SOCIO PREDICTOR

PROJECT REPORT

Submitted by

SHIVCHARAN.B	**Register No.: 15TD0336**
JAGAN.R	**Register No.: 15TI0214**
GLADSON SOLOMON.S	**Register No.: 15TD0245**

Under the guidance of

Dr.J.MADHUSUDANAN

in partial fulfillment of the requirements for the degree

of

BACHELOR OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SRI MANAKULA VINAYAGAR ENGINEERING COLLEGE

MADAGADIPET, PUDUCHERRY-605107,

APRIL 2019

SRI MANAKULA VINAYAGAR ENGINEERING COLLEGE

PONDICHERRY UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that the project work entitled “**SOCIO PREDICTOR**” is a bonafide work done by **SHIVCHARAN.B [15TD0336], JAGAN.R [15TI0214], GLADSON SOLOMON.S [15TDO245]** in partial fulfillment of the requirement, for the award of B.Tech Degree in Computer Science and Engineering by Pondicherry University during the academic year 2018-2019.

PROJECT GUIDE **HEAD OF THE DEPARTMENT**

Submitted for the University Examination held on ……………………………..

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

We are very thankful and grateful to our beloved guide, **Dr.J.MADHUSUDANAN** whose great support in valuable advices, suggestions and tremendous help enabled us in completing our project. He/She has been a great source of inspiration to us.

We also sincerely thank our Head of the Department, **Mr. K. PREMKUMAR** whose continuous encouragement and sufficient comments enabled us to complete our project report.

We thank all our **Staff members** who have been by our side always and helped us with our project. We also sincerely thank all the lab technicians for their help as in the course of our project development.

We would also like to extend our sincere gratitude and grateful thanks to our Director cum Principal, **Dr. V. S. K. VENKATACHALAPATHY** for having extended the Research and Development facilities of the department.

We are grateful to our Founder Chairman **Shri. N. KESAVAN**. He has been a constant source of inspiration right from the beginning.

We would like to express our faithful and grateful thanks to our Chairman & Managing Director, **Shri. M. DHANASEKARAN** for his support.

We would also like to thank our Vice Chairman **Shri. S.V. SUGUMARAN**, for providing us with pleasant learning environment.

We would like to thank our Secretary **Dr. K. NARAYANASAMY** for his support.

We wish to thank our **family members and friends** for their constant encouragement, constructive criticisms and suggestions that has helped us in timely completion of this project.

Last but not the least; we would like to thank the **ALMIGHTY** for His grace and blessings over us throughout the project.

ABSTRACT:

In Current World where technology is playing it’s part Perfectly We Humans are always in progressing state each and every day.In field of computers Automation has made most of the jobs easy which are huge ones in previous generation.Machines have became a part of each and every day routine life of each individual living in earth.In field of medicine though every advancements are made daily it is still a Challenge to feed a Knowledge to a machine which in turn allows it to act as a Decision Maker.There are many challenges in data Preprocessing to Obtain Accurate Results.

Several machine Learning Algorithms are used to train datasets after preprocessing into Train and Test variables with process of Fitting the inputs to that particular Algorithm despite of Regressor Or Classifier.Several Machine Learning models are implemented in field of Medicine which Includes Cancer Prediction,Survivability determination etc.Our Machine Learning Model is a Diabetic Predictor which performs data preprocessing in Independent variables like Glucose,TricepsThickness,Age,BMI and Blood Pressure to Determine the outcome of Diabetes.

Most of the Diabetic Models are feeded with model inputs,whereas we feeded our Machine Learning model with datasets of 15,000 real patients from Government website.We Used Python FLASK concept to make our machine learning model into a FLASK Web Application.Along with the existing feature of Diabetes detection,we added a health advisor report Generated automatically based on the Patient’s data.It provides the current status of patient with his Symptoms,Diseases and Conditions that can attack him with tips and Suggestions to overcome it.

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER-1

MACHINE LEARNING

INTRODUCTION:

Machine learning is an artificial intelligence (AI) discipline geared toward the technological development of human knowledge. Machine learning allows computers to handle new situations via analysis, self-training, observation and experience.Machine learning facilitates the continuous advancement of computing through exposure to new scenarios, testing and adaptation, while employing pattern and trend detection for improved decisions in subsequent (though not identical) situations.Machine learning is often confused with data mining and knowledge discovery in databases (KDD), which share a similar methodology. Machine learning applications include syntactic pattern recognition, natural language processing, search engines, computer vision and machine perception. Machine learning applications include syntactic pattern recognition, natural language processing, search engines, computer vision and machine perception.It's difficult to replicate human intuition in a machine, primarily because human beings often learn and execute decisions unconsciously.Like children, machines require an extended training period when developing broad algorithms geared toward the dictation of future behavior. Training techniques include rote learning, parameter adjustment, macro-operators, chunking, explanation-based learning, clustering, mistake correction, case recording, multiple model management, back propagation, reinforcement learning and genetic algorithms. We know that there is a relation between the features and the labels, in a sense that the value of the features somehow determines the value of the label, or that the value of YY is conditioned on the value of XX.Formally, we express that by saying that there's a conditional probability P(Y|X)P(Y|X). We can utilize this to compress the two distributions we had in the 1st point into a single joint distribution

$$P(X,Y)=P(X)P(Y|X)P(X,Y)=P(X)P(Y|X).$$

With these points, we can define a statistical model that formalizes everything we know about the learning problem.We know that the values of (xi,yi)(xi,yi) in the dataset are just a random sample from a bigger population. We can formalize this fact by saying that the values of xixi and yiyi are realizations of two random variables X and Y with probability distributions PXPX and PYPY respectively. Note that from now on, unless it's pointed otherwise, we'll refer to random variables with uppercase letters X,Y,Z,…X,Y,Z,… while we refer to the values that these variables could take with lowercase letters x,y,z,…x,y,z,….We know that there are some rules on the values of XX and YY that we expect any realization of them to follow. In the diabetes example, we know that a value of a blood glucose test (a component of the xx vector) cannot be negative, so it belongs to a *space* of positive numbers. We also know that value of the label can either be 0 (non-diabetic) or 1 (diabetic), so it belongs to a *space* containing only 0 and 1.These kind of rules define what we formally call a *space*. We say that X takes values drawn from the **input space** XX, and YY from the **output space**YY.In the process to estimate the target function from the sample dataset, because we cannot investigate every single function that could exist in the universe (there is an infinite kinds of them), we attempt to make a hypothesis about the form of ff. We can hypothesize that ff is a linear function of the input, or a cubic one, or some sophisticated non-linear function represented by a neural network. Whatever a form we hypothesize about the the target function ff, it defines a space of possible functions we call the **hypothesis space** HH.

If we hypothesize that the function ff takes the form ax+bax+b, then we're actually defining a hypothesis space HH where:

H={h:X→Y|h(x)=ax+b}H={h:X→Y|h(x)=ax+b}

That is the set of all functions hh mapping the input space to the output space, taking the form ax+bax+b. The task of the machine learning process now is to pick from HH a single concrete function hh that best estimates the target function ff.One way of evaluating how well a hypothesis function is estimating the target function is by noticing that miss-labeling by the hypothesis should be discouraged, we obviously don't want our hypothesis function to make too many mistakes! This is the role of the **loss function** L=L(y,y^)L=L(y,y^) (also called the cost function).

The loss function takes the true label yy of some feature vector xx, and the estimated label by our hypothesis y^=h(x)y^=h(x), then it examines how far our estimate is the from the true value and reports how much do we lose by using that hypothesis.

Using the loss function, we can calculate the performance of a hypothesis function hh on the entire dataset by taking the mean of the losses on each sample. We call this quantity the **in-sample error**, or **the empirical risk**:

Remp(h)=1m∑i=1mL(yi,h(xi))Remp(h)=1m∑i=1mL(yi,h(xi))

It is *empirical* because we calculate it form the empirical data we sampled in the dataset, but why don't we call an empirical error?. The reason behind that is notational; if we used the term *error* we will end up using EE to refer to it in the math, and this could lead into confusion with the notation for the expected value EE, hence we use *risk* and RR instead. The notational choice is also justified by the fact that *error* and *risk* are semantically close.

1.1CLASSIFICATION OF MACHINE LEARNING:

1.1.1 Supervised Learning in the context of artificial intelligence , is a type of system in which both input and desired output data are provided. Input and output data are labelled for classification to provide a learning basis for future data processing.Training data for supervised learning includes a set of examples with paired input subjects and desired output (which is also referred to as the supervisory signal). In supervised learning for image processing, for example, an AI system might be provided with labelled pictures of vehicles in categories such as cars and trucks. After a sufficient amount of observation, the system should be able to distinguish between and categorize unlabeled images, at which time training can be said to be complete.

Supervised learning models have some advantages over the unsupervised approach, but they also have limitations. The systems are more likely to make judgments that humans can relate to, for example, because humans have provided the basis for decisions. However, in the case of a retrieval-based method, supervised learning systems have trouble dealing with new information. If a system with categories for cars and trucks is presented with a bicycle, for example, it would have to be incorrectly lumped in one category or the other. If the AI system was generative, however, it may not know what the bicycle is but would be able to recognize it as belonging to a separate category.

The roots of machine learning algorithms come from Thomas Bayes, who was English statistician who lived in the 18th century. His paper '*An Essay Towards Solving a Problem in the Doctrine of Chances* underpins Bayes' Theorem, which is widely applied in the field of statistics.In the 19th century, Pierre-Simon Laplace published '*Théorieanalytique des probabilités'*, expanding on the work of Bayes and defining what we know of today as Bayes' Theorem. Shortly before that, Adrien-Marie Legendre had described the "least squares" method, also widely used today in supervised learning.The 20th century is the period when the majority of publicly known discoveries have been made in this field. Andrey Markov invented Markov chains, which he used to analyze poems. Alan Turing proposed a learning machine that could become artificially intelligent, basically foreshadowing genetic algorithms. Frank Rosenblatt invented the *Perceptron*, sparking huge excitement and great coverage in the media.

But then the 1970s saw a lot of pessimism around the idea of AI—and thus, reduced funding—so this period is called an *AI winter*. The rediscovery of backpropagation in the 1980s caused a resurgence in machine learning research. And today, it's a hot topic once again.

The late Leo Breiman distinguished between two statistical modeling paradigms: Data modeling and algorithmic modeling. "Algorithmic modeling" means more or less the machine learning algorithms like the *random forest*.

Machine learning and statistics are closely related fields. The ideas of machine learning, from methodological principles to theoretical tools, have had a long prehistory in statistics. *Data science* as a placeholder term for the overall problem that machine learning specialists and statisticians are both implicitly working on.The machine learning field stands on two main pillars called *supervised learning* and *unsupervised learning*. Some people also consider a new field of study—*deep learning*—to be separate from the question of supervised vs. unsupervised learning.

Supervised learning is when a computer is presented with examples of inputs and their desired outputs. The goal of the computer is to learn a general formula which maps inputs to outputs. This can be further broken down into Semi-supervised learning, which is when the computer is given an incomplete training set with some outputs missing Active learning, which is when the computer can only obtain training labels for a very limited set of instances. When used interactively, their training sets can be presented to the user for labeling.Reinforcement learning, which is when the training data is only given as feedback to the program's actions in the dynamic environment, such as driving a vehicle or playing a game against an opponent

In contrast, unsupervised learning is when no labels are given at all and it's up to the algorithm to find the structure in its input. Unsupervised learning can be a goal in itself when we only need to discover hidden patterns.

Deep learning is a new field of study which is inspired by the structure and function of the human brain and based on artificial neural networks rather than just statistical concepts. Deep learning can be used in both supervised and unsupervised approaches.

The easiest possible algorithm is linear regression. Sometimes this can be graphically represented as a straight line, but despite its name, if there's a polynomial hypothesis, this line could instead be a curve. Either way, it models the relationships between scalar dependent variable yy and one or more explanatory values denoted by xx.

1.1.2 Unsupervised learning is the training of an artificial intelligence using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.In unsupervised learning, an AI system may group unsorted information according to similarities and differences even though there are no categories provided. AI systems capable of unsupervised learning are often associated with generative learning models, although they may also use a retrieval-based approach (which is most often associated with the systems that may use either supervised or unsupervised learning approaches.In unsupervised learning, an AI system is presented with unlabeled, uncategorised data and the system's algorithms act on the data without prior training. The output is dependent upon the coded algorithms. Subjecting a system to unsupervised learning is one way of testing AI.Unsupervised learning algorithms can perform more complex processing tasks than supervised learning systems. However, unsupervised learning can be more unpredictable than the alternate model.

While an unsupervised learning AI system might, for example, figure out on its own how to sort - cats from dogs, it might also add unforeseen and undesired categories to deal with unusual breeds, creating clutter instead of order.Consider a machine (or living organism) which receives some sequence of inputs x1, x2, x3, . . ., where xt is the sensory input at time t. This input, which we will often call the data, could correspond to an image on the retina, the pixels in a camera, or a sound waveform. It could also correspond to less obviously sensory data, for example the words in a news story, or the list of items in a supermarket shopping basket. One can distinguish between four different kinds of machine learning.

In supervised learning the machine1 is also given a sequence of desired outputs y1, y2, . . . , and the goal of the machine is to learn to produce the correct output given a new input. This output could be a class label (in classification) or a real number (in regression). In reinforcement learning the machine interacts with its environment by producing actions a1, a2, ... These actions affect the state of the environment, which in turn results in the machine receiving some scalar rewards (or punishments) r1, r2, ... The goal of the machine is to learn to act in a way that maximises the future rewards it receives (or minimises the punishments) over its lifetime. Reinforcement learning is closely related to the fields of decision theory (in statistics and management science), and control theory (in engineering).

The fundamental problems studied in these fields are often formally equivalent, and the solutions are the same, although different aspects of problem and solution are usually emphasised. A third kind of machine learning is closely related to game theory and generalized reinforcement learning. Here again the machine gets inputs, produces actions, and receives rewards. However, the environment the machine interacts with is not some static world, but rather it can contain other machines which can also sense, act, receive rewards, and learn. Thus the goal of the machine is to act so as to maximise rewards in light of the other machines' current and future actions. Although there is a great deal of work in game theory for simple systems, the dynamic case with multiple adapting machines remains an active and challenging area of research. Finally, in unsupervised learning the machine simply receives inputs x1, x2, . . ., but obtains neither supervised target outputs, nor rewards from its environment. It may seem somewhat mysterious to imagine what the machine could possibly learn given that it doesn't get any feedback from its environment.

However, it is possible to develop of formal framework for unsupervised learning based on the notion that the machine's goal is to build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, etc. In a sense, unsupervised learning can be thought of as finding patterns in the data above and beyond what would be considered pure unstructured noise. Two very simple classic examples of unsupervised learning are clustering and dimensionality reduction. Let us consider how unsupervised learning relates to statistics and information theory.PCA produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.

Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization. The first principal component of a set of features X1, X2, . . . ,Xp is the normalized linear combination of the features Z1 = φ11X1 + φ21X2 + . . . + φp1Xp that has the largest variance. By P normalized, we mean that p j=1 φ 2 j1 = 1.We refer to the elements φ11, . . . , φp1 as the loadings of the first principal component; together, the loadings make up the principal component loading vector, φ1 = (φ11 φ21 . . . φp1) T .We constrain the loadings so that their sum of squares is equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

Almost all work in unsupervised learning can be viewed in terms of learning a probabilistic model of the data. Even when the machine is given no supervision or reward, it may make sense for the machine to estimate a model that represents the probability distribution for a new input xt given previous inputs x1, . . . , xt−1. In simpler cases where the order in which the inputs arrive is irrelevant or unknown, the machine can build a model of the data which assumes that the data points x1, x2, . . . are independently and identically drawn from some distribution P(x) 2 . Such a model can be used for outlier detection or monitoring. Let x represent patterns of sensor readings from a nuclear power plant and assume that P(x) is learned from data collected from a normally functioning plant. This model can be used to evaluate the probability of a new sensor reading; if this probability is abnormally low, then either the model is poor or the plant is behaving abnormally, in which case one may want to shut it down.

A probabilistic model can also be used for classification. Assume P1(x) is a model of the attributes of credit card holders who paid on time, and P2(x) is a model learned from credit card holders who defaulted on their payments. By evaluating the relative probabilities P1(x 0) and P2(x 0) on a new applicant x 0 , the machine can decide to classify her into one of these two categories. With a probabilistic model one can also achieve efficient communication and data compression. Imagine that we want to transmit, over a digital communication line, symbols x randomly drawn from P(x). For example, x may be letters of the alphabet, or images, and the communication line may be the internet. Intuitively, we should encode our data so that symbols which occur more frequently have code words with fewer bits in them, otherwise we are wasting bandwidth. Shannon's source coding theorem quantifies this by telling us that the optimal number of bits to use to encode a symbol with probability P(x) is − log2 P(x). Using these number of bits for each symbol, the expected coding cost is the entropy of the distribution P. H(P) def = − X x P(x) log2 P(x)

Bayes rule, P(y|x) = P(x|y)P(y) P(x) which follows from the equality P(x, y) = P(x)P(y|x) = P(y)P(x|y), can be used to motivate a coherent statistical framework for machine learning. The basic idea is if the machine is to represent the strength of its beliefs by real numbers, then the only reasonable and coherent way of manipulating these beliefs is to have them satisfy the rules of probability, such as Bayes rule. Therefore, P(X = x) can be used not only to represent the frequency with which the variable X takes on the value x (as in so-called frequentist statistics) but it can also be used to represent the degree of belief that X = x. Similarly, P(X = x|Y = y) can be used to represent the degree of belief that X = x given that one knowns Y = y. 3 From Bayes rule we derive the following simple framework for machine learning. Assume a universe of models Ω; let Ω = {1, . . . , M} although it need not be finite or even countable. The machines starts with some prior beliefs over models m ∈ Ω (we will see many examples of models later), such that PM m=1 P(m) = 1. A model is simply some probability distribution over data points, i.e. P(x|m). For simplicity, let us further assume that in all the models the data is taken to be independently and identically distributed .

1.1.3 Reinforcement learning, in the context of artificial intelligence, is a type of dynamic programming that trains algorithms using a system of reward and punishment.A reinforcement learning algorithm, or agent, learns by interacting with its environment. The agent receives rewards by performing correctly and penalties for performing incorrectly. The agent learns without intervention from a human by maximizing its reward and minimizing its penalty.Reinforcement learning is an approach to machine learning that is inspired by behaviorist psychology. It is similar to how a child learns to perform a new task. Reinforcement learning contrasts with other machine learning approaches in that the algorithm is not explicitly told how to perform a task, but works through the problem on its own.As an agent, which could be a self-driving car or a program playing chess, interacts with its environment, receives a reward state depending on how it performs, such as driving to destination safely or winning a game. Conversely, the agent receives a penalty for performing incorrectly, such as going off the road or being checkmated.The agent over time makes decisions to maximize its reward and minimize its penalty using dynamic programming. The advantage of this approach to artificial intelligence is that it allows an AI program to learn without a programmer spelling out how an agent should perform the task.

Reinforcement Learning, in the context of AI, is a type of dynamic programming that teaches you algorithms using a system of reward and punishment. Deep Reinforcement Learning (DRL) is a fast evolving subdivision of Artificial Intelligence that aims at solving many of our problems. On the one hand, it mirrors human learning by exploring and receiving feedback from environments, much in the lines of artificial general intelligence, or AGI, Reinforcement Learning has also demonstrated success of dramatic game changes, where bipedal agents learn to walk in a simulation.

While supervised Machine Learning trains models based on known answers, Reinforcement Learning, and researchers train the model through an agent, which interacts with the environment. The agent is rewarded every time its actions produce positive results.

Reinforcement Learning though has its roots in reinforcement theories of animal learning has evolved as a solution for the betterment of mankind. Personalization Travel Support System, for example, is a solution that applies the reinforcement learning to analyze and learn customer behaviors and list out the products that the customers wish to buy. If the system selects the right item that the customer wishes to buy then it errands rewards, and gets a penalty, if it fails to do so. In this way, the system learns about user behavior and preferences, which help it redefine its actions, for particular users.

Some of commonly used RL algorithms are **Q-Learning: Q-Learning is an off-policy, model-free RL algorithm based on the well-known Bellman Equation E** in the above equation refers to the expectation, while ƛ refers to the discount factor. We can rewrite it in the form of Q-value.The optimal Q-value, denoted as Q* can be expressed as Two value update methods that are closely related to Q-learning are Policy Iteration and Value Iteration.**State-Action-Reward-State-Action (SARSA):**SARSA, another popular RL algorithm, is quite similar to Q-learning. The key difference between SARSA and Q-learning is that SARSA is an on-policy algorithm. It implies that SARSA learns the Q-value based on the action performed by the current policy instead of the greedy policy.

The Sarsa algorithm is an On-Policy algorithm for TD-Learning. The primary difference between it and Q-Learning is that the maximum reward for the next state is not necessarily used for updating the Q-values. Instead, a new action, and therefore reward is selected using the same policy that determined the original action.**Deep Q Network (DQN):**DQN leverages a Neural Network to estimate the Q-value function. The input for the network is the current, while the output is the corresponding Q-value for each of the action.In 2013, DeepMind applied DQN to Atari game. The input is the raw image of the current game situation. It went through several layers including convolution layer as well as a fully connected layer. The output is the Q-value for each of the actions that the agent can take.Two essential techniques for training DQN are Experience Replay and Separate Target Network.

1.2 HYPER TEXT MARKUP LANGUAGE WITH CASCADING STYLE SHEETS:

HTML is the standard markup language for creating Web pages.

- HTML stands for Hyper Text Markup Language
- HTML describes the structure of Web pages using markup
- HTML elements are the building blocks of HTML pages
- HTML elements are represented by tags
- HTML tags label pieces of content such as "heading", "paragraph", "table", and so on
- Browsers do not display the HTML tags, but use them to render the content of the page

- The versions of HTML is given in table Tab 1.1

Version	Year
HTML	1991
HTML 2.0	1995
HTML 3.2	1997
HTML 4.0.1	1999
XHTML	2000
HTML5	2014

Tab 1.1 Versions Of HTML

HTML elements are the building blocks of HTML pages. With HTML constructs, images and other objects such as interactive forms may be embedded into the rendered page. HTML provides a means to create structured documents by denoting structural semantics for text such as headings, paragraphs, lists, links,, quotes and other items. HTML elements are delineated by *tags*, written using angle brackets.

CASCADING STYLE SHEET:

CSS is the acronym for "Cascading Style Sheet". CSS is a simple design language intended to simplify the process of making web pages presentable. CSS rules are made up of a selector and at least one declaration. A selector is the code that selects the HTML to which you want to apply the style rule. A declaration is made up of at least one CSS property and related property value. CSS properties define the style.CSS handles the look and feel part of a web page. Using CSS, you can control the color of the text, the style of fonts, the spacing between paragraphs, how columns are sized and laid out, what background images or colors are used, layout designs,variations in display for different devices and screen sizes as well as a variety of other effects.CSS is designed to enable the separation of presentation and content, including layout, colors, and fonts.This separation can improve content accessibility, provide more flexibility and control in the specification of presentation characteristics, enable multiple web pages to share formatting by specifying the relevant CSS in a separate .css file, and reduce complexity and repetition in the structural content. Separation of formatting and content also makes it easy to present the same markup page in different styles for different rendering methods, such as on-screen, in print, by voice (via speech-based browser or screen reader), and on Braille-based tactile devices. CSS also has rules for alternate formatting if the content is accessed on a mobile device.TheH1 is the selector for your h1 headers, and the declaration is made up of the color property with a value of red. Simply said, this rule turns all H1 elements red. You'll note that the syntax of the style rule looks different from what you've seen in HTML. The curly braces contain the declarations, each property is followed by a colon, and a semicolon is used after each property. You can have as many properties as you want in a rule.What you just made is a skeleton html document. This is the minimum required information for a web.document and all web documents should contain these basic components.

