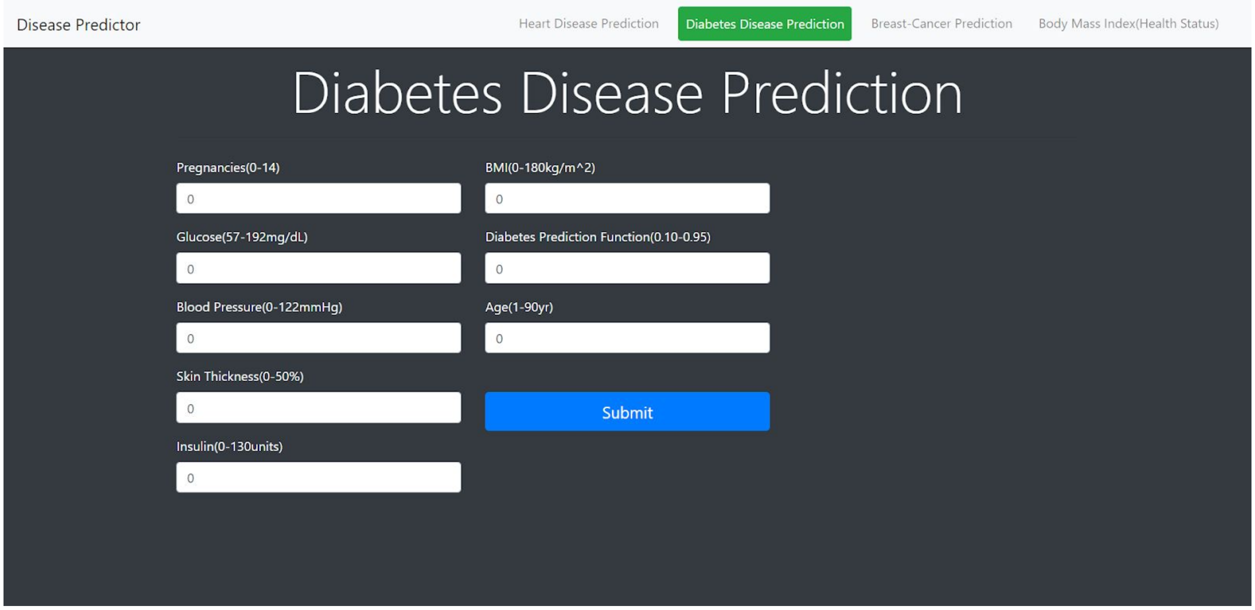
SLOPE(0-2)
0

CA(0-2)
0

THAL(1-3)
0

Figure 28 Heart disease prediction page

Heart disease is a very critical illness. The user can enter values for the attributes asked within the range which is mentioned with the attributes to check if they have heart disease or not. This result will be displayed after they click Submit.