The first tag in your htmldocument is <html>. This tag tells your browser that this is the start of an html document. The last tag in your document is </html>. This tag tells your browser that this is the end of the htmldocument.The text between the <head> tag and the </head> tag is header information. Header information is not displayed in the browser window.The text between the <title> tags is the title of your document. The <title> tag is used to uniquely identify each document and is also displayed in the title bar of the browser window.The text between the <body> tags is the text that will be displayed in your browser.The text between the <b> and </b> tags will be displayed in a bold font.The first web browser, Mosaic, was introduced in 1993. A year later Netscape, based on Mosaic, was introduced and the net began to become popular. HTML was used in both browsers, but there was no "standard" HTML until the introduction of HTML 2.0.HTML 2.0 was first published in 1995.* HTML 3.0 was published two years later and 4.01 two years after that. HTML 4.01 has been the work of the net ever since.The various versions of CSS are given in the table Tab 1.2

Versions	**Year**
CSS 1	1996
CSS 2	1998
CSS 3	1999

Tab 1.2 Versions Of CSS

1.3 PYTHON IN MACHINE LEARNING:

Python is a general-purpose programming language that can be used on any modern computer operating system. It can be used for processing text, numbers, images, scientific data and just about anything else you might save on a computer. It is used daily in the operations of the Google search engine, the video-sharing website YouTube, NASA and the New York Stock Exchange. These are but a few of the places where Python plays important roles in the success of the business, government, and non-profit organizations; there are many others. ne of the key benefits of Python Programming is its interpretive nature. The Python interpreter and standard library are available in binary or source form from the Python website, and can run seamlessly on all major operating systems. Python Programming language is also freely-distributable, and the same site even has tips and other third-party tools, programs, modules and more documentation.The Python interpreter can be easily extended with new data types or functions in C++, C or any other language callable from C.

The Python Programming language works as an extension for customizable applications. What makes this language so easy to learn is the fact that it uses English keywords rather than punctuation, and it has fewer syntax constructions than other programming languages Benefits of Python Programming Language Interpreted language are ,the language is processed by the interpreter at runtime, like PHP or PERL, so you don’t have to compile the program before execution,Interactive ,you can directly interact with the interpreter at the Python prompt for writing your program.Perfect for beginners, for beginner-level programmers, Python is a great choice as it supports the development of applications ranging from games to browsers to text processing.

Python is also one of the older web development languages out there, made by Guido van Rossum at the National Research Institute for Mathematics and Computer Science in the Netherlands in the early 90s. The language borrows heavily from C, C++, SmallTalk, Unix Shell, Modula-3, ABC, Algol-68 and other scripting languages.

Rossum continues to direct the language progress, although a core development team at the institute now maintains most of it.As mentioned before, English language keywords make up most of the programming in Python. If you master them, you have mastered Python for the most part. This will take some practice, and you need to know the basic concepts before you start off. So let’s begin by looking at them:Python is implicitly and dynamically typed, so you do not have to declare variables. The types are enforced, and the variables are also case sensitive, so var and VAR are treated as two separate variables.

If you want to know how any object work, you just need to type the following:help(object)you can also use the dir(object) command to find out all the methods of a particular option, and you can use object.__doc__ to find out its document string.Python is an interpreted language. This means that it is not converted to computer-readable code before the program is run but at runtime. In the past, this type of language was called a scripting language, intimating its use was for trivial tasks. However, programming languages such as Python have forced a change in that nomenclature. Increasingly, large applications are written almost exclusively in Python.

Beyond Python there are a number of open source libraries generally used to facilitate practical machine learning. In general, these are the main so-called *scientific Python libraries* we put to use when performing elementary machine learning tasks:

numpy - *mainly* useful for its *N*-dimensional array objects

pandas - Python data analysis library, including structures such as dataframes

matplotlib - 2D plotting library producing publication quality figures

scikit-learn - the machine learning algorithms used for data analysis and data mining tasks

STEPS IN ANALYZING THE PROBLEM:

- ✓ Examine your problem.
- ✓ Prepare your data (raw data, feature extraction, feature engineering).
- ✓ Spot-check a set of algorithms.
- ✓ Examine your results.Double-down on the algorithms that worked best.

1.3.1 MACHINE LEARNING APPLICATION USING PYTHON FLASK:

Flask is a web-development framework , this is one of the most-popular module of python for web developers . It is easy to use , quick to learn , and also easy to create applications. It uses jinja2 template for html rendering . all though it is a micro-framework one can create medium level applications with this framework.Flask was originally designed and developed by Armin Ronacher . The Flask framework became wildly popular as an alternative to Django projects with their monolithic structure and dependencies.

Flask's success created a lot of additional work in issue tickets and pull requests. Armin eventually created The Pallets Projects collection of open source code libraries after he had been managing Flask under his own GitHub account for several years. The Pallets Project now serves as the community-driven organization that handles Flask and other related Python libraries such as Lektor, Jinja and several others.Siri, Alexa, Google Now are some of the popular examples of virtual personal assistants. As the name suggests, they assist in finding information, when asked over voice. All you need to do is activate them and ask "What is my schedule for today?", "What are the flights from Germany to London", or similar questions. For answering, your personal assistant looks out for the information, recalls your related queries, or send a command to other resources (like phone apps) to collect info. You can even instruct assistants for certain tasks like "Set an alarm for 6 AM next morning", "Remind me to visit Visa Office day after tomorrow".

Machine learning is an important part of these personal assistants as they collect and refine the information on the basis of your previous involvement with them. Later, this set of data is utilized to render results that are tailored to your preferences.Virtual Assistants are integrated to a variety of platforms. For example:

- Smart Speakers: Amazon Echo and Google Home
- Smartphones: Samsung Bixby on Samsung S8
- Mobile Apps: Google Allo

Traffic Predictions: We all have been using GPS navigation services. While we do that, our current locations and velocities are being saved at a central server for managing traffic. This data is then used to build a map of current traffic. While this helps in preventing the traffic and does congestion analysis, the underlying problem is that there are less number of cars that are equipped with GPS. Machine learning in such scenarios helps to estimate the regions where congestion can be found on the basis of daily experiences.

Online Transportation Networks: When booking a cab, the app estimates the price of the ride. When sharing these services, how do they minimize the detours? The answer is machine learning. Jeff Schneider, the engineering lead at Uber ATC reveals in a an interview that they use ML to define price surge hours by predicting the rider demand. In the entire cycle of the services, ML is playing a major role.Imagine a single person monitoring multiple video cameras! Certainly, a difficult job to do and boring as well. This is why the idea of training computers to do this job makes sense.The video surveillance system nowadays are powered by AI that makes it possible to detect crime before they happen. They track unusual behaviour of people like standing motionless for a long time, stumbling, or napping on benches etc. The system can thus give an alert to human attendants, which can ultimately help to avoid mishaps. And when such activities are reported and counted to be true, they help to improve the surveillance services. This happens with machine learning doing its job at the backend.From personalizing your news feed to better ads targeting, social media platforms are utilizing machine learning for their own and user benefits. Here are a few examples that you must be noticing, using, and loving in your social media accounts, without realizing that these wonderful features are nothing but the applications of ML.

People You May Know: Machine learning works on a simple concept: understanding with experiences. Facebook continuously notices the friends that you connect with, the profiles that you visit very often, your interests, workplace, or a group that you share with someone etc. On the basis of continuous learning, a list of Facebook users are suggested that you can become friends with.

Face Recognition: You upload a picture of you with a friend and Facebook instantly recognizes that friend. Facebook checks the poses and projections in the picture, notice the unique features, and then match them with the people in your friend list. The entire process at the backend is complicated and takes care of the precision factor but seems to be a simple application of ML at the front end.

Similar Pins: Machine learning is the core element of Computer Vision, which is a technique to extract useful information from images and videos. Pinterest uses computer vision to identify the objects (or pins) in the images and recommend similar pins accordingly.There are a number of spam filtering approaches that email clients use. To ascertain that these spam filters are continuously updated, they are powered by machine learning. When rule-based spam filtering is done, it fails to track the latest tricks adopted by spammers. Multi Layer Perceptron, C 4.5 Decision Tree Induction are some of the spam filtering techniques that are powered by ML.Over 325, 000 malwares are detected everyday and each piece of code is 90–98% similar to its previous versions. The system security programs that are powered by machine learning understand the coding pattern. Therefore, they detects new malware with 2–10% variation easily and offer protection against them.

A number of websites nowadays offer the option to chat with customer support representative while they are navigating within the site. However, not every website has a live executive to answer your queries. In most of the cases, you talk to a chatbot. These bots tend to extract information from the website and present it to the customers. Meanwhile, the chatbots advances with time. They tend to understand the user queries better and serve them with better answers, which is possible due to its machine learning algorithms.

Google and other search engines use machine learning to improve the search results for you. Every time you execute a search, the algorithms at the backend keep a watch at how you respond to the results. If you open the top results and stay on the web page for long, the search engine assumes that the the results it displayed were in accordance to the query. Similarly, if you reach the second or third page of the search results but do not open any of the results, the search engine estimates that the results served did not match requirement. This way, the algorithms working at the backend improve the search results.

1.4 SUBLIME TEXT EDITOR FOR PYTHON AND HTML INTEGRATION:

Since Python Libraries Cannot be Used normally in Backend or Javascript in HTML,sublime text editor is used to perform integration with frontend.Sublime Text is a text editor for code, HTML, and prose. It features rich selection of editing commands, including indenting or un-indenting , line joining , multiple selections, regular expression search and replace, incremental find as you type, and preserve case on replace. Create macros, snippets, auto complete, and repeat last action. It has build tool integration ability, automatic build on save, and WinSCP integration for editing remote files via SCP and FTP.Sublime Text Editor is a full featured Text editor for editing local files or a code base. It includes various features for editing code base which helps developers to keep track of changes. Various features that are supported by Sublime are as follows

- Syntax Highlight
- Auto Indentation
- File Type Recognition
- Sidebar with files of mentioned directory
- Macros
- Plug-in and Packages

Sublime Text editor is used as an Integrated Development Editor (IDE) like Visual Studio code and NetBeans. The current version of Sublime Text editor is 3.0 and is compatible with various operating systems like Windows, Linux and MacOS.

Sublime Text editor is supported by the following major operating systems −

- Windows
- Linux and its distributions
- OS X

1.5 MACHINE LEARNING TYPES AND ALGORITHMS:

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, biometric identification, document classification etc. There are so many algorithms available that it can feel overwhelming when algorithm names are thrown around and you are expected to just know what they are and where they fit.

I want to give you two ways to think about and categorize the algorithms you may come across in the field.

- The first is a grouping of algorithms by the **learning style**.
- The second is a grouping of algorithms by **similarity** in form or function (like grouping similar animals together).

There are different ways an algorithm can model a problem based on its interaction with the experience or environment or whatever we want to call the input data.It is popular in machine learning and artificial intelligence textbooks to first consider the learning styles that an algorithm can adopt.This taxonomy or way of organizing machine learning algorithms is useful because it forces you to think about the roles of the input data and the model preparation process and select one that is the most appropriate for your problem in order to get the best result.

When crunching data to model business decisions, you are most typically using supervised and unsupervised learning methods.A hot topic at the moment is semi-supervised learning methods in areas such as image classification where there are large datasets with very few labeled examples.

1.5.1 REGRESSOR ALGORITHMS:

Regression is basically a statistical approach to find the relationship between variables. In machine learning, this is used to predict the outcome of an event based on the relationship between variables obtained from the data-set. It predict the output values based on input features from the data fed in the system.It is mostly used for forecasting and finding out cause and effect relationship between variables.

TYPES:

- ✓ Linear Regression
- ✓ Logistic Regression
- ✓ Polynomial Regression
- ✓ Stepwise Regression
- ✓ Ridge Regression,
- ✓ Lasso Regression,
- ✓ Elastic-Net Regression,
- ✓ Gradient boosting Regression
- ✓ Random Forest Regression

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula $y = c + b*x$, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.Sometimes logistic regressions are difficult to interpret; the Intellectus Statistics tool easily allows you to conduct the analysis, then in plain English interprets the output.

This function fits a polynomial regression model to powers of a single predictor by the method of linear least squares. Interpolation and calculation of areas under the curve are also given.

If a polynomial model is appropriate for your study then you may use this function to fit a k order/degree polynomial to your data, where Y caret is the predicted outcome value for the polynomial model with regression coefficients b_1 to k for each degree and Y intercept b_0. The model is simply a general linear regression model with k predictors raised to the power of i where i=1 to k. A second order (k=2) polynomial forms a quadratic expression (parabolic curve), a third order (k=3) polynomial forms a cubic expression and a fourth order (k=4) polynomial forms a quartic expression.

Some general principles:

- The fitted model is more reliable when it is built on large numbers of observations.
- Do not extrapolate beyond the limits of observed values.
- Choose values for the predictor (x) that are not too large as they will cause overflow with higher degree polynomials; scale x down if necessary.
- Do not draw false confidence from low P values, use these to support your model only if the plot looks reasonable.

More complex expressions involving polynomials of more than one predictor can be achieved by using the general linear regression function. For more detail from the regression, such as analysis of residuals, use the general linear regression function. To achieve a polynomial fit using general linear regression you must first create new workbook columns that contain the predictor (x) variable raised to powers up to the order of polynomial that you want. For example, a second order fit requires input data of Y, x and x^2.Subjective goodness of fit may be assessed by plotting the data and the fitted curve. An analysis of variance is given via the analysis option; this reflects the overall fit of the model. Try to use as few degrees as possible for a model that achieves significance at each degree.The plot function supplies a basic plot of the fitted curve and a plot with confidence bands and prediction bands. You can save the fitted Y values with their standard errors, confidence intervals and prediction intervals to a workbook.

The option to calculate the area under the fitted curve employs two different methods. The first method integrates the fitted polynomial function from the lowest to the highest observed predictor (x) value using Romberg's integration. The second method uses the trapezoidal rule directly on the data to provide a crude estimate.Stepwise regression is the step-by-step iterative construction of a regression model that involves automatic selection of independent variables.Stepwise regression can be achieved either by trying out one independent variable at a time and including it in the regression model if it is statistically significant, or by including all potential independent variables in the model and eliminating those that are not statistically significant, or by a combination of both methods. Tests for significance are conducted via F-tests, t-tests, adjusted R squared, and a few other less common methods. The goal is to find a set of independent variables which significantly influence the dependent variable. Conducting these tests automatically can potentially save time for the individual.Stepwise regression has a number of drawbacks, according to some statisticians. These include incorrect results, an inherent bias in the process itself and the necessity for significant computing power to develop complex regression models through iteration..

Ridge regression is one of the most fundamental regularization technique which is not used by many due to the complex science behind it. If you have an overall idea about the concept of multiple regression, it’s not so difficult to explore the science behind Ridge regression. When the overall idea about regression is same, what makes regularization different is the way how the model coefficients are determined.

Lasso regression is a type of **linear regression** that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of muti_collinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

The acronym “LASSO” stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator.Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models. On the other hand, L2 regularization (e.g. Ridge regression) *doesn't* result in elimination of coefficients or sparse models. This makes the Lasso far easier to interpret than the Ridge.Lasso solutions are quadratic programming problems, which are best solved with software (like Matlab). The solution of lasso implementation is given in fig 1.1

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Fig 1.1 Solution Of Lasso Implementation

Which is the same as minimizing the sum of squares with constraint Σ |Bj≤ s. Some of the βs are shrunk to exactly zero, resulting in a regression model that's easier to interpret.

A **tuning parameter**, λ controls the strength of the L1 penalty. λ is basically the amount of shrinkage:

When λ = 0, no parameters are eliminated. The estimate is equal to the one found with linear regression.

As λ increases, more and more coefficients are set to zero and eliminated (theoretically, when λ = ∞, *all* coefficients are eliminated).

As λ increases, bias increases.

As λ decreases, variance increases.

XGBoost is a software library that you can download and install on your machine, then access from a variety of interfaces. Specifically, XGBoost supports the following main interfaces:

- Command Line Interface (CLI).
- C++ (the language in which the library is written).
- Python interface as well as a model in scikit-learn.
- R interface as well as a model in the caret package.
- Julia.
- Java and JVM languages like Scala and platforms like Hadoop.

The library is laser focused on computational speed and model performance, as such there are few frills. Nevertheless, it does offer a number of advanced features

The implementation of the model supports the features of the scikit-learn and R implementations, with new additions like regularization. Three main forms of gradient boosting are supported:

Gradient Boosting algorithm also called gradient boosting machine including the learning rate.

Stochastic Gradient Boosting with sub-sampling at the row, column and column per split levels.

Regularized Gradient Boosting with both L1 and L2 regularization.

The library provides a system for use in a range of computing environments, not least:

Parallelization of tree construction using all of your CPU cores during training.

Distributed Computing for training very large models using a cluster of machines.

Out-of-Core Computing for very large datasets that don’t fit into memory.

Cache Optimization of data structures and algorithm to make best use of hardware.

The implementation of the algorithm was engineered for efficiency of compute time and memory resources. A design goal was to make the best use of available resources to train the model. Some key algorithm implementation features include:

Sparse Aware implementation with automatic handling of missing data values.

Block Structure to support the parallelization of tree construction.

Continued Training so that you can further boost an already fitted model on new data.

XGBoost is free open source software available for use under the permissive Apache-2 license.

Although most of the Kaggle competition winners use stack or ensemble of various models, one particular model that is part of most of the ensembles is some variant of Gradient Boosting (GBM) algorithm. Take for an example the winner of latest Kaggle competition: Michael Jahrer solution with representation learning in **Safe Driver Prediction.** His solution was a blend of 6 models. 1 LightGBM (a variant of GBM) and 5 Neural Nets. Although his success is attributed to the semi-supervised learning that he used for the structured data, but gradient boosting model has done the useful part too.

Even though GBM is being used widely, many practitioners still treat it as complex black-box algorithm and just run the models using pre-built libraries. The purpose of this post is to simplify a supposedly complex algorithm and to help the reader to understand the algorithm intuitively. I am going to explain the pure vanilla version of the gradient boosting algorithm and will share links for its different variants at the end I have taken base Decision Tree code from fast.ai library and on top of that, I have built my own simple version of basic gradient boosting model.When we try to predict the target variable using any machine learning technique, the main causes of difference in actual and predicted values are **noise, variance, and bias**. Ensemble helps to reduce these factors (except noise, which is irreducible error)

An ensemble is just a collection of predictors which come together (e.g. mean of all predictions) to give a final prediction. The reason we use ensembles is that many different predictors trying to predict same target variable will perform a better job than any single predictor alone. Ensembling techniques are further classified into Bagging and Boosting.**Bagging**is a simple ensembling technique in which we build many *independent* predictors/models/learners and combine them using some model averaging techniques. (e.g. weighted average, majority vote or normal average)We typically take random sample/bootstrap of data for each model, so that all the models are little different from each other. *Each observation is chosen with replacement to be used as input for each of the model. So, each model will have different observations based on the bootstrap process.* Because this technique takes many uncorrelated learners to make a final model, it reduces error by reducing variance. Example of bagging ensemble is **Random Forest models.**

Boosting is an ensemble technique in which the predictors are not made independently, but sequentially.This technique employs the logic in which the subsequent predictors learn from the mistakes of the previous predictors. *Therefore, the observations have an unequal probability of appearing in subsequent models and ones with the highest error appear most. (So the observations are not chosen based on the bootstrap process, but based on the error).* The predictors can be chosen from a range of models like decision trees, regressors, classifiers etc. Because new predictors are learning from mistakes committed by previous predictors, it takes less time/iterations to reach close to actual predictions. But we have to choose the stopping criteria carefully or it could lead to overfitting on training data. **Gradient Boosting** is an example of boosting algorithm.

1.5.2 CLASSIFIER ALGORITHMS:

Classification can be performed on structured or unstructured data. Classification is a technique where we categorize data into a given number of classes. The main goal of a classification problem is to identify the category/class to which a new data will fall under.

TYPES:

- ✓ Support Vector Machines
- ✓ ,Decision Trees
- ✓ Boosted Trees
- ✓ Random Forest
- ✓ Neural Networks
- ✓ Logistic Regression Classifier
- ✓ Naive Bayes Classifier
- ✓ Nearest Neighbour Classifier.

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper_plane that differentiate the two classes very well.In SVM, it is easy to have a linear hyper_plane between these two classes. But, another burning question which arises is, should we need to add this feature manually to have a hyper_plane. No, SVM has a technique called the **kernel trick**. These are functions which takes low dimensional input space and transform it to a higher dimensional space i.e. it converts not separable problem to separable problem, these functions are called kernels. It is mostly useful in non-linear separation problem. Simply put, it does some extremely complex data transformations, then find out the process to separate the data based on the labels or outputs you've defined.

Consider a very basic example that uses titanic data set for predicting whether a passenger will survive or not. Below model uses 3 features/attributes/columns from the data set, namely sex, age and sibsp (number of spouses or children along).

A decision tree is drawn upside down with its root at the top. In the image on the left, the bold text in black represents a condition/**internal node**, based on which the tree splits into branches/ **edges**. The end of the branch that doesn't split anymore is the decision/**leaf**, in this case, whether the passenger died or survived, represented as red and green text respectively.

Although, a real dataset will have a lot more features and this will just be a branch in a much bigger tree, but you can't ignore the simplicity of this algorithm. The **feature importance is clear** and relations can be viewed easily. This methodology is more commonly known as **learning decision tree from data** and above tree is called **Classification tree** as the target is to classify passenger as survived or died. **Regression trees** are represented in the same manner, just they predict continuous values like price of a house. In general, Decision Tree algorithms are referred to as CART or Classification and Regression Trees.

So, what is actually going on in the background? Growing a tree involves deciding on **which features to choose** and **what conditions to use** for splitting, along with knowing when to stop. As a tree generally grows arbitrarily, **you will need to trim it down** for it to look beautiful.

In **Recursive Binary Splitting** procedure all the features are considered and different split points are tried and tested using a cost function. The split with the best cost (or lowest cost) is selected.

Consider the earlier example of tree learned from titanic dataset. In the first split or the root, all attributes/features are considered and the training data is divided into groups based on this split. We have 3 features, so will have 3 candidate splits. Now we will ***calculate how much accuracy each split will cost us, using a function***. ***The split that costs least is chosen***, which in our example is sex of the passenger. This ***algorithm is recursive in nature*** as the groups formed can be sub_divided using same strategy. Due to this procedure, this algorithm is also known as the **greedy algorithm**, as we have an excessive desire of lowering the cost. **This makes the root node as best predictor/classifier.**

The human visual system is one of the wonders of the world. the following sequence of handwritten digits is shown in Fig 1.2:

Fig:1.2 Handwritten Digits

Most people effortlessly recognize those digits as 504192. That ease is deceptive. In each hemisphere of our brain, humans have a primary visual cortex, also known as V1, containing 140 million neurons, with tens of billions of connections between them. And yet human vision involves not just V1, but an entire series of visual cortices - V2, V3, V4, and V5 - doing progressively more complex image processing. We carry in our heads a supercomputer, tuned by evolution over hundreds of millions of years, and superbly adapted to understand the visual world. Recognizing handwritten digits isn't easy. Rather, we humans are stupendously, astoundingly good at making sense of what our eyes show us. But nearly all that work is done unconsciously. And so we don't usually appreciate how tough a problem our visual systems solve.

The difficulty of visual pattern recognition becomes apparent if you attempt to write a computer program to recognize digits like those above. What seems easy when we do it ourselves suddenly becomes extremely difficult. Simple intuitions about how we recognize shapes - "a 9 has a loop at the top, and a vertical stroke in the bottom right" - turn out to be not so simple to express algorithmically. When you try to make such rules precise, you quickly get lost in a morass of exceptions and caveats and special cases. It seems hopeless.

Neural networks approach the problem in a different way. The idea is to take a large number of handwritten digits and then develop a system which can learn from those training examples. In other words, the neural network uses the examples to automatically infer rules for recognizing handwritten digits. Furthermore, by increasing the number of training examples, the network can learn more about handwriting, and so improve its accuracy. So while I've shown just 100 training digits above, perhaps we could build a better handwriting recognizer by using thousands or even millions or billions of training examples.

KNN makes predictions using the training dataset directly.Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances. For regression this might be the mean output variable, in classification this might be the mode (or most common) class value.

To determine which of the K instances in the training dataset are most similar to a new input a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance.Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (xi) across all input attributes j.

Euclidean Distance(x, xi) = sqrt(sum((xj – xij)^2))

Other popular distance measures include:

Hamming Distance: Calculate the distance between binary vectors..

Manhattan Distance: Calculate the distance between real vectors using the sum of their absolute difference. Also called City Block Distance.

Minkowski Distance: Generalization of Euclidean and Manhattan distance.

There are many other distance measures that can be used, such as Tanimoto, Jaccard, Mahalanobis and cosine distance. You can choose the best distance metric based on the properties of your data. If you are unsure, you can experiment with different distance metrics and different values of K together and see which mix results in the most accurate models.

Euclidean is a good distance measure to use if the input variables are similar in type (e.g. all measured widths and heights). Manhattan distance is a good measure to use if the input variables are not similar in type (such as age, gender, height, etc.).

The value for K can be found by algorithm tuning. It is a good idea to try many different values for K (e.g. values from 1 to 21) and see what works best for your problem.

The computational complexity of KNN increases with the size of the training dataset. For very large training sets, KNN can be made stochastic by taking a sample from the training dataset from which to calculate the K-most similar instances.

KNN has been around for a long time and has been very well studied. As such, different disciplines have different names for it, for example:

Instance-Based Learning: The raw training instances are used to make predictions. As such KNN is often referred to as instance-based learning or a case-based learning (where each training instance is a case from the problem domain).

Lazy Learning: No learning of the model is required and all of the work happens at the time a prediction is requested. As such, KNN is often referred to as a lazy learning algorithm.

Non-Parametric: KNN makes no assumptions about the functional form of the problem being solved. As such KNN is referred to as a non-parametric machine learning algorithm.

KNN can be used for regression and classification problems.When KNN is used for regression problems the prediction is based on the mean or the median of the K-most similar instances.When KNN is used for classification, the output can be calculated as the class with the highest frequency from the K-most similar instances. Each instance in essence votes for their class and the class with the most votes is taken as the prediction.

Class probabilities can be calculated as the normalized frequency of samples that belong to each class in the set of K most similar instances for a new data instance. For example, in a binary classification problem (class is 0 or 1)

If you are using K and you have an even number of classes it is a good idea to choose a K value with an odd number to avoid a tie. And the inverse, use an even number for K when you have an odd number of classes.

Ties can be broken consistently by expanding K by 1 and looking at the class of the next most similar instance in the training dataset.KNN works well with a small number of input variables (p), but struggles when the number of inputs is very large. Each input variable can be considered a dimension of a p-dimensional input space. For example, if you had two input variables x1 and x2, the input space would be 2-dimensional.As the number of dimensions increases the volume of the input space increases at an exponential rate.

In high dimensions, points that may be similar may have very large distances. All points will be far away from each other and our intuition for distances in simple 2 and 3-dimensional spaces breaks down. This might feel unintuitive at first, but this general problem is called the “Curse of Dimensionality“.

Rescale Data: KNN performs much better if all of the data has the same scale. Normalizing your data to the range [0, 1] is a good idea. It may also be a good idea to standardize your data if it has a Gaussian distribution.

Address Missing Data: Missing data will mean that the distance between samples can not be calculated. These samples could be excluded or the missing values could be imputed.

Lower Dimensionality: KNN is suited for lower dimensional data. You can try it on high dimensional data (hundreds or thousands of input variables) but be aware that it may not perform as well as other techniques. KNN can benefit from feature selection that reduces the dimensionality of the input feature space.

CHAPTER-2

LITERATURE SURVEY

2.1.Disease Predictor by Machine Learning over big data from health-care communities

Author:**MIN CHEN , YIXUE HAO, KAI HWANG, LU WANG, LIN WANG**

Abstract:With Machine Learning growth in biomedical and healthcare communities, accurate analysis of medical data benefits early disease detection, patient care, and community services. However, the analysis accuracy is reduced when the quality of medical data is incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks.The aim to predict diseases with higher accuracy results.

2.2 Machine Learning in medical application

Author:**GEORGE D.MAGOULAS AND ANDRIANA PRENTZA**

Abstract:Machine Learning (ML) provides methods, techniques, and tools that can help solving diagnostic and prognostic problems in a variety of medical domains. ML is being used for the analysis of the importance of clinical parameters and their combinations for prognosis, e.g. prediction of disease progression, extraction of medical knowledge for outcome research, therapy planning and support, and for the overall patient management. ML is also being used for data analysis, such as detection of regularities in the data by appropriately dealing with imperfect data, interpretation of continuous data used in the Intensive Care Unit, and intelligent alarming resulting in effective and efficient monitoring It is argued that the successful implementation of ML methods can help the integration of computer-based systems in the healthcare environment providing opportunities to facilitate and enhance the work of medical experts and ultimately to improve the efficiency and quality of medical care.

2.3 Comparative analysis of Machine Learning methods for classification type decision problem in healthcare

Author:**NAHIT EMANET, HALIL ROZ, NAZAN BAYRAM, DURSUN DELEN**

Abstract:Advanced analytical techniques are gaining popularity in addressing complex classification type decision problems in many fields including healthcare and medicine. In this exemplary study, using digitized signal data, they developed predictive models employing three machine learning methods to diagnose an asthma patient based solely on the sounds acquired from the chest of the patient in a clinical laboratory.

2.4 .A Few Useful Things to Know about Machine Learning

Author:**PEDRO DOMINGOS**

Abstract:Machine learning systems automatically learn programs from data. This is often a very attractive alternative to manually constructing them, and in the last decade the use of machine learning has spread rapidly throughout computer science and beyond. Machine learning is used in Web search, spam filters, recommender systems, ad placement, credit scoring, fraud detection, stock trading, drug design, and many other applications. A recent report from the McKinsey Global Institute asserts that machine learning will be the driver of the next big wave of innovation . Several fine textbooks are available to interested practitioners and researchers However, much of the “folk knowledge” that is needed to successfully develop machine learning applications is not readily available in them. As a result, many machine learning projects take much longer than necessary or wind up producing less- than-ideal results. Yet much of this folk knowledge is fairly easy to communicate.

2.5 .Machine that Learn and Teach Seamlessly.

Author:**GARY STEIN AND CLAYTON BARHAM.**

Abstract:This paper describes an investigation into creating agents that can learn how to perform a task by observing an expert, then seamlessly turn around and teach the same task to a less proficient person. These agents are taught through observation of expert performance and thereafter refined through unsupervised practice of the task, all on a simulated environment. A less proficient human is subsequently taught by the now-trained agent through a third approach—coaching, executed through a haptic device. This approach addresses tasks that involve complex psychomotor skills. A machine-learning algorithm called PIGEON is used to teach the agents. A prototype is built and then tested on a task involving the manipulation of a crane to move large container boxes in a simulated shipyard. Two evaluations were performed—a proficiency test and a learning rate test. These tests were designed to determine whether this approach improves the human learning more than self-experimentation by the human. While the test results do not conclusively show that our approach provides improvement over self-learning, some positive aspects of the results suggest great potential for this approach.

2.6 .Automated Microscopy and Machine Learning for Expert-Level Malaria Field Diagnosis.

Author:**CHARLES B.DELAHUNT AND BENJAMIN K.**

Abstract:The optical microscope is one of the most widely used tools for diagnosing infectious diseases in the developing world. Due to its reliance on trained microscopists, field microscopy often suffers from poor sensitivity, specificity, and reproducibility. The goal of this work, called the Autoscope, is a low-cost automated digital microscope coupled with a set of computer vision and classification algorithms, which can accurately diagnose of a variety of infectious diseases, targeting use-cases in the developing world. Our initial target is malaria, because of the high difficulty of the task and because manual microscopy is currently a central but highly imperfect tool for malaria work in the field. In addition to diagnosis, the algorithm performs species identification and quantitation of parasite load, parameters which are critical in many field applications but which are not effectively determined by rapid diagnostic tests (RDTs). They have built a hardware prototype which can scan approximately 0.1 μL of blood volume in a standard

Giemsa- stained thick smear blood slide in approximately 20 minutes. They have also developed a comprehensive machine learning framework, leveraging computer vision and machine learning techniques including support vector machines (SVMs) and convolutional neural networks (CNNs). The Autoscope has undergone successful initial field testing for malaria diagnosis in Thailand.

2.7 .A Health big data Analytics- A TECHNOLOGY SURVEY

Author:**GASPARD HARERIMANA AND HUNG KOOK PARK**

Abstract::Because of the vast availability of data, there has been an additional focus on the health industry and an increasing number of studies that aim to leverage the data to improve healthcare have been conducted. The health data are growing increasingly large, more complex, and its sources have increased tremendously to include computerized physician order entry, electronic medical records, clinical notes, medical images, cyber-physical systems, medical Internet of Things, genomic data, and clinical decision support systems. New types of data from sources like social network services and genomic data are used to build personalized healthcare systems, hence health data are obtained in various forms, from varied sources, contexts, technologies, and their nature can impede a proper analysis. Any analytical research must overcome these obstacles to mine data and produce meaningful insights to save lives. In this paper, we investigate the key challenges, data sources, techniques, technologies, as well as future directions in the field of big data analytics in healthcare. We provide a do-it-yourself review that delivers a holistic, simplified, and easily understandable view of various technologies that are used to develop an integrated health analytic application.

2.8 LITERATURE SURVEY CONCLUSION WITH PROBLEM DESCRIPTION:

From the above papers , developing an **Python flaskapplication** for a disease in medical field for health care application is necessary and developing an machine learning model for **detecting diabetes in patients** with its type and a detailed health advisory report will be useful in machine learning field **in replacement of doctors** with machines in future

CHAPTER-3
SYSTEM STUDY

3.1 EXISTING SYSTEM:

In existing system the machine learning developing model for diabetes prediction (diab-predo) uses either classification or regression algorithms to train and test a random data set and with a help of fit method a straight line is drawn to fit all the points in that particular graph for plotting using matplotlib library function in python as well as data frames and numpylibraries.This model developed will be usually implemented into an application hosted in a website or localhost as a software.The outcome predicted(either positive or negative for a particular patient) will be displayed in a result page whereas the home page get inputs like patient name, glucose value and insulin value which will be processed in a machine learning model to return a output to the result page

3.1.1 **ISSUES IN EXISTING SYSTEM:**

Lots of machine learning model that are developed into an application can be used only for a single purpose.In frequent years machine learning models were improved much when compared to old models and can be used to perform and solve several issues underlying in the society and in the field of medical applications.The machine learning models in the field of health care for many applications were trained with insufficient data sets for both training and testing which leads to inaccurate results which decreases the liability of the model since different data sets are available in online loads of data can be used to train a prediction model which yields higher accuracy value for certain algorithms that matches the inputs perfectly.

Most of the prediction models will predict the output for that particular instance and cannot be recorded or documented this will be a major issue in future as documentation plays a major role in data science in medical field.

3.1.2 SOLUTIONS:

- A disease prediction model should serve both as a prediction and suggestion model. Suggestion must be given to the patients based on their current medical data values and existing reports

- Sufficient amount of data sets must be trained to a model before implementing it as an application for protection

.

- The suggestion and reports of patients given by a machine learning model must be integrated with existing records and must be confidential.

- Appropriate machine learning algorithms must be used to increase the accuracy which is based on the type of dataset used.Data pre-processing plays a major role in machine learning and all the non-numeric data should be converted to a numeric data for easy evaluation.

3.2 PROPOSED SYSTEM:

We have developed a machine learning flask application to determine the outcome of diabetes in patients aged from 10 to 69 with appropriate algorithm usage with sufficient amount of medical data from data sets in government website.Python library packages cannot be easily integrated with the web application and requires a Application Programming Interface.This API can be used in Hypertext Markup Language with the help of json for integration.Java script cannot be used in back-end of HTML as it cannot import python library functions.

To overcome the documentation issues of patients record our ML Flask application is trained to predict the outcome of the model with the detailed analysis report for the particular patient.Python flask application allowed us to implement our machine learning model for production.The detailed health analysis report for a patient with a independent variable of Glucose, BMI, Insulin value ,Triceps thickness, Age, Diastolic Blood Pressure, Diabetes pedigree function and will determine

- Either the patient is either positive or negative for diabetes

- Type-I(Diabetes Mellitus) or Type-II(Diabetes Insipidus) if the patient is positive for diabetes

- Determination of pre-diabetes.

- Diastolic blood pressure range along with its causes,diets,medication,symptoms and measures to be taken for low,normal,high and very high blood pressure.

- Serum Insulin range with remedies for lower and higher insulin value along with a diet,medication,causes,symptoms and measures to be taken.

- BMI value which determines the overall body fat with diet and exercises to be followed.

- Patient name,hisdetails,TricepsValue,Diabetes pedigree function and Age

- This report can be downloaded for patients documentation purpose and can be used for future reference.

CHAPTER-4

SYSTEM REQUIREMENTS

4.1 HARDWARE REQUIREMENTS:

- A Computer with stable internet connection.

4.2 SOFTWARE REQUIREMENTS:

- Google Collab
- Anaconda 3.1.9(Jupyter Notebook).
- Diabetes patients dataset uploaded in CSV format.
- Sublime text editor.

4.3 TECHNOLOGIES USED:

- Python for Machine Learning.
- Python Flask.
- Hyper text Markup language.
- Cascading style sheets.
- Machine Learning.
- Random Forest Classifier Algorithm

4.4 OBJECT ORIENTED ANALYSIS:

4.4.1 Use Case Diagram:

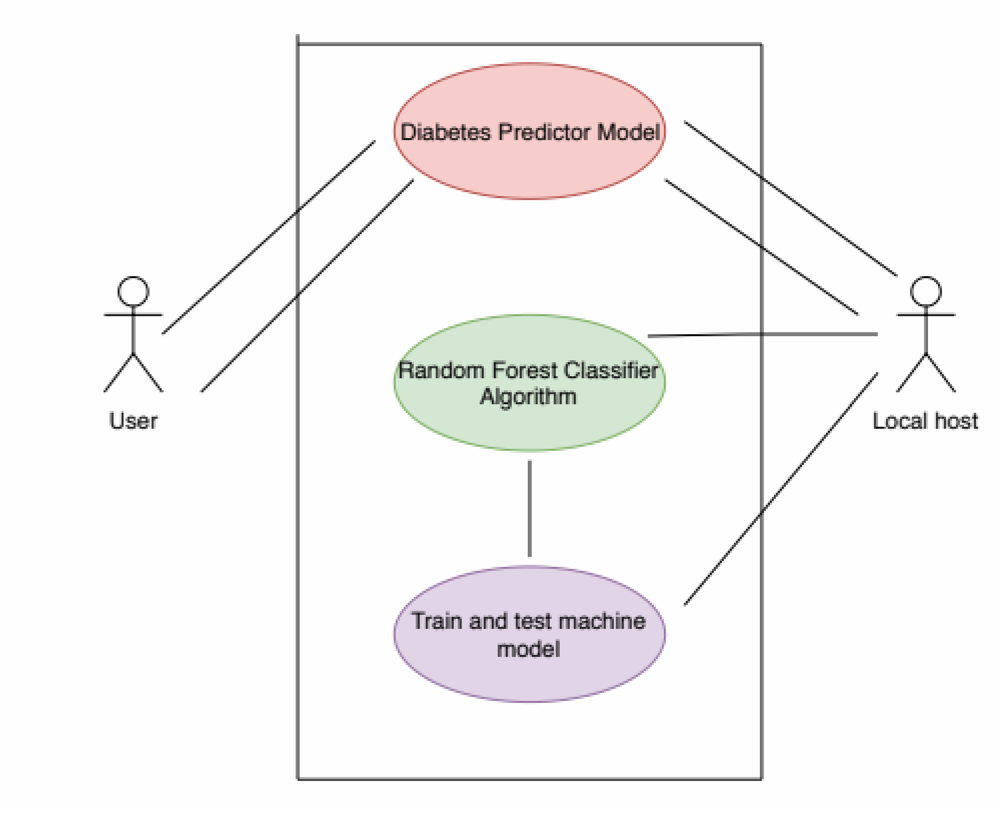

Diabetes Predictor Model
Random Forest Classifier Algorithm
Train and test machine model
User
Local host

Fig : 4.1 Use case Diagam

The purpose of the use case diagram is to demonstrate the different possible ways that a user or an external system interacts with your system. The use case diagram consists of the system which is represented using the rectangular boxes consisting of the use-cases which is represented using the ellipse labelled with the name of the use-case. It also consists of one or more actors which is represented using stick person icon. The user interacts with the local host with help of diabetes predictor model which uses Random Forest Classifier Algorithm to train and test the machine.

4.4.2 Activity Diagram

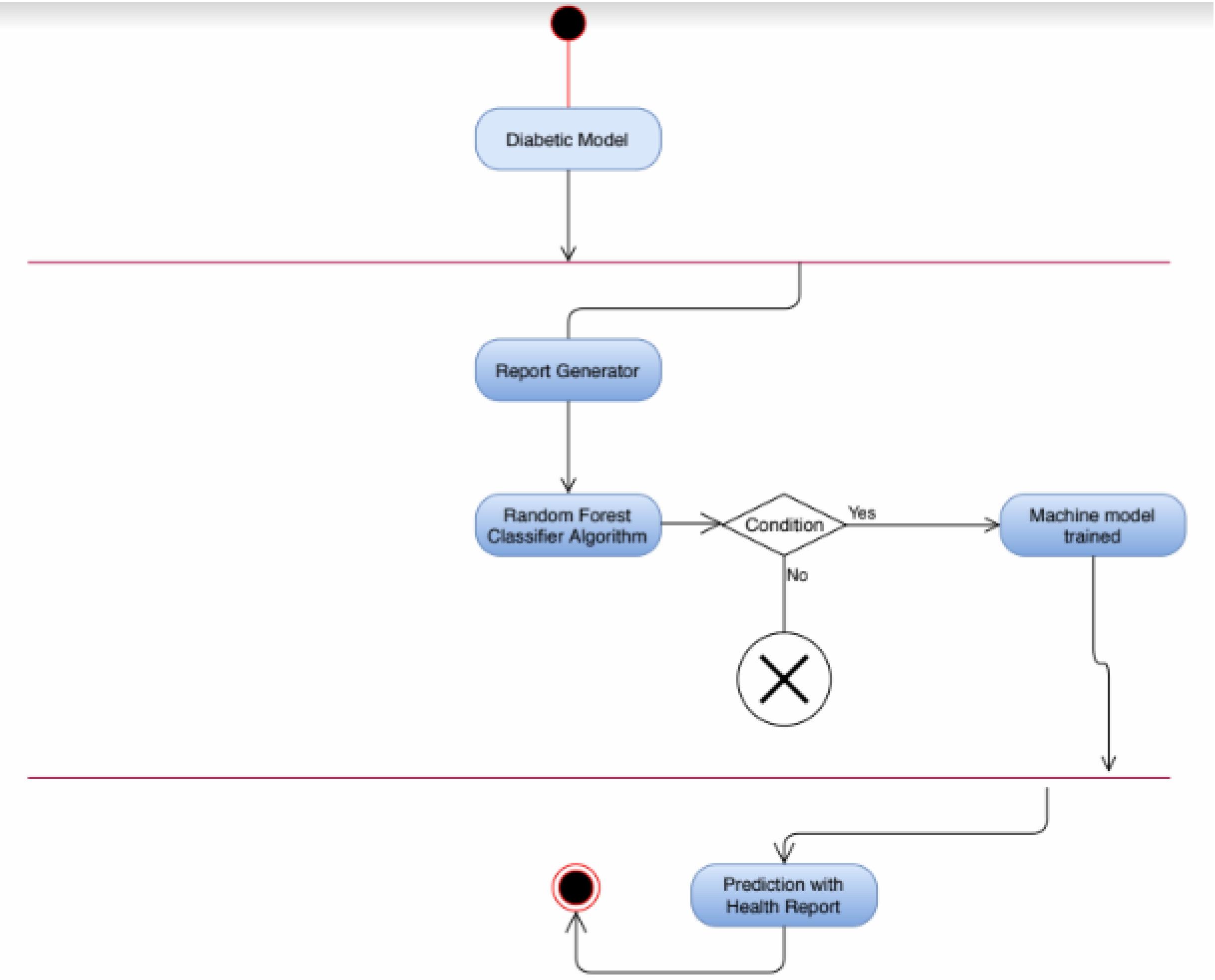


Fig : 4.2 Activity Diagram

Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. The control flow is drawn from one operation to another. This flow can be sequential, branched, or concurrent. When the control determines yes condition model is trained and report is generated, if not then it skips the process.

CHAPTER-5

DESIGN AND IMPLEMENTATION

5.1DESIGN:

The Machine learning model which allows to determine the diabetic condition of a patient can be implemented in a three stage process. First developing a machine learning model followed by integration with machine learning python along with local host and web technology at the end to give an complete machine learning flask application.

To overcome the documentation issues of patients record our ML Flask application is trained to predict the outcome of the model with the detailed analysis report for the particular patient.Python flask application allowed us to implement our machine learning model for production.The detailed health analysis report for a patient with a independent variable of Glucose,BMI,Insulin value,Triceps thickness,Age,Diastolic Blood Pressure,Diabetes pedigree function and will determine whether the patient is Positive or Negative for diabetes.The architecture for this model is given in fig 5.1

ARCHITECTURE:

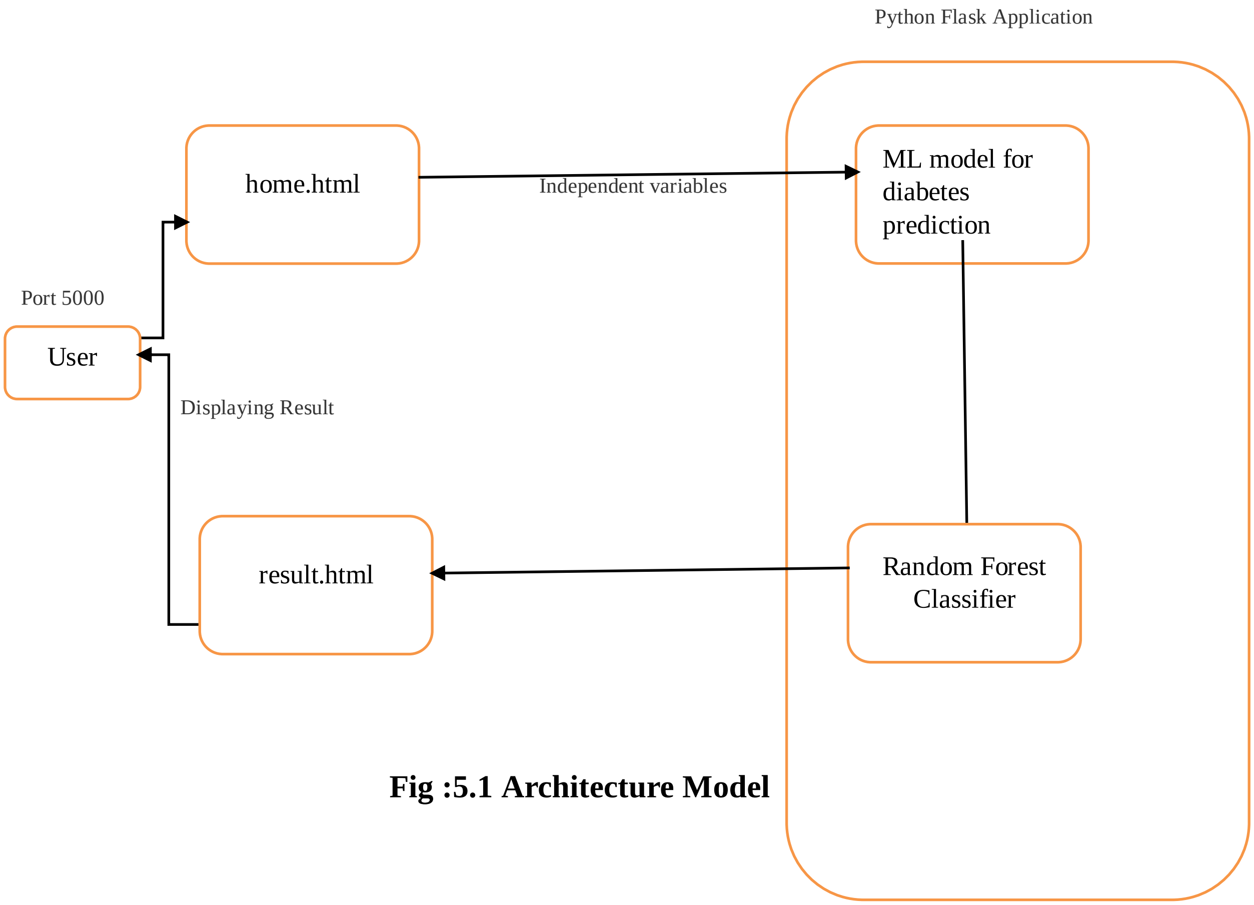


Fig :5.1 Architecture Model

5.2 PROBLEM STATEMENT:

" To develop a machine learning model ,a diabetic predictor which performs data pre-preprocessing an independent to determine the outcome of the diabetes along with its type and health advice report"

5.2.1 WHY THIS PROBLEM STATEMENT?

The python flask application which allows us to implement the machine learning model for production determines only the diabetes when coming to an existing model.In future replacement of humans in every field is gonna happen and each and every data in an instruction originating from an intelligence of human mind will be feeded in a machine as soon as possible which will drive the artificial intelligence to another level so the first step in upgrading the existing model can be approached in a way where the instruction about this particular diabetic topic(which has a consistent information) can be feeded to a machine to replace human presence.