Diabetes Disease Prediction

Pregnancies(0-14)
0

Glucose(57-192mg/dL)
0

Blood Pressure(0-122mmHg)
0

Skin Thickness(0-50%)
0

Insulin(0-130units)
0

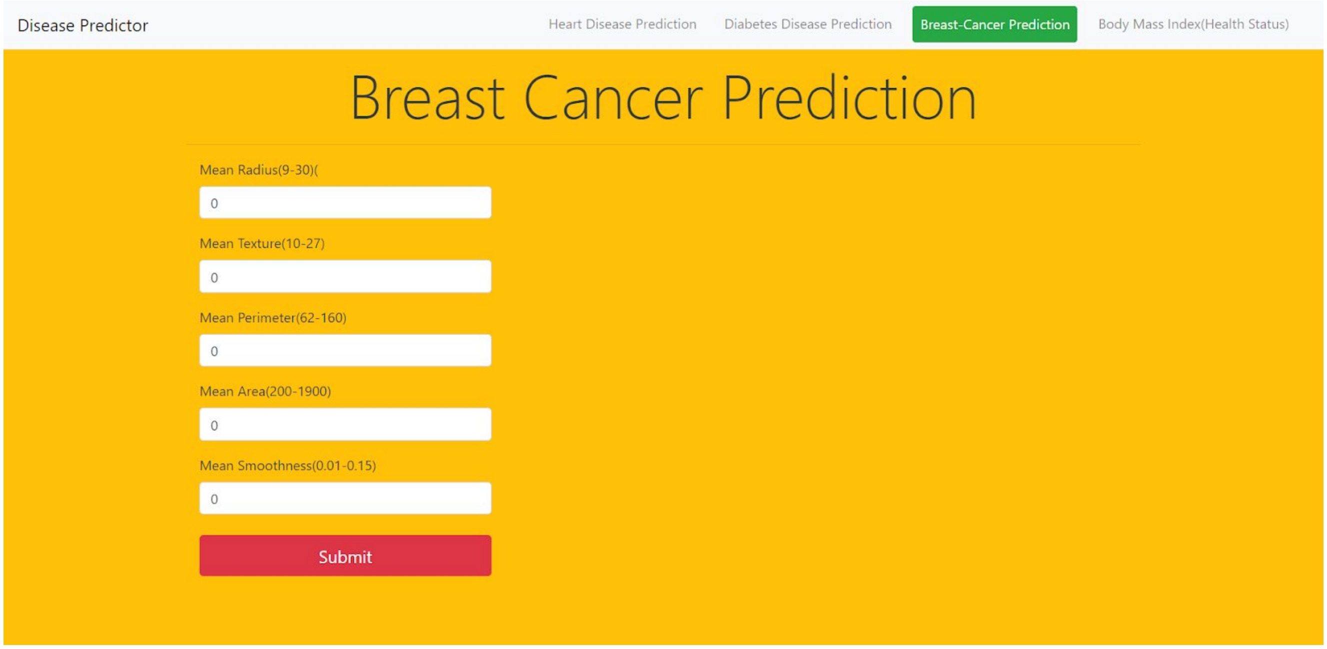
BMI(0-180kg/m^2)
0

Diabetes Prediction Function(0.10-0.95)
0

Age(1-90yr)
0

Figure 29 Diabetes disease prediction page

Similar to heart disease, the user can enter their health data asked and can check if they have the illness or not.



Breast Cancer Prediction

Mean Radius(9-30)(
0

Mean Texture(10-27)
0

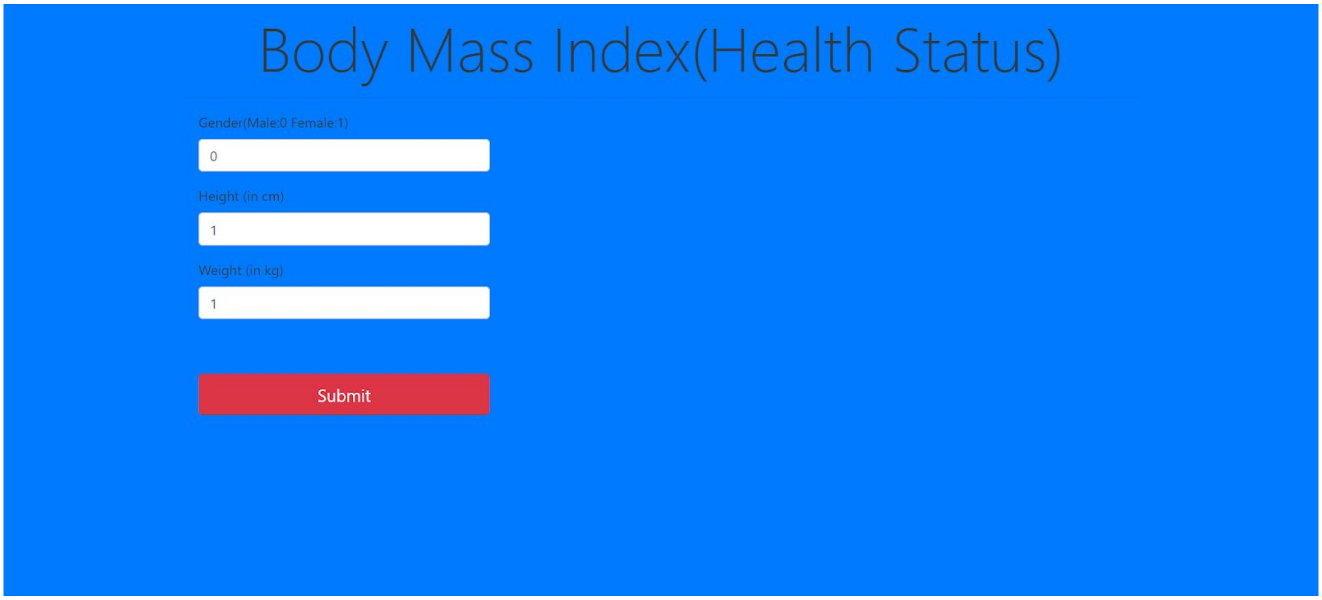
Mean Perimeter(62-160)
0

Mean Area(200-1900)
0

Mean Smoothness(0.01-0.15)
0

Figure 30 Breast cancer prediction page

By entering the attributes asked within the range, the user can check whether they have breast cancer or not.