5.3 MACHINE LEARNING PROCESS WITH LIBRARY IMPORT:

We started to implement and execute our ideas on machine learning model with help of collaboratoryservices,Googlecolab which offers 40GB RAM with python 2.7 back-end engine is really useful in handling data-sets with huge amount of data pandas,numpy and seaborn libraries are imported along with train_test_split,mean_squared_error,Accuracy_squared,metrics and required machine learning algorithms Joblib and pickle packages in python allows our machine learning model data with its parameter to be stored in physical drive which can be imported again while running the python application at the back-end of sublime text editor which merges HTML with python packages through machine learning. The import statement is given in fig A1

5.4 IMPORTING DATA SETS:

The datasets through which data for the problem statement is provided are passed as CSV files for python back-end engine to process.For google colab upload method is used to feed the machine learning model its required inputs and can be read using read_csvmethod.In this case it is diabetes.csv. The code to upload file to Google colab is given below fig A2

The above CSV file format is the real time data of 15000 persons with survey of Prediction of Diabetes and must contain all values in it.The necessary data that affects the dependent variable Diabetic are trained to model to predict variety of outcomes for each and every new observations. The dataset for my model is given in fig A3

5.5 DETERMINATION OF NULL VALUES:

Determination of null values is necessary in order to perform data pre-processing in data set and can be performed with the help of isna().This method returns true or false values where true means empty cell and false means occupied.sum() method along with isna() method will return a numerical value with the count of number of empty cell in the series there are two methods to eradicate this problem label encoding can be performed to replace values both in numerical and string format methods such as mean(numerical data),median and mode(for certain string data).One heart encoding is used to convert strings into binary numbers and this concept is known as degree of freedom.ROC curve and cross-validation are other methods to determine the performance of the model. The code to check for null values and their null values and their output is given in fig A4

In the above piece of code there are no null values in the provided data set series and can be pre-processed for further evaluation.

5.6 PLOTTING DATA:

The datasets can be plotted using scatter method and can be represented in the form of pie chart,bar graph and histogram ,the correlation between the input and the output is determined and similarities between the inputs are identified and redundancy along with the out-layer are removed.The below comparison of the independent variables age,BMI,pedigreefunction,blood pressure and triceps thickness represent the variation within the input and unwanted outlayer can easily identified with the help of this representation and eradicated. The plotting of data between various inputs is given in fig A5

5.7 HEATMAP DETERMINATION:

The correlation between the inputs are essential to determine the usability of the dataset and it must be minimum in order to get higher efficiency for that particular dataset and can be implemented using corr method in SNS.heatmap library.

In the diagram A6 the correlation between the diastolic blood pressure,age,serum,insulin and triceps thickness will determine whether the patient is suffering from diabetes or not there must be possible lowest value of correlation between two inputs for higher efficiency of datasets.

In the below diagram the correlation between the diabetic series and age series are quiet close which determines the output is closely linked with the age factor and must be eradicated since it does not affect the input it is not necessary to change the value of that particular dataset.

The comparison of various inputs is shown in fig A6

5.8 ACCURACY DETERMINATION:

The overall accuracy of the dataset is determined by fitting the output variable diabetic to the difference of sum of independent variable with the help of SMF.OLS library using summary method.The adjacent R-square value is taken as overall accuracy value for the dataset and the probability value for each series must be less than 0.05 which determines less correlation between the inputs.The 25%,50%,75%,standard deviation,overall mean for the dataset is useful to determine correlation between the inputs additionally,R-square value is not used to determine accuracy of the given dataset because it includes extra processing values which yields more accuracy than the original accuracy.

This OLS regression result uses least squares method to determine the standard error and probability value the co-variant value is non-robust therefore it is really tough to fit the points in a line with a equation of a straight line further machine learning algorithms must be implemented to fit a line in a point to get a best fit as a result. The accuracy for this model is shown in fig A7

5.9 TRAINING AND TESTING VALUES:

The values in the dataset is classified into dependent and independent variableswhich are further splitted into x and y for training and testing purpose.The x variable contains the set of series except the dependent variable diabetic whereas the y variable contains only the dependent variable diabetic.The x and y variable are further divided into x-train,x-test,y-train, y-test with the help of train test split method which has the parameter test size and random state test_size method determines the percentage of inputs with respect to output(For eg:0.3 determines 30% for testing the data whereas the other 70% is used for training).The random state method takes the state of training and testing inputs with specified test_size randomly which helps to predict outcome for new datasets. The code to seperate the dataset for both training and testing is shown in fig A8

NOTE: Patient Id and pregnancies columns do not contribute any value to predication of output and can be removed.

5.10 ALGORITHMS:

Both classification and regression algorithms can suit a problem statement and we implement a regression algorithms like linear regression,lasso regression and gradient boosting regression which yielded accuracy less than 40% and made as move towards classifier.

The yes or no model for diabetes prediction model prefers to move towards classification algorithms rather than regression algorithms where x varies with y .The major classification algorithms used in machine learning models are **Decision trees,Logistic,Know_your_nearest neighbour** and **Random forest.**For the provided dataset the logistic regression provides the accuracy of 76% for train values and 76%values for test values. Checking the accuracy for training data is shown in fig A9 and for testing data is shown in fig A10

The Know_your_nearest algorithm provides the accuracy of 82% for train data and 80% for test data.The accuracy of data using KNN algorithm is shown in fig A11

The n-neighbours value is identified by passing values with the set of range for that algorithm to determine the highest possible accuracy for that particular data.

5.10.1 RANDOM FOREST CLASSIFIER:

Random forest classifier algorithm uses trees that contains left and right nodes to process the inputs for better prediction than other classifier algorithms.A input passed to a random forest classifier is verified again and again by tuning the parameter n-estimators and max depth parameter which yields higher accuracy.

5.10.2 WHY RFC ALGORITHM:

The inputs given to Random Forest classifier will yield an output which is given a input to next random forest classifier.This process continues for n times and the exact output is obtained for verified training datasets.Similarly gradient boosting algorithm uses the same concept to achieve higher accuracy of machine learning model which falls under regressor category.It consists of combination of yes and no decision trees which makes RFC algorithm superior over others. The accuracy of data using Random Forest Classifier is shown in fig A12

The inputs given to Random Forest classifier will yield an output which is given a input to next random forest classifier.This process continues for n times and the exact output is obtained for verified training datasets.Similarly gradient boosting algorithm uses the same concept to achieve higher accuracy of machine learning model which falls under regressor category.It consists of combination of yes and no decision trees which makes RFC algorithm superior over others.

5.10.3 ERROR HISTOGRAM:

Outlayer is defined as a difference between the actual output and obtained output.Error histogram is determined by plt.hist method which shows the error difference between the actual input and obtained input with the help of RMSE(Root Mean Square Error) it determines the distance between the best fit line and the plotted point for train values,train points that are closer to the best fit line will be included along with the sum of other points to determine the RMSE value of the train and similar method is used to determine the RMSE value for test.

The standard bell-curve deviation says the nature of data set is higher.

Average point and lower in less and more ranges.The quality of the data set can be identified by using this curve.The samples must be taken from all the ranges to have a appropriate standard deviation. The histogram data is shown is fig A13

5.11 JOBLIB AND PICKLE :

The joblib and pickle package in python allows us to store the machine learning model in the local physical drive along the operating system so that the machine learning model need not to be executed each and every time.Joblib is used for sequence of numerical values and pickle is used for reading and writing of string in a file which can be imported later or whenever required.Bothjoblib and pickle are used in this model to write the machine learning model into a file(random.pkl) which can be implemented again in read mode with important parameters to run that machine learning model in local host.This will further allow the machine learning model to be converted as a flask application using python flask in the back-end

5.12 HOME PAGE FEATURES:

The homepage of Python Flask application consist of text boxes to collect information about the patients health record and this data is passed to the backend python programming where machine learning algorithms have already been implemented which makes the job much easier.

The homepage consist of Columns for obtaining Patients name,Patient ID no,Plasma Glucose value ,Diastolic Blood Pressure,Triceps Thickness ,Body Mass Index and Age. Placeholders are used to determine the data to be filled in the text box and each text box is referred with a name through which the objects can be thrown and caught easily that other backend request process.

When the submit button is clicked the form performs a action which involves using POST method to pass string data and this helps the python program at the back to gets the values through form.request method and pass it to the classifier and throw the results back to the result page. The html code for the homepage is shown in fig A14

The above image is a example for the homepage of Diab Preddo Application running in local host of production server at 127.0.00.5000. The design of front page is shown in fig A15

5.13 RESULT PAGE FEATURES:

The values after being processed by Random Forest classifier is thrown back to result page as string.The values in the form of string can only be seen or displayed in Result Page and cannot be manipulated.To perform further processing the only Possible way is to convert the string into an integer to perform operations and check conditions.

The following components will be displayed in home page after processing of data

- Either the patient is either positive or negative for diabetes

- Type-I(Diabetes Mellitus) or Type-II(Diabetes Insipidus) if the patient is positive for diabetes

- Determination of pre-diabetes.

- Diastolic blood pressure range along with its causes,diets,medication,symptoms and measures to be taken for low,normal,high and very high blood pressure.

- Serum Insulin range with remedies for lower and higher insulin value along with a diet,medication,causes,symptoms and measures to be taken.

- BMI value which determines the overall body fat with diet and exercises to be followed.

- Patient name,hisdetails,TricepsValue,Diabetes pedigree function and Age

- This report can be downloaded for patients documentation purpose and can be used for future reference.

5.14 PYTHON FLASK APPLICATION:

Machine Learning models can be applied in production with help of python flask concept.**Python flask converts the machine learning model into an API** which will help in production of applications related to web.In diab.py Python flask is implemented by importing flak,rendertemplate,url and request libraries.Render template method allows the navigation of home page to result page and vice versa.diab.route determines the path for local host and post method allows the application to request and response with request.methodfunction.In learning process df is a dataframe that has a CSV file of dataset.X contains the series except the output and Y contains output variable.

Train test split method allows us to split data values into train and test which will be useful while training the model with available inputs.Random Forest Classifier is used to perform data processing and the parameters are stored in random.pkl file which is imported back in the python program in read mode.the Observation is passed to the opt after performing the predict function and render template method throws the results back to the result page where it can be monitored and manipulated further.

Flask is a web-development framework , this is one of the most-popular module of python for web developers . It is easy to use , quick to learn , and also easy to create applications. It uses jinja2 template for html rendering . all though it is a micro-framework one can create medium level applications with this framework.

The Workflow for our model is shown in fig 5.2

Request.form method allows to get parameters for Name,Age,Diastolic blood pressure,Seruminsulin,Body Mass Index,Triceps thickness and passes the values to classifier.predict method .it determines the possibility of diabetes for a person and throws an output to result page as either 0 or 1 meaning 0 for Negative and 1 for Positive.When the prediction for input is 1 and age of patients is less than 24,it is more likely that the condition may intimate type 1 Mellitus whereas age greater than or equal to 24 indicates type 2 Mellitus condition.The average weight for BMI is 18.8 to 24.4 and normal blood pressure ranges from 80 to 120.

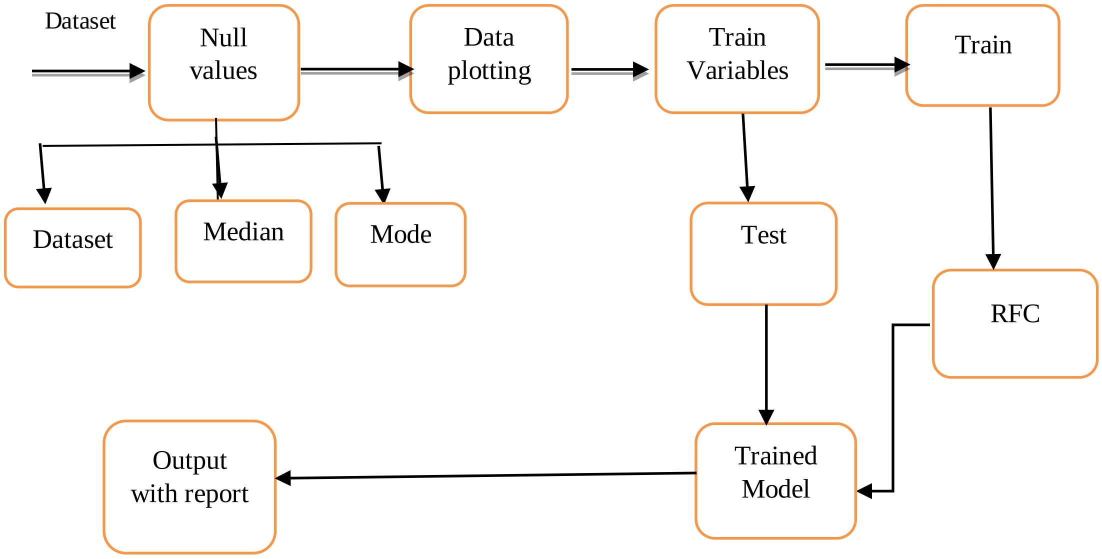


Fig 5.2 Workflow Model

The piece of code shown in fig A21 determines the python representation of machine learning model with diab.run method to run the application in command prompt. Random state is used to train and test split the data randomly and test size determines the size of test data to be taken.

5.15 IMPLEMENTATION USING LOCAL HOST:

Command prompt is set to path which is a folder in Desktop called ML APP.The change directory command is used to change the path in command prompt. The path setting is shown in fig:A22

The python program diab.py is executed and local host with server at port 5000 is allocated in system is shown in fig: A23

The Debugger pin will be active and Server will be running on http://127.0.0.1:5000/ which will take the control to the homepage (predictor.html) of Python Flask Application.

5.16 REPORT GENERATION:

The report generated by the Diabetes model can be saved to the system and can be viewed anytime.Further enhancements can be made by Report generation in PDF format using PHP and hosting the application in cloud. The below report determines the status of patient abc with Plasma glucose 170 ,Blood pressure 180,Triceps value 25,Serum insulin value 153,Body Mass Index 43 and Age 44 with Tests to be taken ,diet suggestion and diabetes type with medication.

5.17 MODULE 1:

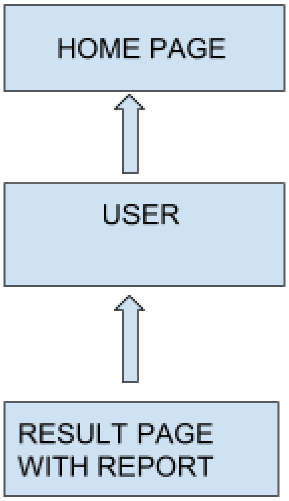


Fig : 5.3 Architecture For Module 1

User has access to home page where the patients data is feeded and obtains the result with report through the result page. The data pre-processing machine learning algorithm is implemented at back-end of the application.

5.18 MODULE 2:

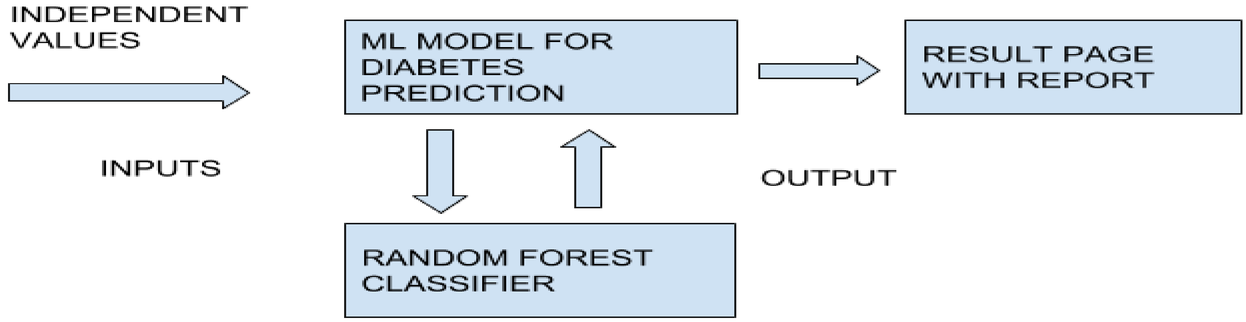


Fig 5.4: Architecture Of Module 2

The independent values given by a patient will be passed to machine learning model for diabetes prediction and Random Forest Classifier algorithm gets the values determines the result whether it is positive or negative and throws back the value to the machine learning model which in turn passes the result to the result page with the report for user to view.

CHAPTER-6

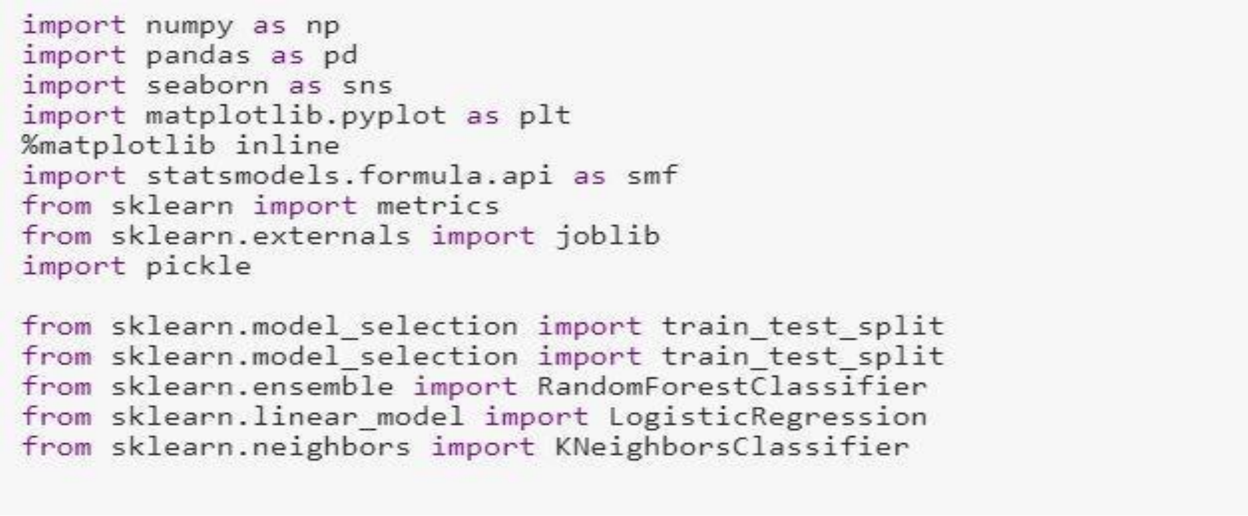
CONCLUSION AND FUTURE ENHANCEMENT :

Thus the diabetes prediction model to determine the presence of diabetes mellitus or diabetes insipidus is successfully implemented with an automatic report generation mechanism for patients documentation purpose. ML Flask application is trained to predict the outcome of the model with the detailed analysis report for the particular patient.Python flask application allowed us to implement our machine learning model for production.Machine learning applications must be implemented in near future and will replace humans which will reduce the workload and artificial intelligence will play a key role in future.

The generated health report for a patient can be difficult for him to carry or understand the specification and medical technical terms it contains so an assistant or a bot to advice the patients will be required as most of the technologies and human work would be digitalized in future.

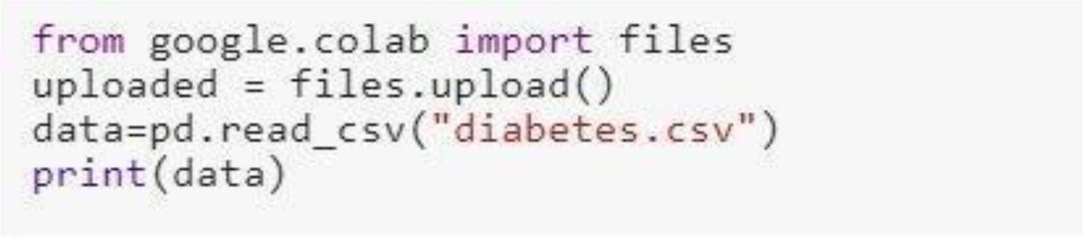
APPENDICES:

APPENDIX 1:



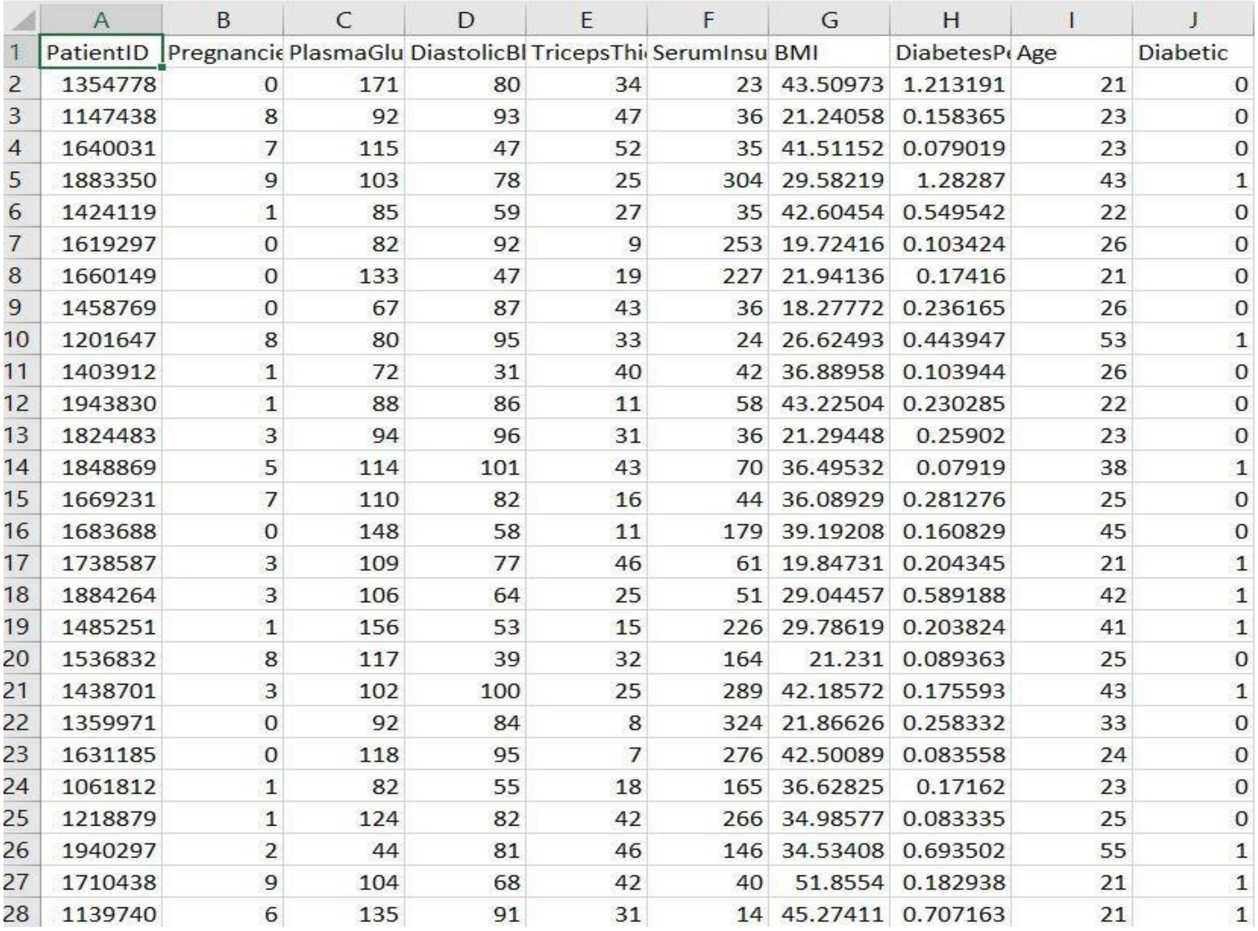
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import statsmodels.formula.api as smf
from sklearn import metrics
from sklearn.externals import joblib
import pickle

from sklearn.model_selection import train_test_split
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
```

Fig: A1 Importing Packages



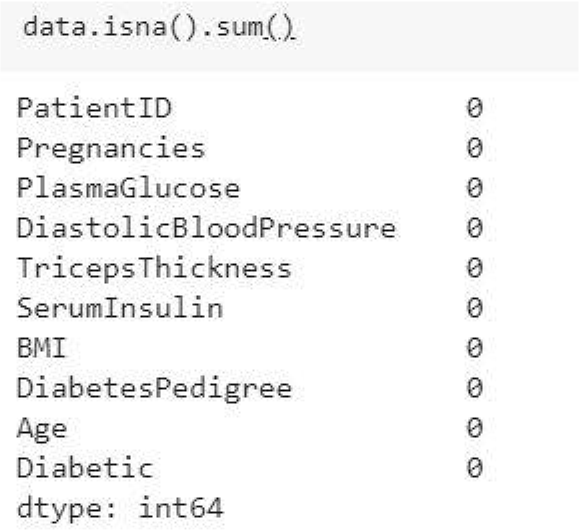
```
from google.colab import files
uploaded = files.upload()
data=pd.read_csv("diabetes.csv")
print(data)
```

Fig: A2 Uploading File to Google Collab



	A	B	C	D	E	F	G	H	I	J
1	PatientID	Pregnancie	PlasmaGlu	DiastolicBl	TricepsThi	SerumInsu	BMI	DiabetesP	Age	Diabetic
2	1354778	0	171	80	34	23	43.50973	1.213191	21	0
3	1147438	8	92	93	47	36	21.24058	0.158365	23	0
4	1640031	7	115	47	52	35	41.51152	0.079019	23	0
5	1883350	9	103	78	25	304	29.58219	1.28287	43	1
6	1424119	1	85	59	27	35	42.60454	0.549542	22	0
7	1619297	0	82	92	9	253	19.72416	0.103424	26	0
8	1660149	0	133	47	19	227	21.94136	0.17416	21	0
9	1458769	0	67	87	43	36	18.27772	0.236165	26	0
10	1201647	8	80	95	33	24	26.62493	0.443947	53	1
11	1403912	1	72	31	40	42	36.88958	0.103944	26	0
12	1943830	1	88	86	11	58	43.22504	0.230285	22	0
13	1824483	3	94	96	31	36	21.29448	0.25902	23	0
14	1848869	5	114	101	43	70	36.49532	0.07919	38	1
15	1669231	7	110	82	16	44	36.08929	0.281276	25	0
16	1683688	0	148	58	11	179	39.19208	0.160829	45	0
17	1738587	3	109	77	46	61	19.84731	0.204345	21	1
18	1884264	3	106	64	25	51	29.04457	0.589188	42	1
19	1485251	1	156	53	15	226	29.78619	0.203824	41	1
20	1536832	8	117	39	32	164	21.231	0.089363	25	0
21	1438701	3	102	100	25	289	42.18572	0.175593	43	1
22	1359971	0	92	84	8	324	21.86626	0.258332	33	0
23	1631185	0	118	95	7	276	42.50089	0.083558	24	0
24	1061812	1	82	55	18	165	36.62825	0.17162	23	0
25	1218879	1	124	82	42	266	34.98577	0.083335	25	0
26	1940297	2	44	81	46	146	34.53408	0.693502	55	1
27	1710438	9	104	68	42	40	51.8554	0.182938	21	1
28	1139740	6	135	91	31	14	45.27411	0.707163	21	1

Fig : A3 Dataset



```
data.isna().sum()
```

```
PatientID                 0
Pregnancies               0
PlasmaGlucose             0
DiastolicBloodPressure    0
TricepsThickness          0
SerumInsulin              0
BMI                       0
DiabetesPedigree          0
Age                       0
Diabetic                  0
dtype: int64
```

Fig: A4 Checking For Null Values

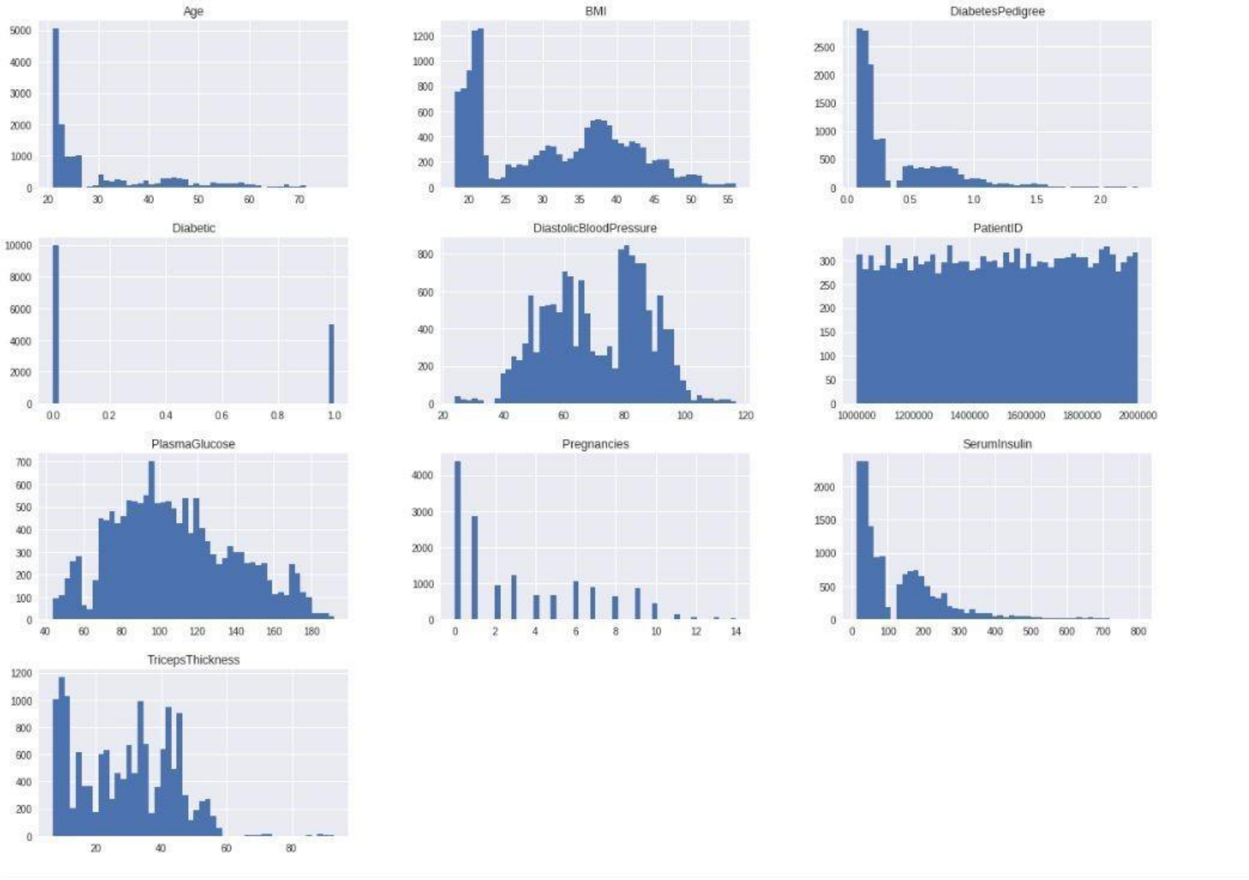


Fig: A5 Comparison Between Various Inputs

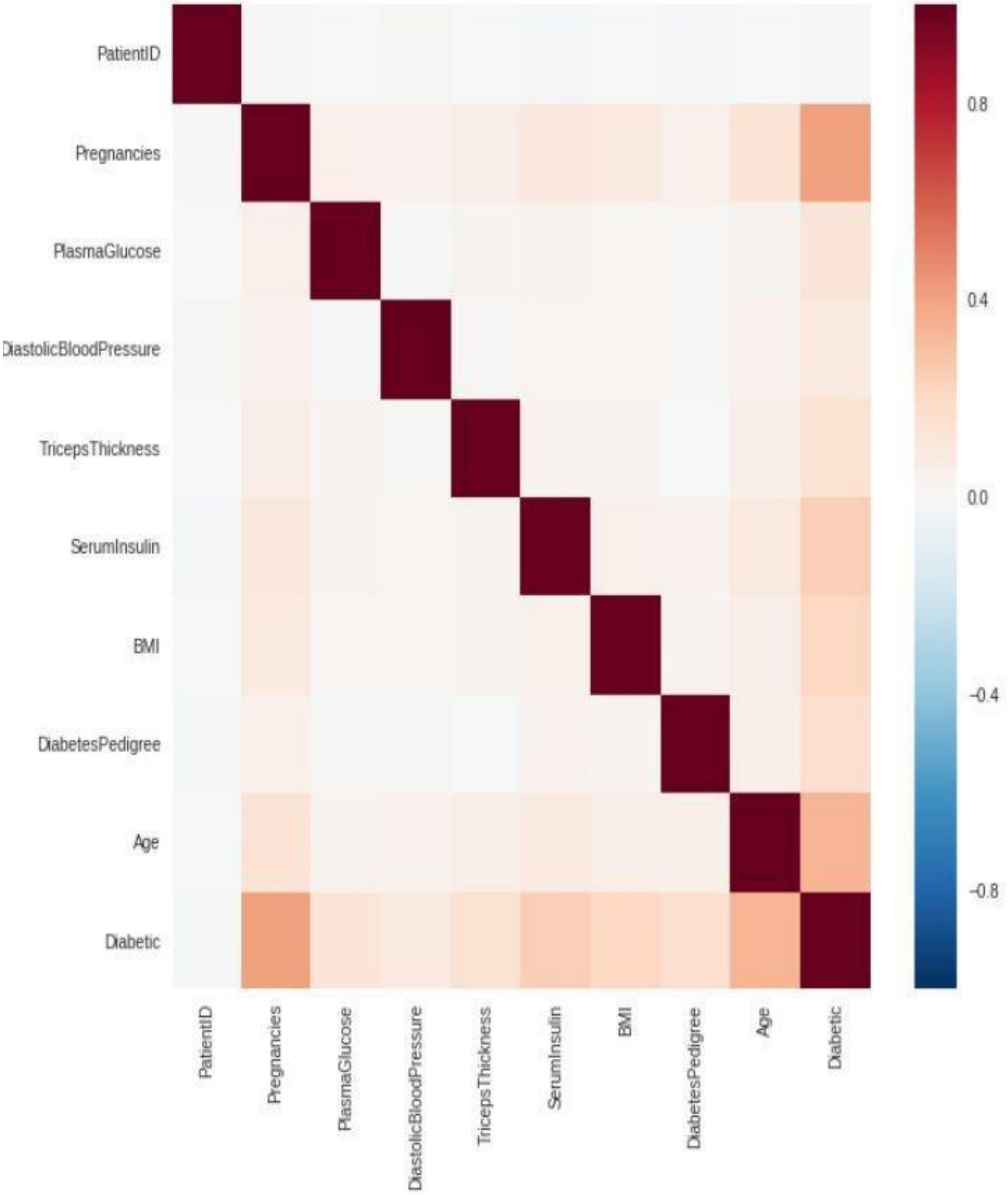
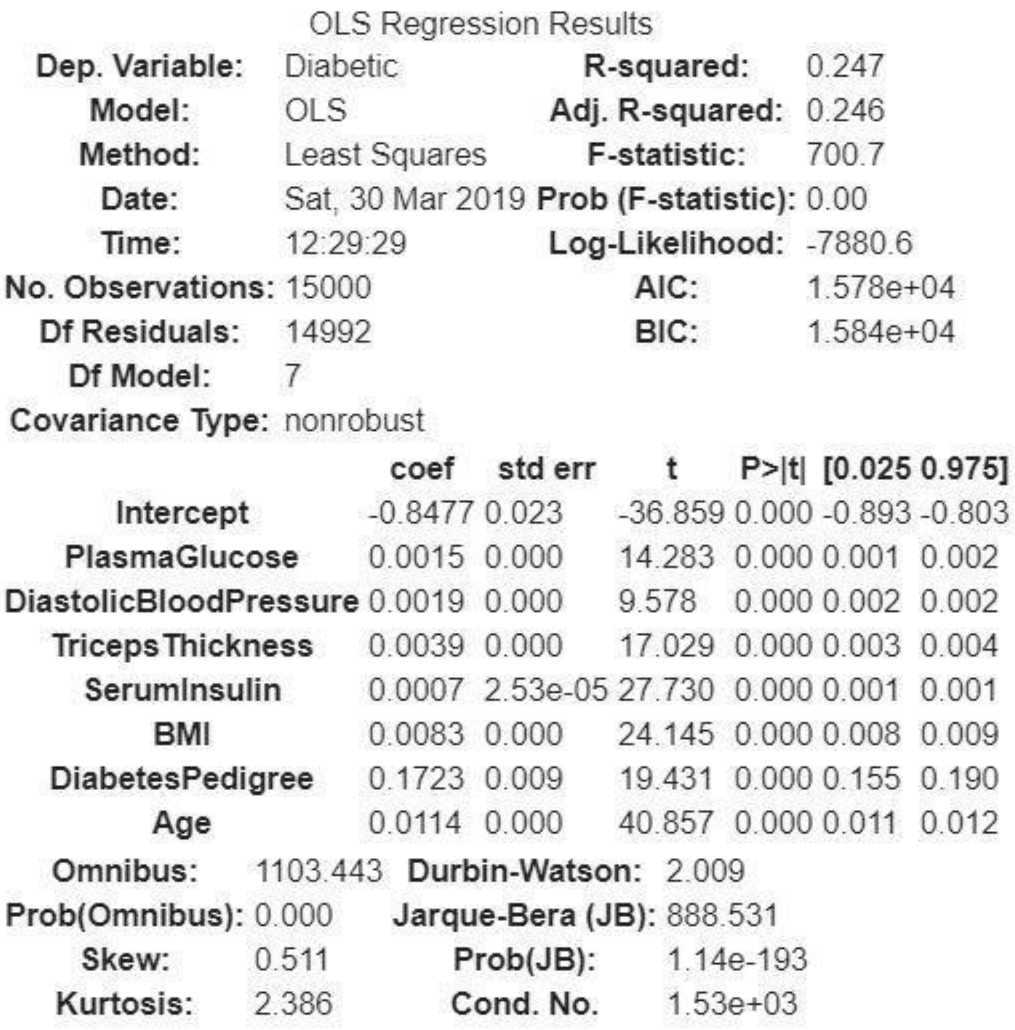
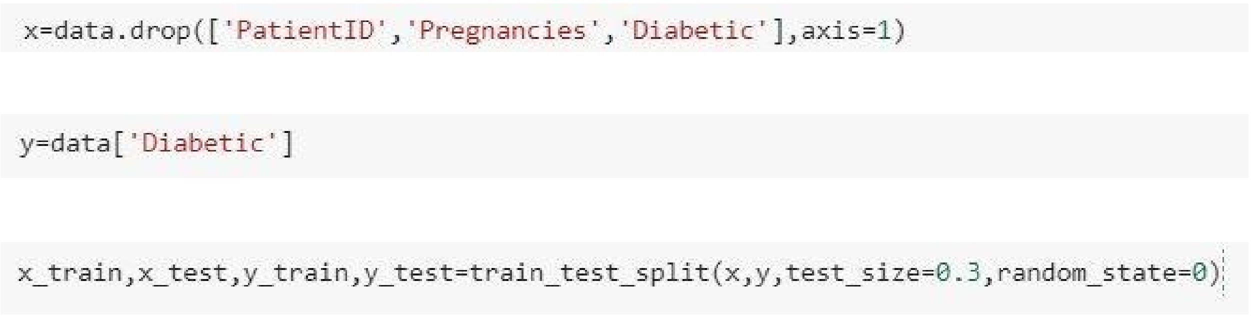


Fig: A6 Comparison Of Inputs Using Heat Map



```
                        OLS Regression Results
    Dep. Variable:   Diabetic              R-squared:          0.247
           Model:    OLS               Adj. R-squared:         0.246
          Method:    Least Squares          F-statistic:       700.7
            Date:    Sat, 30 Mar 2019 Prob (F-statistic): 0.00
            Time:    12:29:29          Log-Likelihood:     -7880.6
No. Observations:    15000                      AIC:           1.578e+04
    Df Residuals:    14992                      BIC:           1.584e+04
        Df Model:    7
 Covariance Type:    nonrobust
                              coef    std err     t       P>|t|  [0.025  0.975]
       Intercept             -0.8477  0.023      -36.859  0.000  -0.893  -0.803
     PlasmaGlucose            0.0015  0.000       14.283  0.000   0.001   0.002
DiastolicBloodPressure        0.0019  0.000       9.578   0.000   0.002   0.002
    TricepsThickness          0.0039  0.000       17.029  0.000   0.003   0.004
      SerumInsulin            0.0007  2.53e-05    27.730  0.000   0.001   0.001
          BMI                 0.0083  0.000       24.145  0.000   0.008   0.009
    DiabetesPedigree          0.1723  0.009       19.431  0.000   0.155   0.190
          Age                 0.0114  0.000       40.857  0.000   0.011   0.012
     Omnibus:       1103.443  Durbin-Watson:      2.009
Prob(Omnibus):      0.000     Jarque-Bera (JB):   888.531
        Skew:       0.511          Prob(JB):      1.14e-193
    Kurtosis:       2.386          Cond. No.      1.53e+03
```

Fig: A7 Accuracy Of model

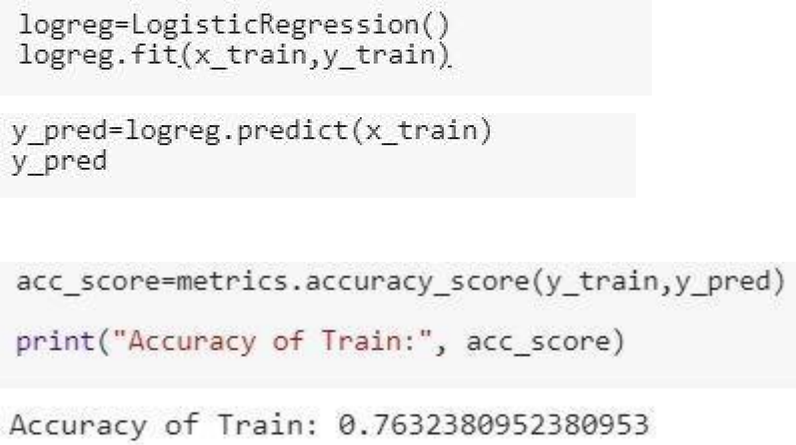


```
x=data.drop(['PatientID','Pregnancies','Diabetic'],axis=1)
```

```
y=data['Diabetic']
```

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=0)
```

Fig: A8 Seperating Dataset For Training And Testing



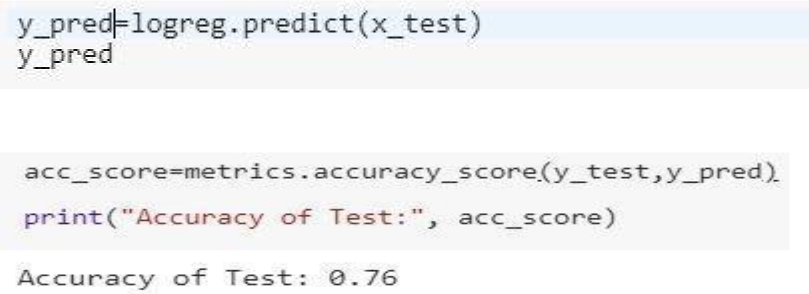
```
logreg=LogisticRegression()
logreg.fit(x_train,y_train)
```

```
y_pred=logreg.predict(x_train)
y_pred
```

```
acc_score=metrics.accuracy_score(y_train,y_pred)

print("Accuracy of Train:", acc_score)
```

```
Accuracy of Train: 0.7632380952380953
```

Fig: A9 Accuracy For Training Data

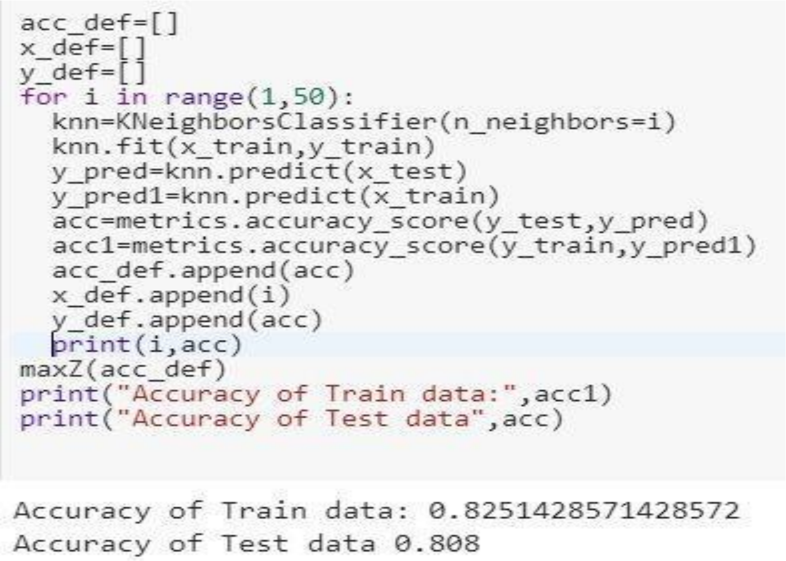


```
y_pred=logreg.predict(x_test)
y_pred
```

```
acc_score=metrics.accuracy_score(y_test,y_pred)

print("Accuracy of Test:", acc_score)
```

```
Accuracy of Test: 0.76
```

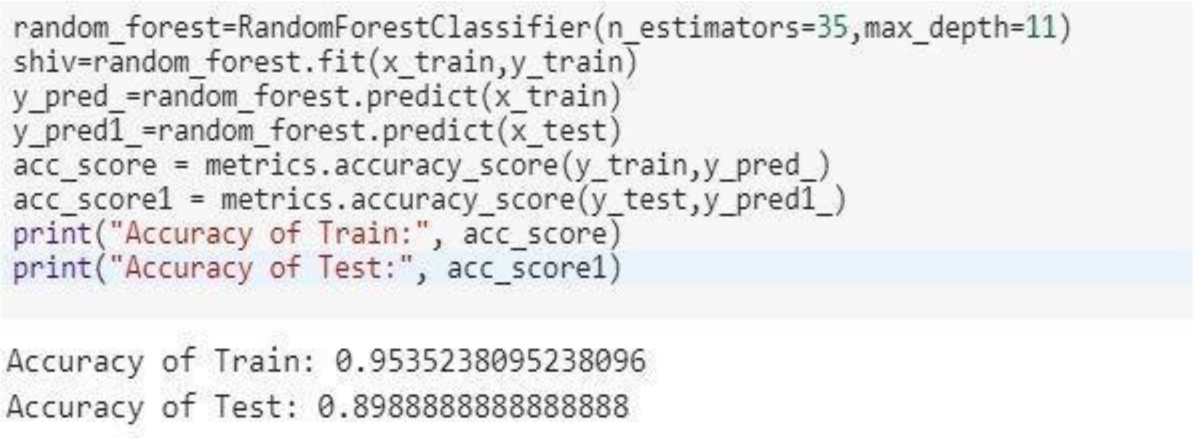
Fig: A10 Accuracy For Testing Data



```
acc_def=[]
x_def=[]
y_def=[]
for i in range(1,50):
  knn=KNeighborsClassifier(n_neighbors=i)
  knn.fit(x_train,y_train)
  y_pred=knn.predict(x_test)
  y_pred1=knn.predict(x_train)
  acc=metrics.accuracy_score(y_test,y_pred)
  acc1=metrics.accuracy_score(y_train,y_pred1)
  acc_def.append(acc)
  x_def.append(i)
  y_def.append(acc)
  print(i,acc)
maxZ(acc_def)
print("Accuracy of Train data:",acc1)
print("Accuracy of Test data",acc)
```

```
Accuracy of Train data: 0.8251428571428572
Accuracy of Test data 0.808
```