Body Mass Index(Health Status)

Gender(Male:0 Female:1)
0

Height (in cm)
1

Weight (in kg)
1

Figure 31 Body Mass Index status page

Users can get their body mass index status by entering the attributes asked. Status such as ‘overweight’,’underweight’,’normal’,’obese’ will be displayed according to the user’s inputs.

CONCLUSION

We have implemented this system to provide a user friendly and interactive environment for early diagnosis of the disease the patient is suffering from. This system can not only be used by patients but also by doctors and any other medical staff to identify the disease and give immediate first aid. Through this system, the time required to detect the disease will get decreased and more lives can be saved. This report presents the methodology for implementing a disease prediction system using the symptoms of the patients with the technology of Chatbot for the betterment of the healthcare industry. The system built is very easy to access and can be updated with modern technology from time to time. Early and accurate diagnosis of the disease along with online doctor recommendation can help in providing the best and on-time treatment to the patients.

REFERENCES

[1] Dharwadkar, R., & Deshpande, N. A. (2018). A medical ChatBot. *International Journal of Computer Trends and Technology (IJCTT)*, *60*(1), 41-45.

[2] Kumar, A., & Soni, H. K. (2020). *Early Diagnosis of Human Disease using Artificial Intelligence* (No. 3459). EasyChair

[3] Dahiya, M. (2017). A tool of conversation: Chatbot. *International Journal of Computer Sciences and Engineering*, *5*(5), 158-161.

[4] Laumer, S., Maier, C., & Gubler, F. T. (2019). Chatbot acceptance in healthcare: Explaining user adoption of conversational agents for disease diagnosis.

[5] Mathew, R. B., Varghese, S., Joy, S. E., & Alex, S. S. (2019, April). Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 851-856). IEEE.

[6] Bali, M., Mohanty, S., Chatterjee, S., Sarma, M., & Puravankara, R. Diabot: A Predictive Medical Chatbot using Ensemble Learning.

[7] Bozic, J., Tazl, O. A., & Wotawa, F. (2019, April). Chatbot testing using AI planning. In *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)* (pp. 37-44). IEEE.

[8] Shakya, S. (2020). The Disease prediction system using Machine learning. *International Journal of Engineering and Computer Science*, *9*(2), 24948-24952.

[9] KC, G. P., Ranjan, S., Ankit, T., & Kumar, V. A Personalized Medical Assistant Chatbot: MediBot.

[10] Amato, F., Marrone, S., Moscato, V., Piantadosi, G., Picariello, A., & Sansone, C. (2017). Chatbots Meet eHealth: Automatizing Healthcare. In *WAIAH@ AI* IA* (pp. 40-49).

[11] Sharma, M., Singh, G., & Singh, R. (2019). An advanced conceptual diagnostic healthcare framework for diabetes and cardiovascular disorders. *arXiv preprint arXiv:1901.10530*.

[12] GARG, A., JINDAL, R., ASHISH, S., & SHAFALI, S. Y. A PAPER ON CHATBOT FOR MEDICAL DIAGNOSIS.

[13] Chaki, J., Ganesh, S. T., Cidham, S. K., & Theertan, S. A. (2020). Machine Learning and Artificial Intelligence based Diabetes Mellitus Detection and Self-Management: A Systematic Review. *Journal of King Saud University-Computer and Information Sciences*.

[14] Shetty, S. V., Karthik, G. A., & Ashwin, M. (2019, July). Symptom Based Health Prediction using Data Mining. In *2019 International Conference on Communication and Electronics Systems (ICCES)* (pp. 744-749). IEEE.

[15]Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2020). HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System. *IEEE Access*, *8*, 133034-133050.

[16]Patil, P. B., Shastry, P. M., & Ashokumar, P. S. (2020). MACHINE LEARNING BASED ALGORITHM FOR RISK PREDICTION OF CARDIO VASCULAR DISEASE (CVD). *Journal of Critical Reviews*, *7*(9), 836-844.

[17]Manogaran, G., Varatharajan, R., & Priyan, M. K. (2018). Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system. *Multimedia tools and applications*, *77*(4), 4379-4399.

[18]Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists
Yun Liu, PhD; Timo Kohlberger, PhD; Mohammad Norouzi, PhD; George E. Dahl, PhD; Jenny L. Smith, DO:

[19]Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System Alejandro Rodríguez-Ruiz, Elizabeth Krupinski, Jan-Jurre Mordang, Kathy Schilling, Sylvia H. Heywang-Köbrunner, Ioannis Sechopoulos, Ritse M. Mann

[20]Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists
Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Mireille Broeders, Gisella Gennaro, Paola Clauser, Thomas H Helbich,JNCI: Journal of the National Cancer Institute, Volume 111, Issue 9, September 2019.