Fig: A11 Accuracy Of Data Using KNN Algorithm



```
random_forest=RandomForestClassifier(n_estimators=35,max_depth=11)
shiv=random_forest.fit(x_train,y_train)
y_pred_=random_forest.predict(x_train)
y_pred1_=random_forest.predict(x_test)
acc_score = metrics.accuracy_score(y_train,y_pred_)
acc_score1 = metrics.accuracy_score(y_test,y_pred1_)
print("Accuracy of Train:", acc_score)
print("Accuracy of Test:", acc_score1)
```

```
Accuracy of Train: 0.9535238095238096
Accuracy of Test: 0.8988888888888888
```

Fig: A12 Accuracy Of Data Using Random forest Classifier

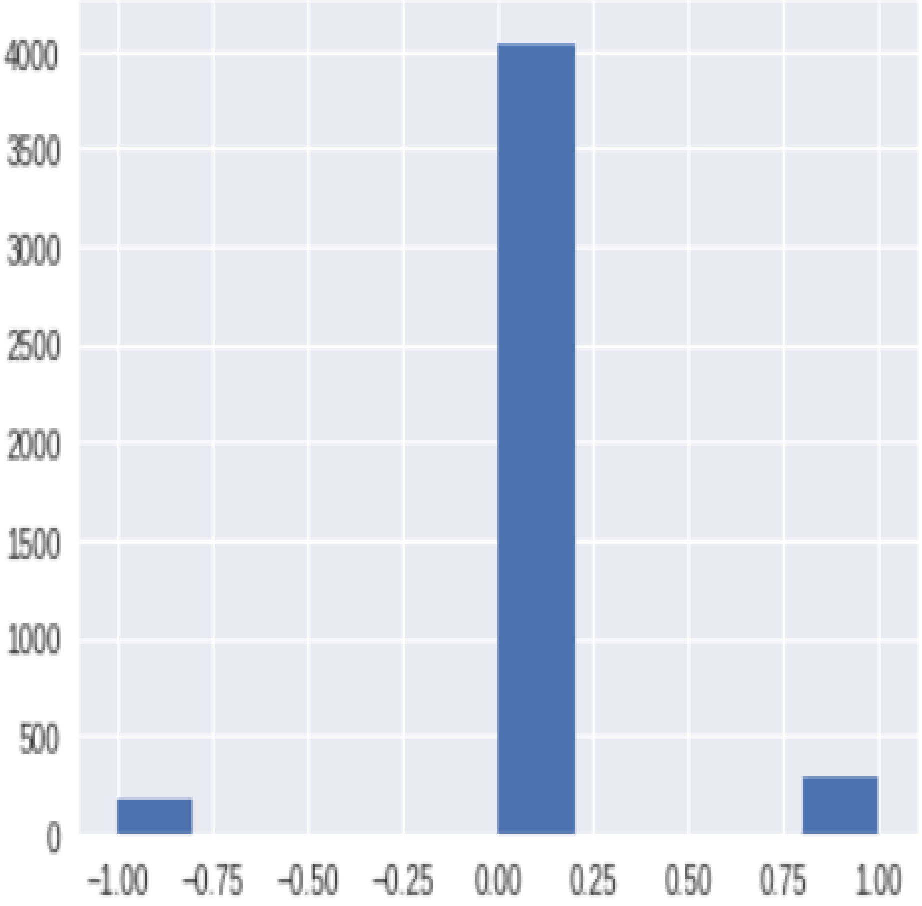
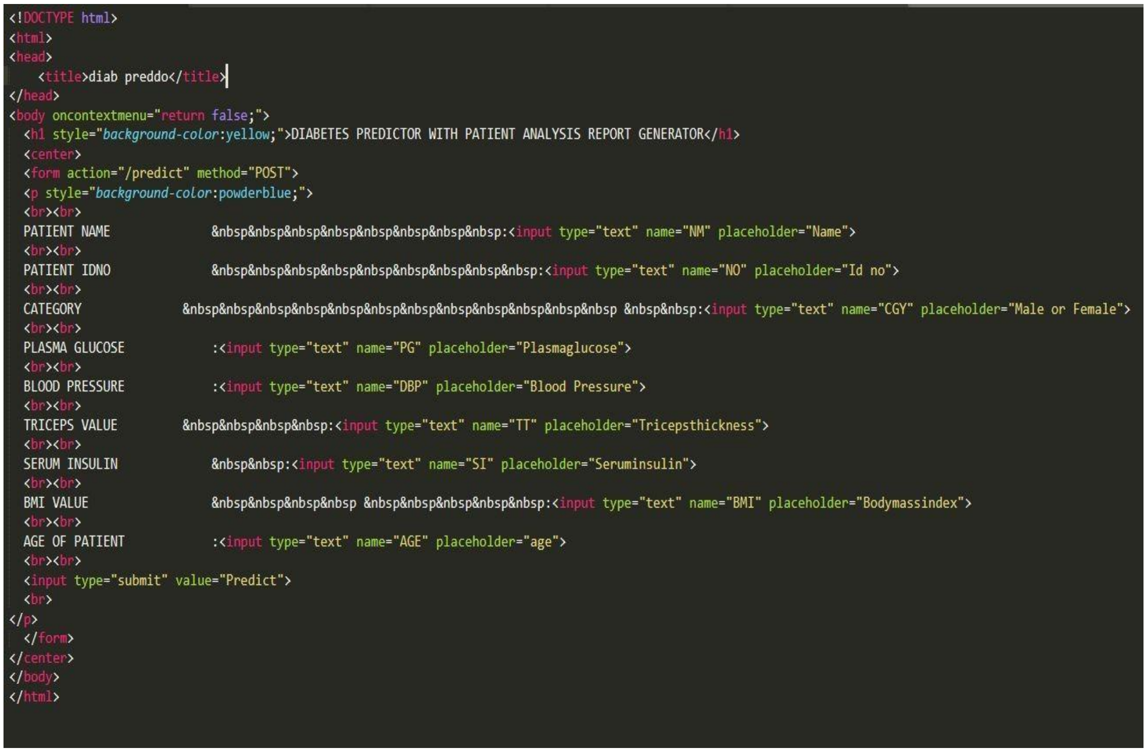
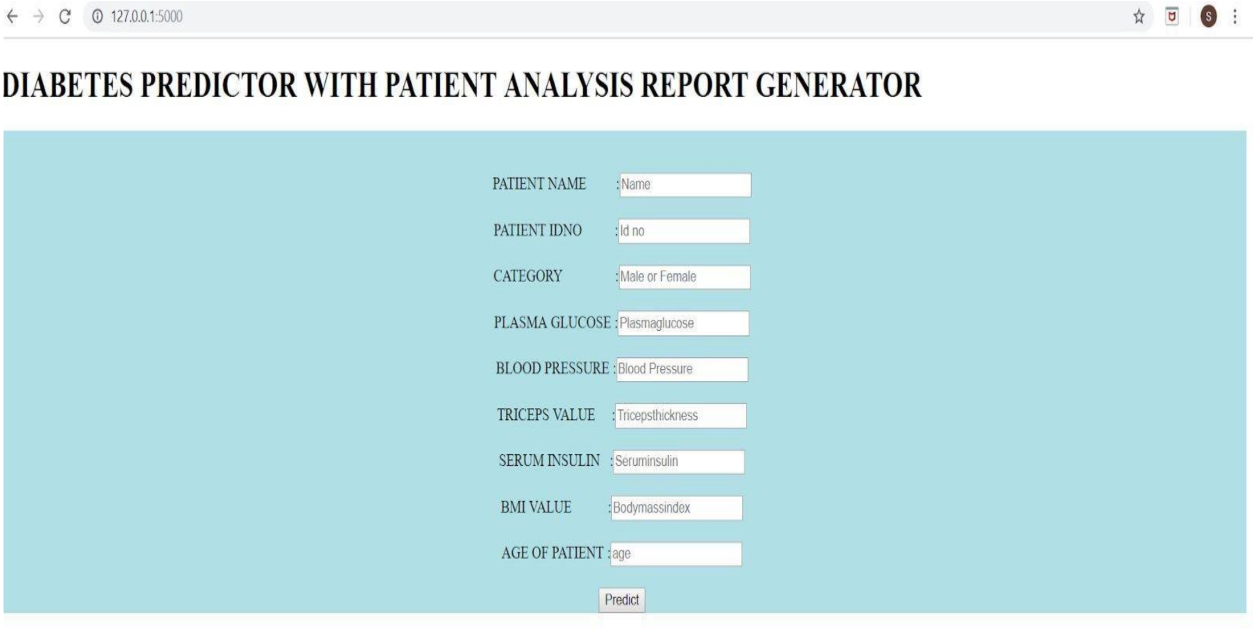


Fig: A13 Representation Of Data Using Histogram



```
<!DOCTYPE html>
<html>
<head>
    <title>diab preddo</title>
</head>
<body oncontextmenu="return false;">
  <h1 style="background-color:yellow;">DIABETES PREDICTOR WITH PATIENT ANALYSIS REPORT GENERATOR</h1>
  <center>
  <form action="/predict" method="POST">
  <p style="background-color:powderblue;">
  <br><br>
  PATIENT NAME              &nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp:<input type="text" name="NM" placeholder="Name">
  <br><br>
  PATIENT IDNO              &nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp:<input type="text" name="NO" placeholder="Id no">
  <br><br>
  CATEGORY              &nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp &nbsp&nbsp:<input type="text" name="CGY" placeholder="Male or Female">
  <br><br>
  PLASMA GLUCOSE            :<input type="text" name="PG" placeholder="Plasmaglucose">
  <br><br>
  BLOOD PRESSURE            :<input type="text" name="DBP" placeholder="Blood Pressure">
  <br><br>
  TRICEPS VALUE         &nbsp&nbsp&nbsp&nbsp:<input type="text" name="TT" placeholder="Tricepsthickness">
  <br><br>
  SERUM INSULIN             &nbsp&nbsp:<input type="text" name="SI" placeholder="Seruminsulin">
  <br><br>
  BMI VALUE                 &nbsp&nbsp&nbsp&nbsp &nbsp&nbsp&nbsp&nbsp&nbsp:<input type="text" name="BMI" placeholder="Bodymassindex">
  <br><br>
  AGE OF PATIENT            :<input type="text" name="AGE" placeholder="age">
  <br><br>
  <input type="submit" value="Predict">
  <br>
</p>
  </form>
</center>
</body>
</html>
```

Fig: A14 Source Code For Home Page



DIABETES PREDICTOR WITH PATIENT ANALYSIS REPORT GENERATOR

PATIENT NAME :Name

PATIENT IDNO :Id no

CATEGORY :Male or Female

PLASMA GLUCOSE :Plasmaglucose

BLOOD PRESSURE :Blood Pressure

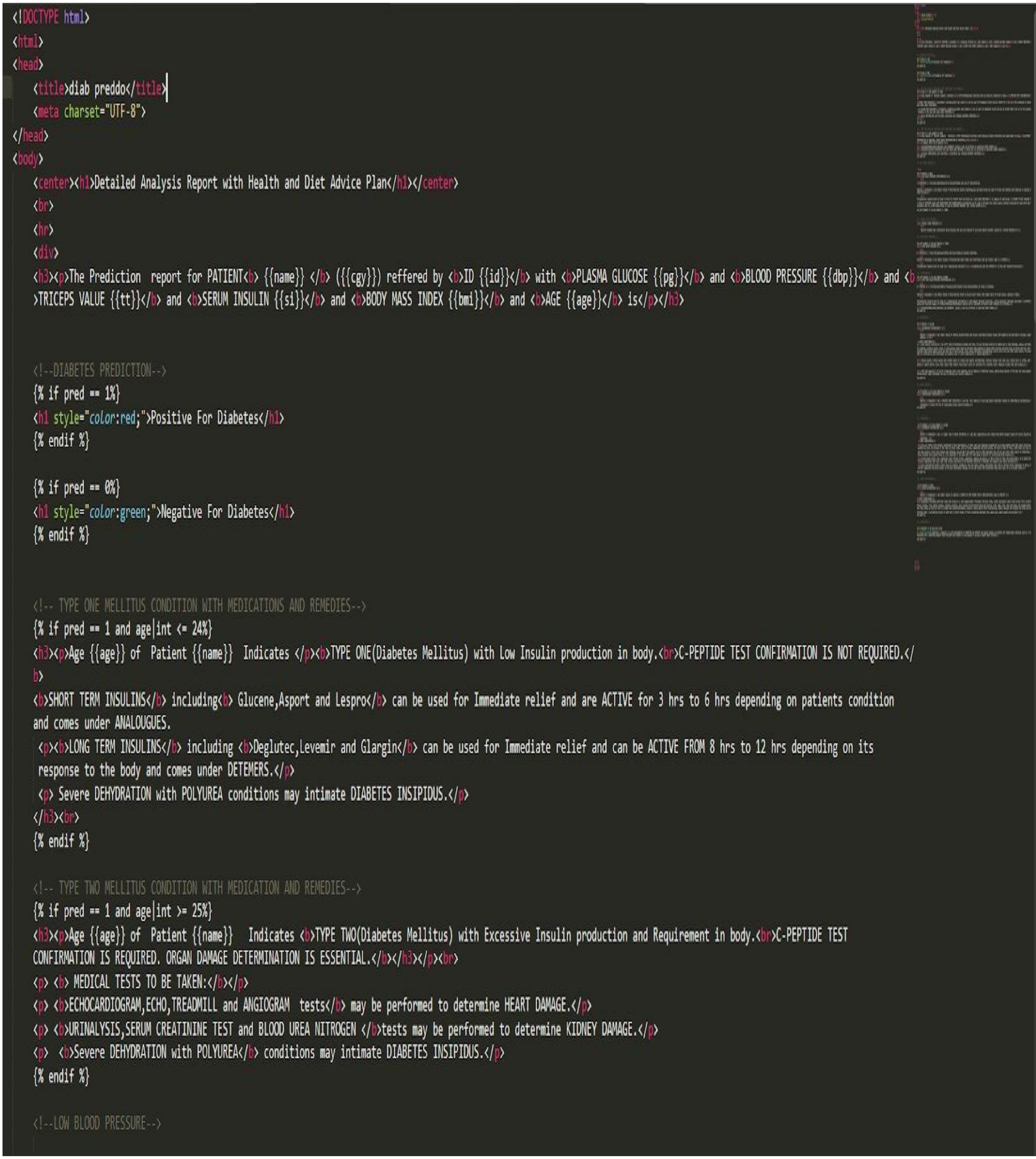
TRICEPS VALUE :Tricepsthickness

SERUM INSULIN :Seruminsulin

BMI VALUE :Bodymassindex

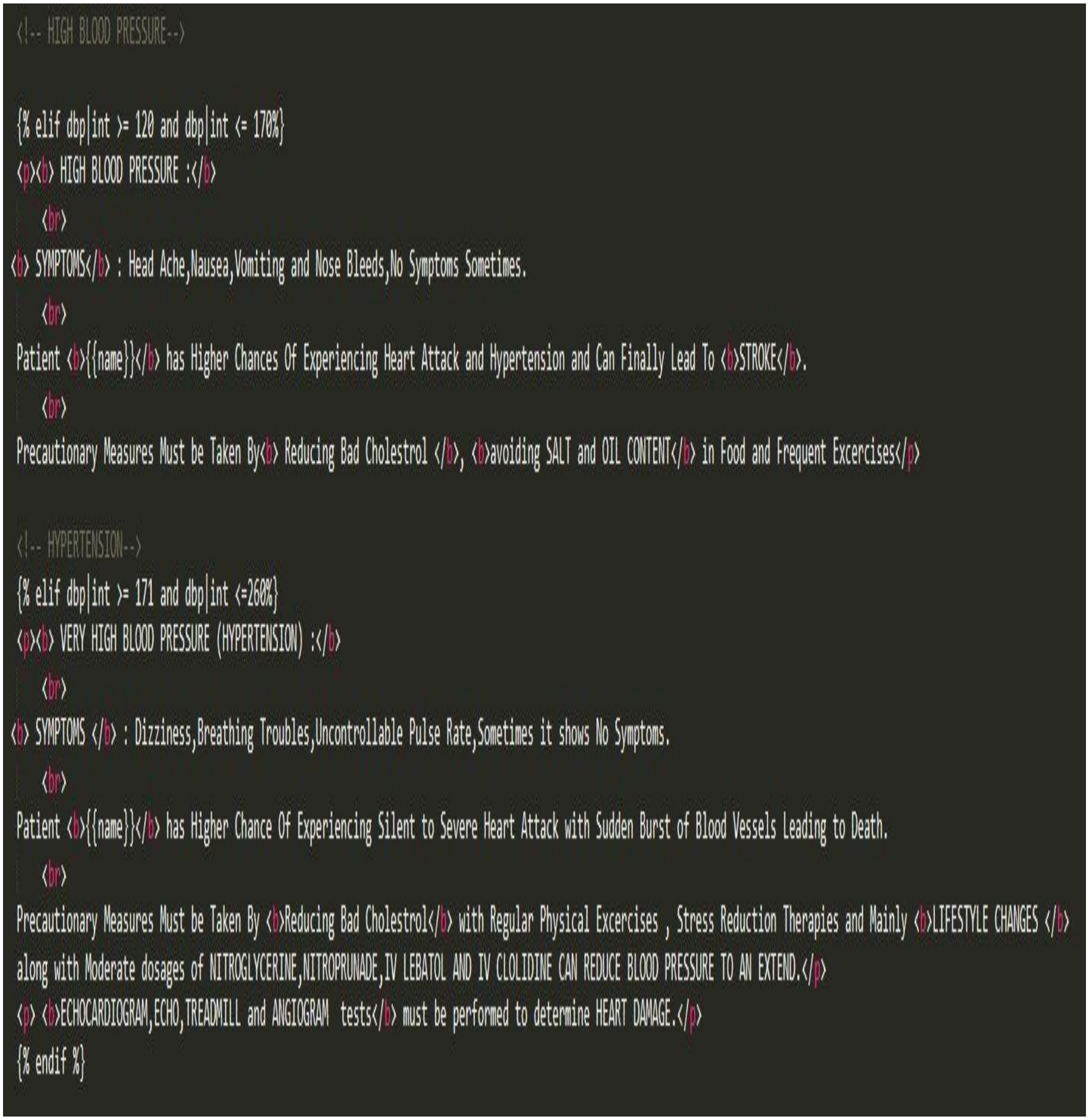
AGE OF PATIENT :age

Fig: A15 Design Of Front-end



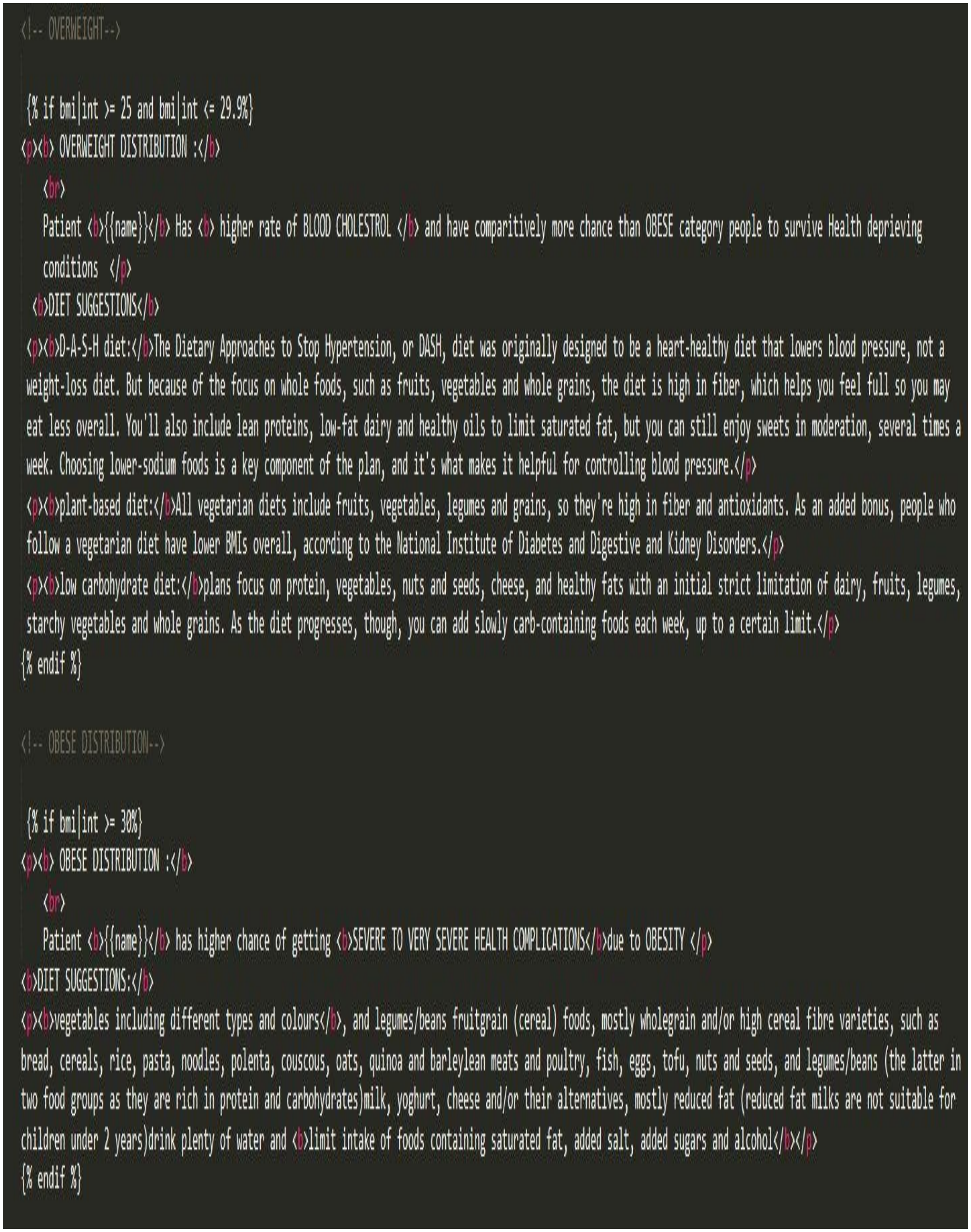
```
<!DOCTYPE html>
<html>
<head>
    <title>diab preddo</title>
    <meta charset="UTF-8">
</head>
<body>
    <center><h1>Detailed Analysis Report with Health and Diet Advice Plan</h1></center>
    <br>
    <hr>
    <div>
    <h3><p>The Prediction  report for PATIENT<b> {{name}} </b> ({{cgy}}) reffered by <b>ID {{id}}</b> with <b>PLASMA GLUCOSE {{pg}}</b> and <b>BLOOD PRESSURE {{dbp}}</b> and <b
    >TRICEPS VALUE {{tt}}</b> and <b>SERUM INSULIN {{si}}</b> and <b>BODY MASS INDEX {{bmi}}</b> and <b>AGE {{age}}</b> is</p></h3>


    <!--DIABETES PREDICTION-->
    {% if pred == 1%}
    <h1 style="color:red;">Positive For Diabetes</h1>
    {% endif %}

    {% if pred == 0%}
    <h1 style="color:green;">Negative For Diabetes</h1>
    {% endif %}


    <!-- TYPE ONE MELLITUS CONDITION WITH MEDICATIONS AND REMEDIES-->
    {% if pred == 1 and age|int <= 24%}
    <h3><p>Age {{age}} of  Patient {{name}}  Indicates </p><b>TYPE ONE(Diabetes Mellitus) with Low Insulin production in body.<br>C-PEPTIDE TEST CONFIRMATION IS NOT REQUIRED.</
    b>
    <b>SHORT TERM INSULINS</b> including<b> Glucene,Asport and Lespro</b> can be used for Immediate relief and are ACTIVE for 3 hrs to 6 hrs depending on patients condition
    and comes under ANALOUGUES.
     <p><b>LONG TERM INSULINS</b> including <b>Deglutec,Levemir and Glargin</b> can be used for Immediate relief and can be ACTIVE FROM 8 hrs to 12 hrs depending on its
     response to the body and comes under DETEMERS.</p>
     <p> Severe DEHYDRATION with POLYUREA conditions may intimate DIABETES INSIPIDUS.</p>
    </h3><br>
    {% endif %}

    <!-- TYPE TWO MELLITUS CONDITION WITH MEDICATION AND REMEDIES-->
    {% if pred == 1 and age|int >= 25%}
    <h3><p>Age {{age}} of  Patient {{name}}   Indicates <b>TYPE TWO(Diabetes Mellitus) with Excessive Insulin production and Requirement in body.<br>C-PEPTIDE TEST
    CONFIRMATION IS REQUIRED. ORGAN DAMAGE DETERMINATION IS ESSENTIAL.</b></h3></p><br>
    <p> <b> MEDICAL TESTS TO BE TAKEN:</b></p>
    <p> <b>ECHOCARDIOGRAM,ECHO,TREADMILL and ANGIOGRAM  tests</b> may be performed to determine HEART DAMAGE.</p>
    <p> <b>URINALYSIS,SERUM CREATININE TEST and BLOOD UREA NITROGEN </b>tests may be performed to determine KIDNEY DAMAGE.</p>
    <p>  <b>Severe DEHYDRATION with POLYUREA</b> conditions may intimate DIABETES INSIPIDUS.</p>
    {% endif %}

    <!--LOW BLOOD PRESSURE-->
```

Fig: A16 Python code For Diabetes Type Identification



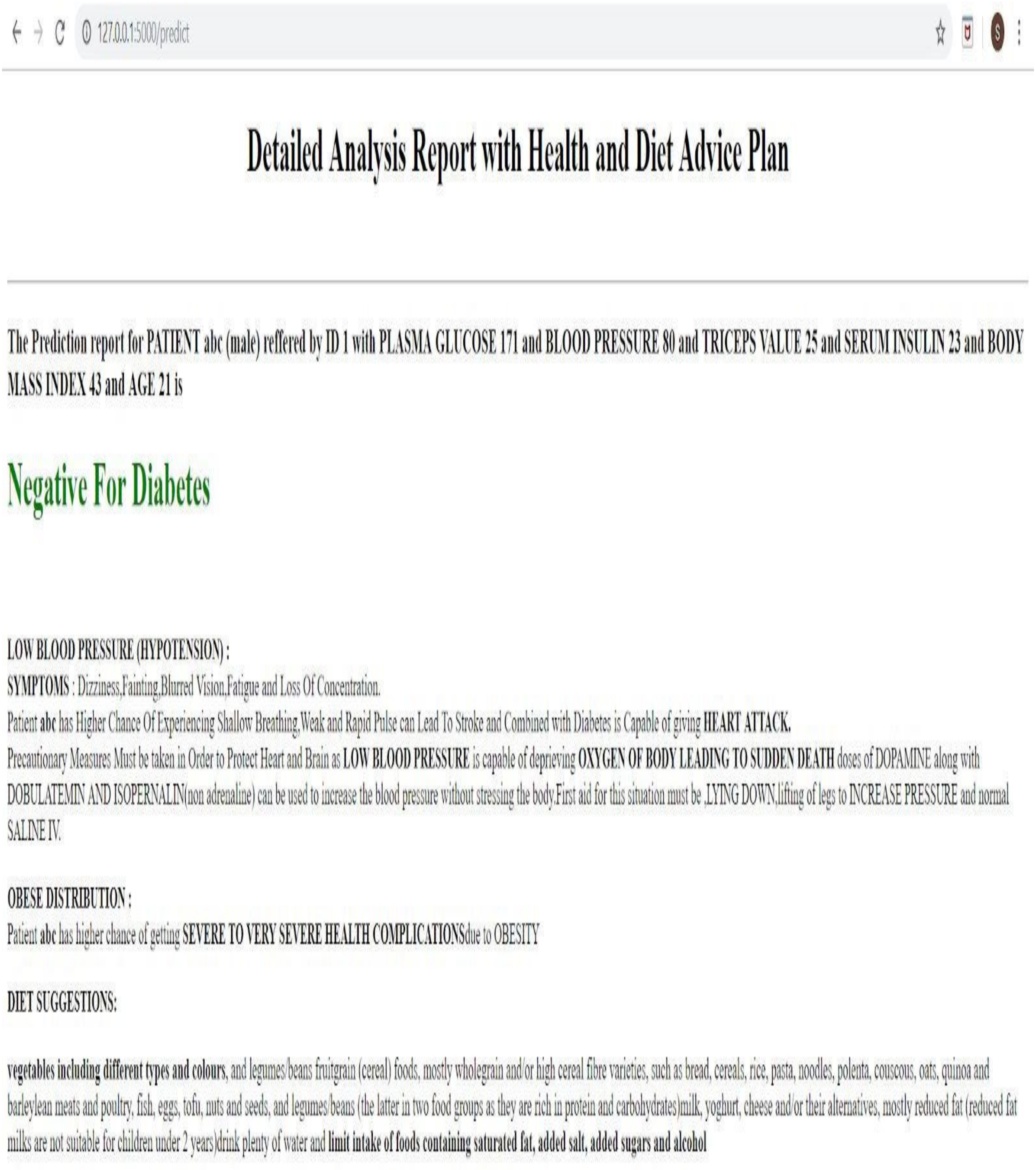
```
<!-- HIGH BLOOD PRESSURE-->


{% elif dbp|int >= 120 and dbp|int <= 170%}
<p><b> HIGH BLOOD PRESSURE :</b>
    <br>
<b> SYMPTOMS</b> : Head Ache,Nausea,Vomiting and Nose Bleeds,No Symptoms Sometimes.
    <br>
Patient <b>{{name}}</b> has Higher Chances Of Experiencing Heart Attack and Hypertension and Can Finally Lead To <b>STROKE</b>.
    <br>
Precautionary Measures Must be Taken By<b> Reducing Bad Cholestrol </b>, <b>avoiding SALT and OIL CONTENT</b> in Food and Frequent Excercises</p>


<!-- HYPERTENSION-->
{% elif dbp|int >= 171 and dbp|int <=260%}
<p><b> VERY HIGH BLOOD PRESSURE (HYPERTENSION) :</b>
    <br>
<b> SYMPTOMS </b> : Dizziness,Breathing Troubles,Uncontrollable Pulse Rate,Sometimes it shows No Symptoms.
    <br>
Patient <b>{{name}}</b> has Higher Chance Of Experiencing Silent to Severe Heart Attack with Sudden Burst of Blood Vessels Leading to Death.
    <br>
Precautionary Measures Must be Taken By <b>Reducing Bad Cholestrol</b> with Regular Physical Excercises , Stress Reduction Therapies and Mainly <b>LIFESTYLE CHANGES </b>
along with Moderate dosages of NITROGLYCERINE,NITROPRUNADE,IV LEBATOL AND IV CLOLIDINE CAN REDUCE BLOOD PRESSURE TO AN EXTEND.</p>
<p> <b>ECHOCARDIOGRAM,ECHO,TREADMILL and ANGIOGRAM  tests</b> must be performed to determine HEART DAMAGE.</p>
{% endif %}
```

Fig: A17 Symptoms And Precautions For Diabetes



```
<!-- OVERWEIGHT-->


 {% if bmi|int >= 25 and bmi|int <= 29.9%}
<p><b> OVERWEIGHT DISTRIBUTION :</b>
   <br>
   Patient <b>{{name}}</b> Has <b> higher rate of BLOOD CHOLESTROL </b> and have comparitively more chance than OBESE category people to survive Health deprieving
   conditions  </p>
  <b>DIET SUGGESTIONS</b>
 <p><b>D-A-S-H diet:</b>The Dietary Approaches to Stop Hypertension, or DASH, diet was originally designed to be a heart-healthy diet that lowers blood pressure, not a
 weight-loss diet. But because of the focus on whole foods, such as fruits, vegetables and whole grains, the diet is high in fiber, which helps you feel full so you may
 eat less overall. You'll also include lean proteins, low-fat dairy and healthy oils to limit saturated fat, but you can still enjoy sweets in moderation, several times a
 week. Choosing lower-sodium foods is a key component of the plan, and it's what makes it helpful for controlling blood pressure.</p>
 <p><b>plant-based diet:</b>All vegetarian diets include fruits, vegetables, legumes and grains, so they're high in fiber and antioxidants. As an added bonus, people who
 follow a vegetarian diet have lower BMIs overall, according to the National Institute of Diabetes and Digestive and Kidney Disorders.</p>
 <p><b>low carbohydrate diet:</b>plans focus on protein, vegetables, nuts and seeds, cheese, and healthy fats with an initial strict limitation of dairy, fruits, legumes,
 starchy vegetables and whole grains. As the diet progresses, though, you can add slowly carb-containing foods each week, up to a certain limit.</p>
{% endif %}


<!-- OBESE DISTRIBUTION-->


 {% if bmi|int >= 30%}
<p><b> OBESE DISTRIBUTION :</b>
   <br>
   Patient <b>{{name}}</b> has higher chance of getting <b>SEVERE TO VERY SEVERE HEALTH COMPLICATIONS</b>due to OBESITY </p>
<b>DIET SUGGESTIONS:</b>
<p><b>vegetables including different types and colours</b>, and legumes/beans fruitgrain (cereal) foods, mostly wholegrain and/or high cereal fibre varieties, such as
bread, cereals, rice, pasta, noodles, polenta, couscous, oats, quinoa and barleylean meats and poultry, fish, eggs, tofu, nuts and seeds, and legumes/beans (the latter in
two food groups as they are rich in protein and carbohydrates)milk, yoghurt, cheese and/or their alternatives, mostly reduced fat (reduced fat milks are not suitable for
children under 2 years)drink plenty of water and <b>limit intake of foods containing saturated fat, added salt, added sugars and alcohol</b></p>
{% endif %}
```

Fig: A18 Python Code To Determine The Type Of Diabetes



Detailed Analysis Report with Health and Diet Advice Plan

The Prediction report for PATIENT abc (male) reffered by ID 1 with PLASMA GLUCOSE 171 and BLOOD PRESSURE 80 and TRICEPS VALUE 25 and SERUM INSULIN 23 and BODY MASS INDEX 43 and AGE 21 is

Negative For Diabetes

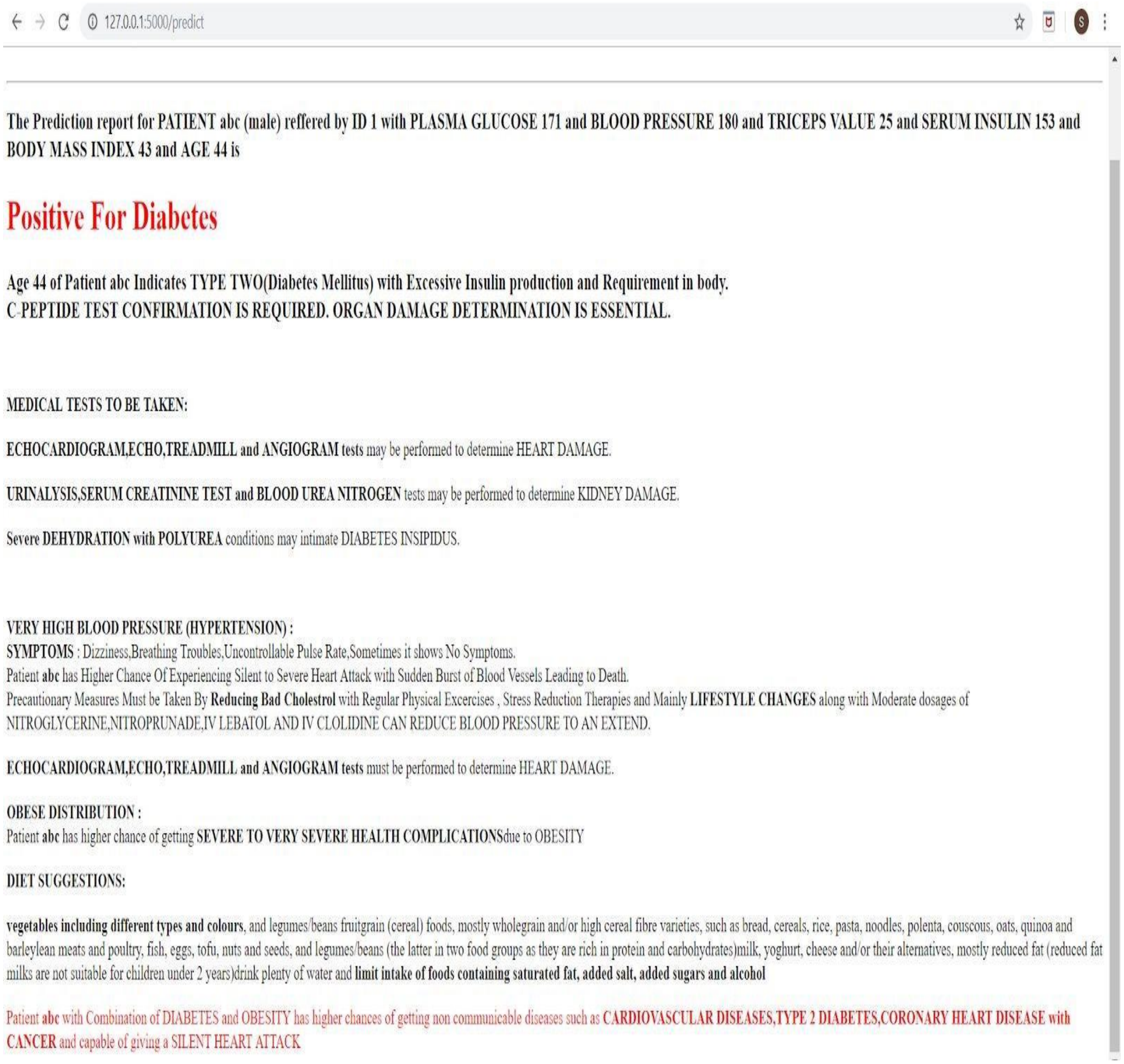
LOW BLOOD PRESSURE (HYPOTENSION) :
SYMPTOMS : Dizziness,Fainting,Blurred Vision,Fatigue and Loss Of Concentration.
Patient **abc** has Higher Chance Of Experiencing Shallow Breathing,Weak and Rapid Pulse can Lead To Stroke and Combined with Diabetes is Capable of giving **HEART ATTACK.**
Precautionary Measures Must be taken in Order to Protect Heart and Brain as **LOW BLOOD PRESSURE** is capable of deprieving **OXYGEN OF BODY LEADING TO SUDDEN DEATH** doses of DOPAMINE along with DOBULATEMIN AND ISOPERNALIN(non adrenaline) can be used to increase the blood pressure without stressing the body.First aid for this situation must be ,LYING DOWN,lifting of legs to INCREASE PRESSURE and normal SALINE IV.

OBESE DISTRIBUTION :
Patient **abc** has higher chance of getting **SEVERE TO VERY SEVERE HEALTH COMPLICATIONS**due to OBESITY

DIET SUGGESTIONS:

vegetables including different types and colours, and legumes/beans fruitgrain (cereal) foods, mostly wholegrain and/or high cereal fibre varieties, such as bread, cereals, rice, pasta, noodles, polenta, couscous, oats, quinoa and barleylean meats and poultry, fish, eggs, tofu, nuts and seeds, and legumes/beans (the latter in two food groups as they are rich in protein and carbohydrates)milk, yoghurt, cheese and/or their alternatives, mostly reduced fat (reduced fat milks are not suitable for children under 2 years)drink plenty of water and **limit intake of foods containing saturated fat, added salt, added sugars and alcohol**

Fig: 5.19 Output 1



The Prediction report for PATIENT abc (male) reffered by ID 1 with PLASMA GLUCOSE 171 and BLOOD PRESSURE 180 and TRICEPS VALUE 25 and SERUM INSULIN 153 and BODY MASS INDEX 43 and AGE 44 is

Positive For Diabetes

Age 44 of Patient abc Indicates TYPE TWO(Diabetes Mellitus) with Excessive Insulin production and Requirement in body.
C-PEPTIDE TEST CONFIRMATION IS REQUIRED. ORGAN DAMAGE DETERMINATION IS ESSENTIAL.

MEDICAL TESTS TO BE TAKEN:

ECHOCARDIOGRAM,ECHO,TREADMILL and ANGIOGRAM tests may be performed to determine HEART DAMAGE.

URINALYSIS,SERUM CREATININE TEST and BLOOD UREA NITROGEN tests may be performed to determine KIDNEY DAMAGE.

Severe DEHYDRATION with POLYUREA conditions may intimate DIABETES INSIPIDUS.

VERY HIGH BLOOD PRESSURE (HYPERTENSION) :
SYMPTOMS : Dizziness,Breathing Troubles,Uncontrollable Pulse Rate,Sometimes it shows No Symptoms.
Patient **abc** has Higher Chance Of Experiencing Silent to Severe Heart Attack with Sudden Burst of Blood Vessels Leading to Death.
Precautionary Measures Must be Taken By **Reducing Bad Cholestrol** with Regular Physical Excercises , Stress Reduction Therapies and Mainly **LIFESTYLE CHANGES** along with Moderate dosages of NITROGLYCERINE,NITROPRUNADE,IV LEBATOL AND IV CLOLIDINE CAN REDUCE BLOOD PRESSURE TO AN EXTEND.

ECHOCARDIOGRAM,ECHO,TREADMILL and ANGIOGRAM tests must be performed to determine HEART DAMAGE.

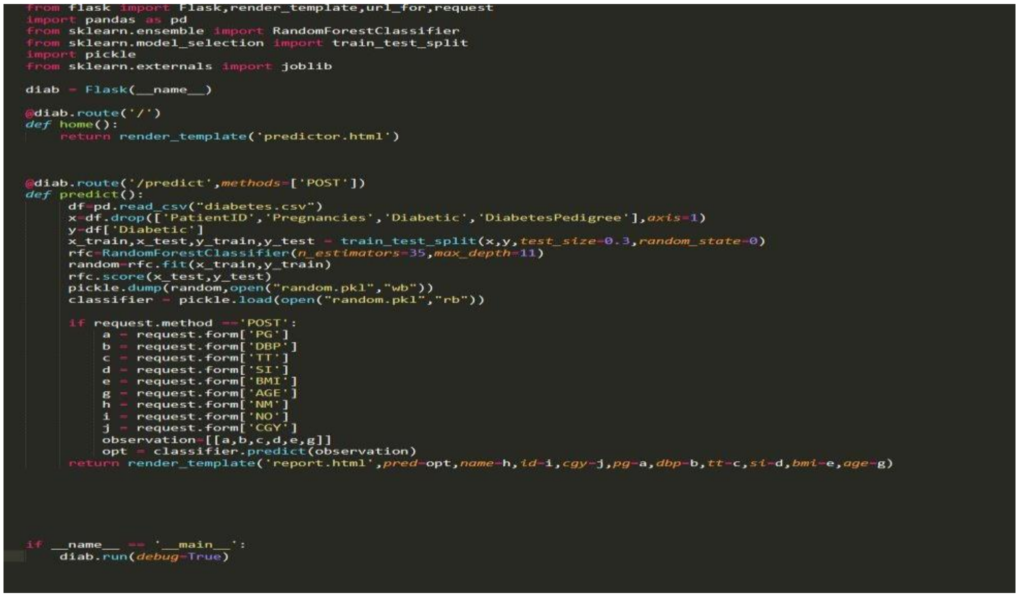
OBESE DISTRIBUTION :
Patient **abc** has higher chance of getting **SEVERE TO VERY SEVERE HEALTH COMPLICATIONS**due to OBESITY

DIET SUGGESTIONS:

vegetables including different types and colours, and legumes/beans fruitgrain (cereal) foods, mostly wholegrain and/or high cereal fibre varieties, such as bread, cereals, rice, pasta, noodles, polenta, couscous, oats, quinoa and barleylean meats and poultry, fish, eggs, tofu, nuts and seeds, and legumes/beans (the latter in two food groups as they are rich in protein and carbohydrates)milk, yoghurt, cheese and/or their alternatives, mostly reduced fat (reduced fat milks are not suitable for children under 2 years)drink plenty of water and **limit intake of foods containing saturated fat, added salt, added sugars and alcohol**

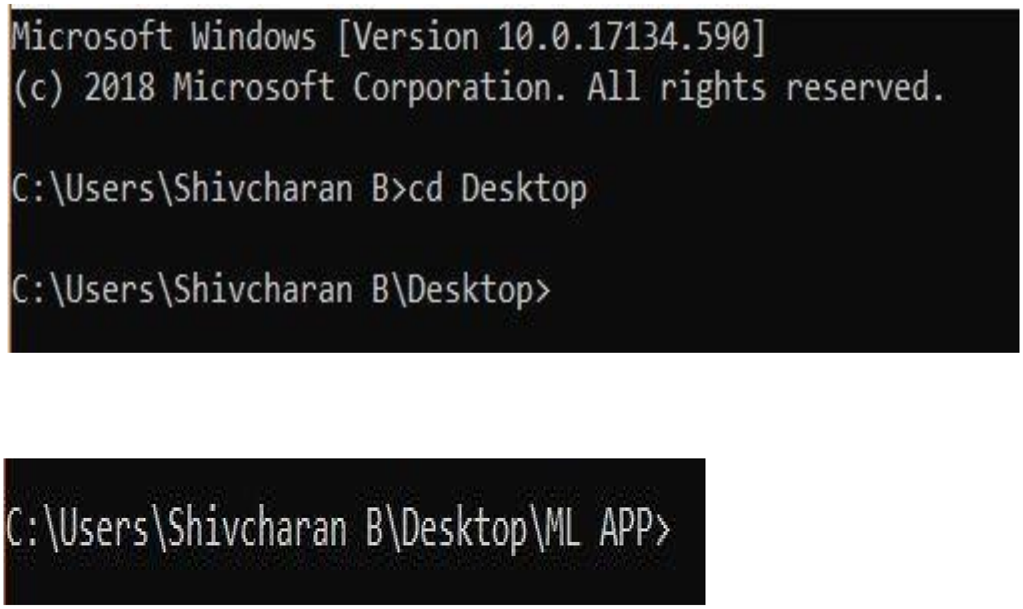
Patient **abc** with Combination of DIABETES and OBESITY has higher chances of getting non communicable diseases such as **CARDIOVASCULAR DISEASES,TYPE 2 DIABETES,CORONARY HEART DISEASE with CANCER** and capable of giving a SILENT HEART ATTACK

Fig: 5.20 Output 2



```
from flask import Flask,render_template,url_for,request
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
import pickle
from sklearn.externals import joblib

diab = Flask(__name__)

@diab.route('/')
def home():
    return render_template('predictor.html')


@diab.route('/predict',methods=['POST'])
def predict():
        df=pd.read_csv("diabetes.csv")
        x=df.drop(['PatientID','Pregnancies','Diabetic','DiabetesPedigree'],axis=1)
        y=df['Diabetic']
        x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.3,random_state=0)
        rfc=RandomForestClassifier(n_estimators=35,max_depth=11)
        random=rfc.fit(x_train,y_train)
        rfc.score(x_test,y_test)
        pickle.dump(random,open("random.pkl","wb"))
        classifier = pickle.load(open("random.pkl","rb"))

        if request.method =='POST':
            a = request.form['PG']
            b = request.form['DBP']
            c = request.form['TT']
            d = request.form['SI']
            e = request.form['BMI']
            g = request.form['AGE']
            h = request.form['NM']
            i = request.form['NO']
            j = request.form['CGY']
            observation=[[a,b,c,d,e,g]]
            opt = classifier.predict(observation)
        return render_template('report.html',pred=opt,name=h,id=i,cgy=j,pg=a,dbp=b,tt=c,si=d,bmi=e,age=g)



if __name__ == '__main__':
    diab.run(debug=True)
```

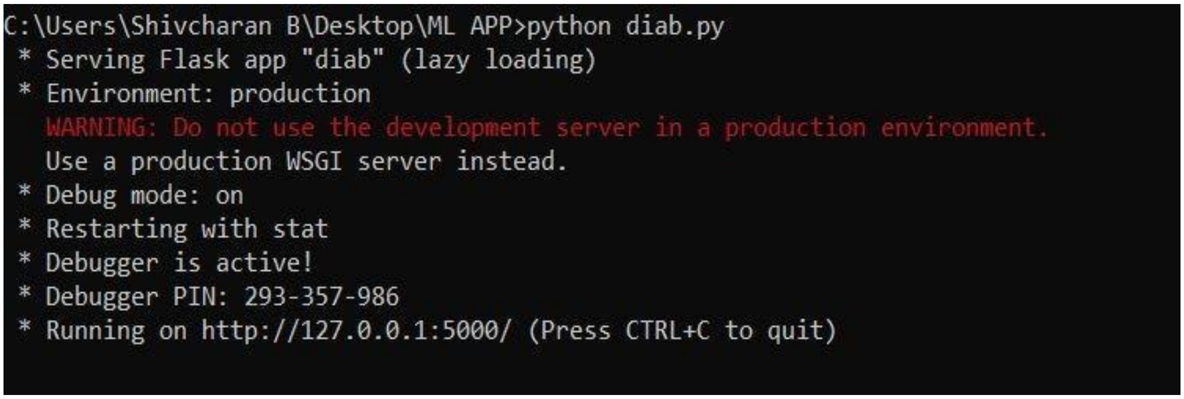
Fig: A21 Diab.run



```
Microsoft Windows [Version 10.0.17134.590]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\Shivcharan B>cd Desktop

C:\Users\Shivcharan B\Desktop>
```

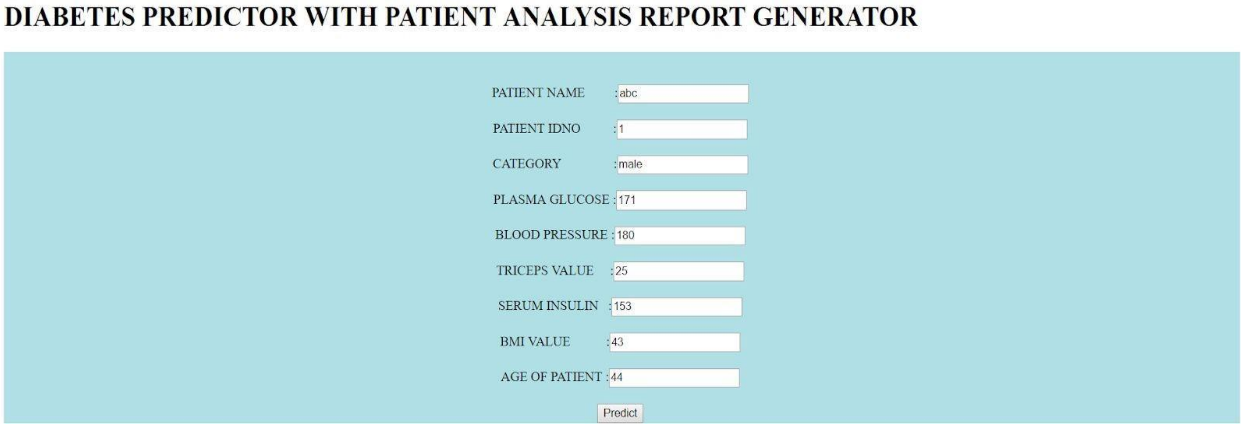
```
C:\Users\Shivcharan B\Desktop\ML APP>
```

Fig: A22 Path Setting



```
C:\Users\Shivcharan B\Desktop\ML APP>python diab.py
 * Serving Flask app "diab" (lazy loading)
 * Environment: production
   WARNING: Do not use the development server in a production environment.
   Use a production WSGI server instead.
 * Debug mode: on
 * Restarting with stat
 * Debugger is active!
 * Debugger PIN: 293-357-986
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

Fig: A23 Server Running



DIABETES PREDICTOR WITH PATIENT ANALYSIS REPORT GENERATOR

PATIENT NAME :abc

PATIENT IDNO :1

CATEGORY :male

PLASMA GLUCOSE :171

BLOOD PRESSURE :180

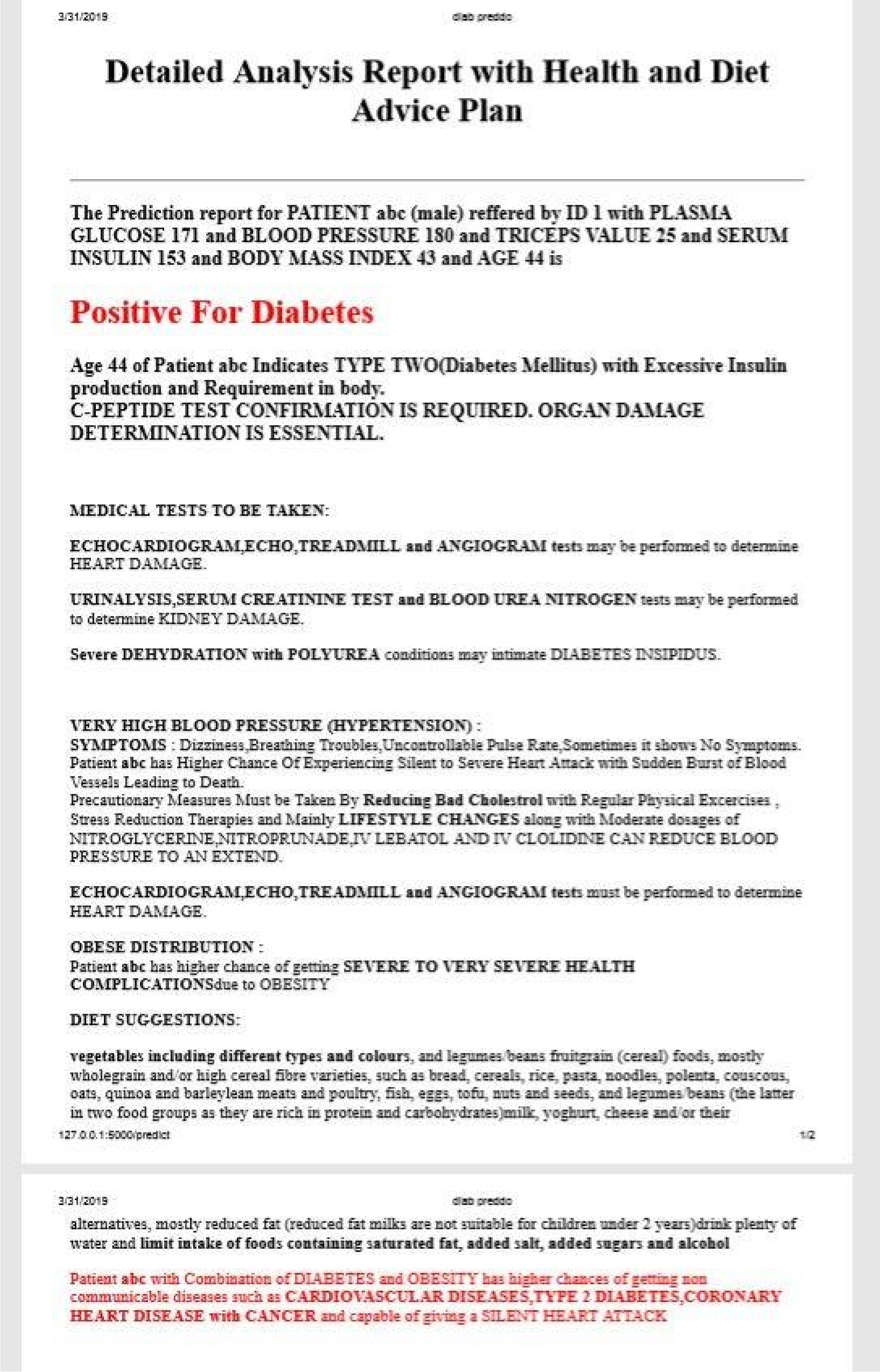
TRICEPS VALUE :25

SERUM INSULIN :153

BMI VALUE :43

AGE OF PATIENT :44

Fig: A24 Feeding Of Data



3/31/2019 diab preddo

Detailed Analysis Report with Health and Diet Advice Plan

The Prediction report for PATIENT abc (male) reffered by ID 1 with PLASMA GLUCOSE 171 and BLOOD PRESSURE 180 and TRICEPS VALUE 25 and SERUM INSULIN 153 and BODY MASS INDEX 43 and AGE 44 is

Positive For Diabetes

Age 44 of Patient abc Indicates TYPE TWO(Diabetes Mellitus) with Excessive Insulin production and Requirement in body.
C-PEPTIDE TEST CONFIRMATION IS REQUIRED. ORGAN DAMAGE DETERMINATION IS ESSENTIAL.

MEDICAL TESTS TO BE TAKEN:

ECHOCARDIOGRAM,ECHO,TREADMILL and ANGIOGRAM tests may be performed to determine HEART DAMAGE.

URINALYSIS,SERUM CREATININE TEST and BLOOD UREA NITROGEN tests may be performed to determine KIDNEY DAMAGE.

Severe DEHYDRATION with POLYUREA conditions may intimate DIABETES INSIPIDUS.

VERY HIGH BLOOD PRESSURE (HYPERTENSION) :
SYMPTOMS : Dizziness,Breathing Troubles,Uncontrollable Pulse Rate,Sometimes it shows No Symptoms.
Patient **abc** has Higher Chance Of Experiencing Silent to Severe Heart Attack with Sudden Burst of Blood Vessels Leading to Death.
Precautionary Measures Must be Taken By **Reducing Bad Cholestrol** with Regular Physical Excercises , Stress Reduction Therapies and Mainly **LIFESTYLE CHANGES** along with Moderate dosages of NITROGLYCERINE,NITROPRUNADE,IV LEBATOL AND IV CLOLIDINE CAN REDUCE BLOOD PRESSURE TO AN EXTEND.

ECHOCARDIOGRAM,ECHO,TREADMILL and ANGIOGRAM tests must be performed to determine HEART DAMAGE.

OBESE DISTRIBUTION :
Patient **abc** has higher chance of getting **SEVERE TO VERY SEVERE HEALTH COMPLICATIONS**due to OBESITY

DIET SUGGESTIONS:

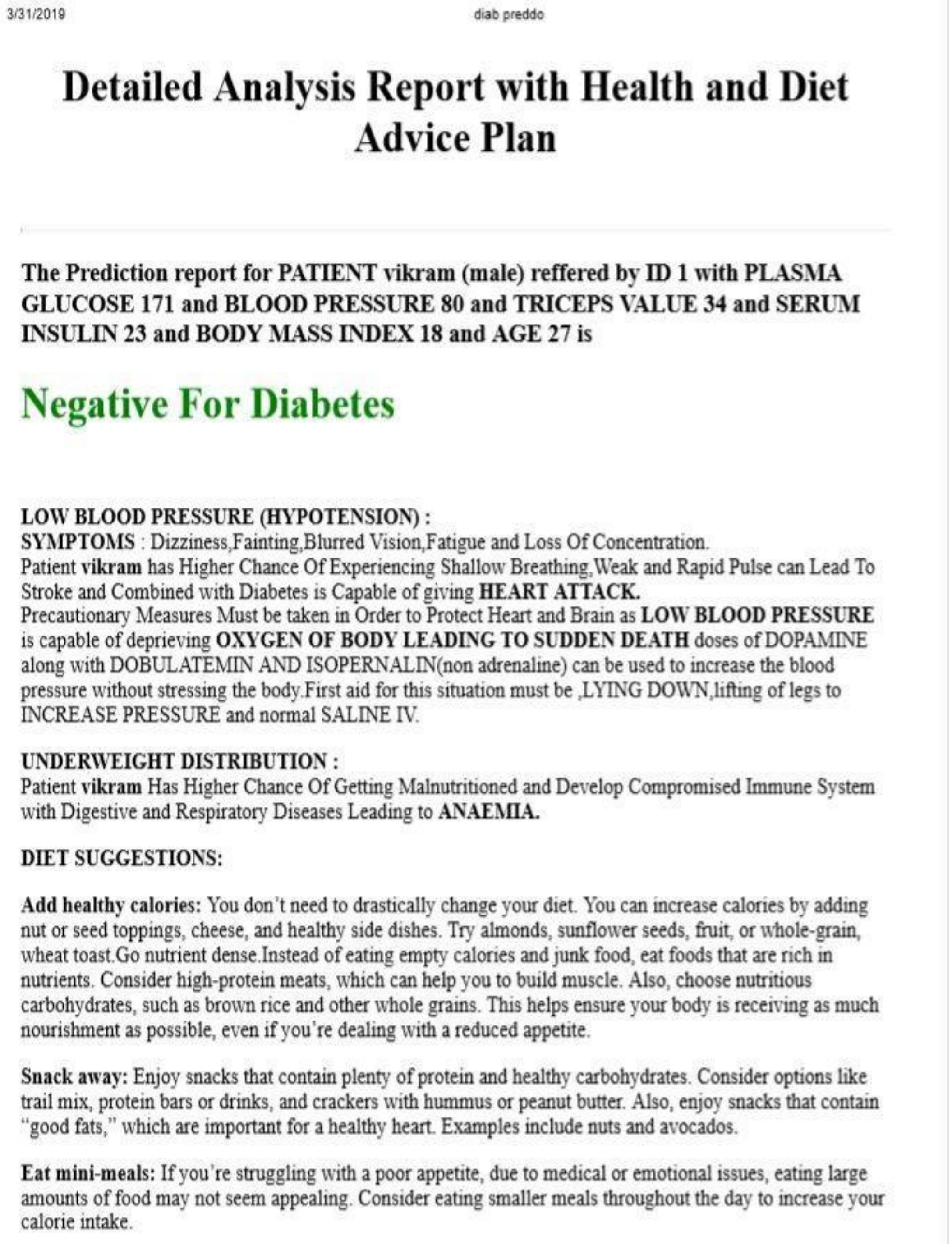
vegetables including different types and colours, and legumes/beans fruitgrain (cereal) foods, mostly wholegrain and/or high cereal fibre varieties, such as bread, cereals, rice, pasta, noodles, polenta, couscous, oats, quinoa and barleylean meats and poultry, fish, eggs, tofu, nuts and seeds, and legumes/beans (the latter in two food groups as they are rich in protein and carbohydrates)milk, yoghurt, cheese and/or their alternatives, mostly reduced fat (reduced fat milks are not suitable for children under 2 years)drink plenty of water and **limit intake of foods containing saturated fat, added salt, added sugars and alcohol**

127.0.0.1:5000/predict 1/2

3/31/2019 diab preddo

Patient **abc** with Combination of DIABETES and OBESITY has higher chances of getting non communicable diseases such as **CARDIOVASCULAR DISEASES,TYPE 2 DIABETES,CORONARY HEART DISEASE with CANCER** and capable of giving a SILENT HEART ATTACK

Fig: A25 Sample Output



3/31/2019 diab preddo

Detailed Analysis Report with Health and Diet Advice Plan

The Prediction report for PATIENT vikram (male) reffered by ID 1 with PLASMA GLUCOSE 171 and BLOOD PRESSURE 80 and TRICEPS VALUE 34 and SERUM INSULIN 23 and BODY MASS INDEX 18 and AGE 27 is

Negative For Diabetes

LOW BLOOD PRESSURE (HYPOTENSION) :
SYMPTOMS : Dizziness,Fainting,Blurred Vision,Fatigue and Loss Of Concentration.
Patient **vikram** has Higher Chance Of Experiencing Shallow Breathing,Weak and Rapid Pulse can Lead To Stroke and Combined with Diabetes is Capable of giving **HEART ATTACK.**
Precautionary Measures Must be taken in Order to Protect Heart and Brain as **LOW BLOOD PRESSURE** is capable of deprieving **OXYGEN OF BODY LEADING TO SUDDEN DEATH** doses of DOPAMINE along with DOBULATEMIN AND ISOPERNALIN(non adrenaline) can be used to increase the blood pressure without stressing the body.First aid for this situation must be ,LYING DOWN,lifting of legs to INCREASE PRESSURE and normal SALINE IV.

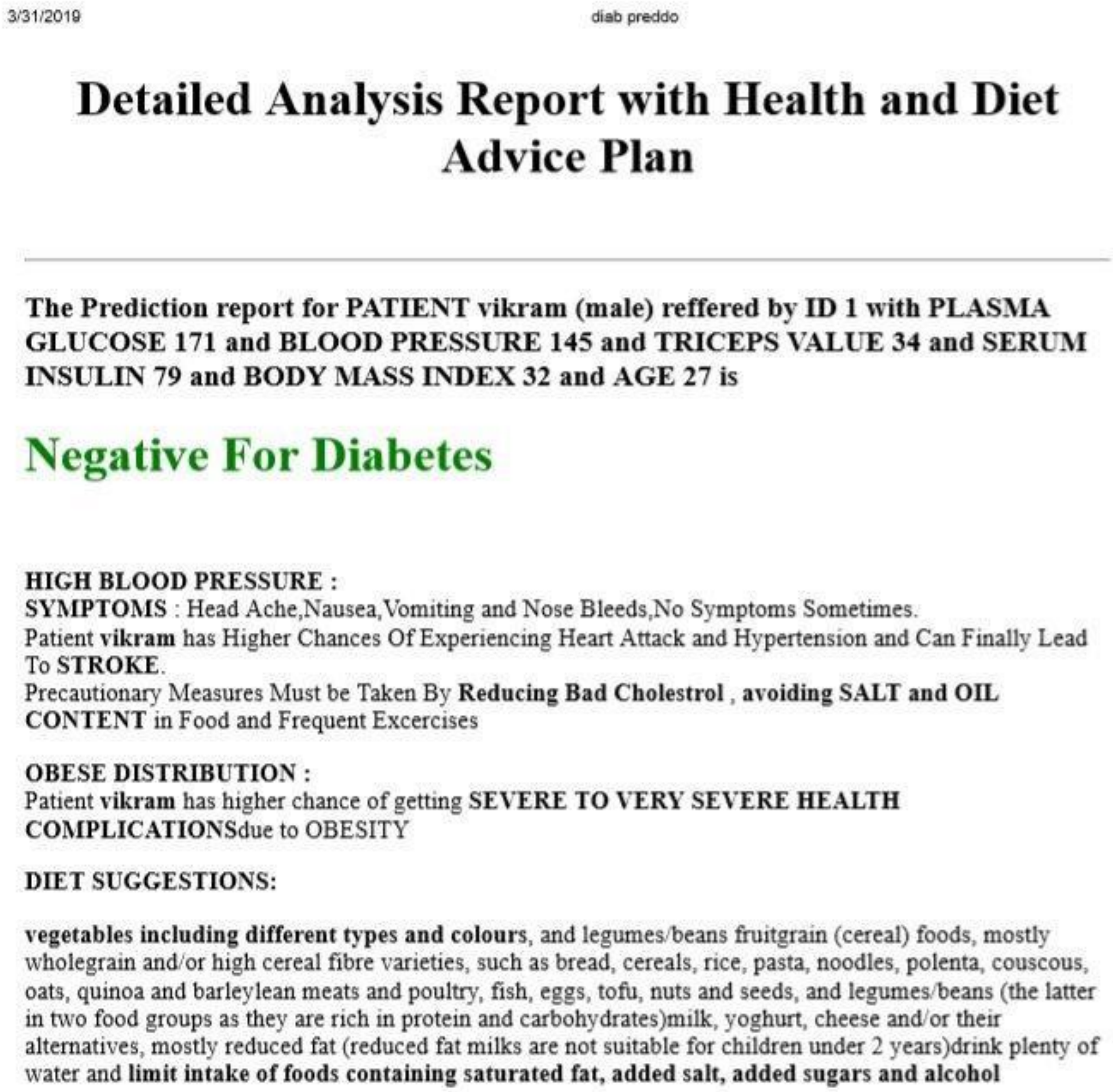
UNDERWEIGHT DISTRIBUTION :
Patient **vikram** Has Higher Chance Of Getting Malnutritioned and Develop Compromised Immune System with Digestive and Respiratory Diseases Leading to **ANAEMIA.**

DIET SUGGESTIONS:

Add healthy calories: You don't need to drastically change your diet. You can increase calories by adding nut or seed toppings, cheese, and healthy side dishes. Try almonds, sunflower seeds, fruit, or whole-grain, wheat toast.Go nutrient dense.Instead of eating empty calories and junk food, eat foods that are rich in nutrients. Consider high-protein meats, which can help you to build muscle. Also, choose nutritious carbohydrates, such as brown rice and other whole grains. This helps ensure your body is receiving as much nourishment as possible, even if you're dealing with a reduced appetite.

Snack away: Enjoy snacks that contain plenty of protein and healthy carbohydrates. Consider options like trail mix, protein bars or drinks, and crackers with hummus or peanut butter. Also, enjoy snacks that contain "good fats," which are important for a healthy heart. Examples include nuts and avocados.

Eat mini-meals: If you're struggling with a poor appetite, due to medical or emotional issues, eating large amounts of food may not seem appealing. Consider eating smaller meals throughout the day to increase your calorie intake.



3/31/2019 diab preddo

Detailed Analysis Report with Health and Diet Advice Plan

The Prediction report for PATIENT vikram (male) reffered by ID 1 with PLASMA GLUCOSE 171 and BLOOD PRESSURE 145 and TRICEPS VALUE 34 and SERUM INSULIN 79 and BODY MASS INDEX 32 and AGE 27 is

Negative For Diabetes

HIGH BLOOD PRESSURE :
SYMPTOMS : Head Ache,Nausea,Vomiting and Nose Bleeds,No Symptoms Sometimes.
Patient **vikram** has Higher Chances Of Experiencing Heart Attack and Hypertension and Can Finally Lead To **STROKE**.
Precautionary Measures Must be Taken By **Reducing Bad Cholestrol** , **avoiding SALT and OIL CONTENT** in Food and Frequent Excercises

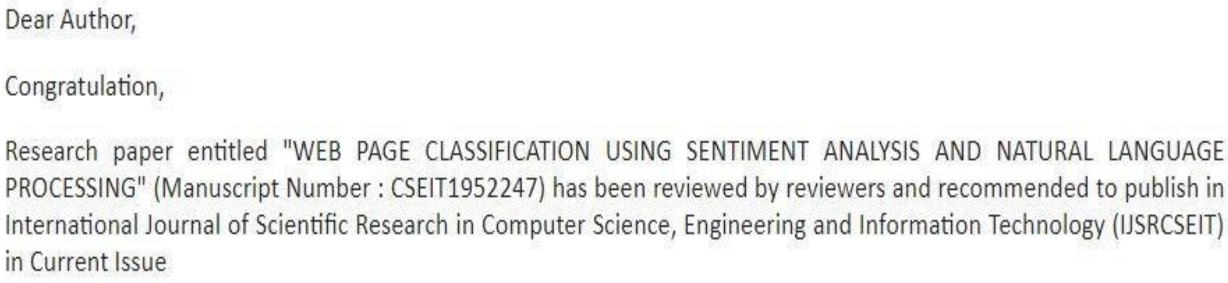
OBESE DISTRIBUTION :
Patient **vikram** has higher chance of getting **SEVERE TO VERY SEVERE HEALTH COMPLICATIONS**due to OBESITY

DIET SUGGESTIONS:

vegetables including different types and colours, and legumes/beans fruitgrain (cereal) foods, mostly wholegrain and/or high cereal fibre varieties, such as bread, cereals, rice, pasta, noodles, polenta, couscous, oats, quinoa and barleylean meats and poultry, fish, eggs, tofu, nuts and seeds, and legumes/beans (the latter in two food groups as they are rich in protein and carbohydrates)milk, yoghurt, cheese and/or their alternatives, mostly reduced fat (reduced fat milks are not suitable for children under 2 years)drink plenty of water and **limit intake of foods containing saturated fat, added salt, added sugars and alcohol**

APPENDIX 2:

PUBLICATIONS:



Dear Author,

Congratulation,

Research paper entitled "WEB PAGE CLASSIFICATION USING SENTIMENT ANALYSIS AND NATURAL LANGUAGE PROCESSING" (Manuscript Number : CSEIT1952247) has been reviewed by reviewers and recommended to publish in International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT) in Current Issue

REFERENCES:

[1]Disease Predictor by Machine Learning over big data from health-care communities ,**MIN CHEN , YIXUE HAO, KAI HWANG, LU WANG, LIN WANG**

[2]Machine Learning in medical application,**GEORGE D.MAGOULAS AND ANDRIANA PRENTZA**

[3]Comparative analysis of Machine Learning methods for classification type decision problem in health care,**NAHIT EMANET, HALIL ROZ, NAZAN BAYRAM, DURSUN DELEN**

[4]A Few Useful Things to Know about Machine Learning ,**PEDRO DOMINGOS**

[5]Machine that Learn and Teach Seamlessly.**GARY STEIN AND CLAYTON BARHAM.**

[6]Automated Microscopy and Machine Learning for Expert-Level Malaria Field Diagnosis. **CHARLES B.DELAHUNT AND BENJAMIN K.**

[7]A Health big data Analytics- A TECHNOLOGY SURVEY,**GASPARD HARERIMANA AND HUNG KOOK PARK**