Diabetic Prediction Using Devops CI/CD Pipeline

Project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering

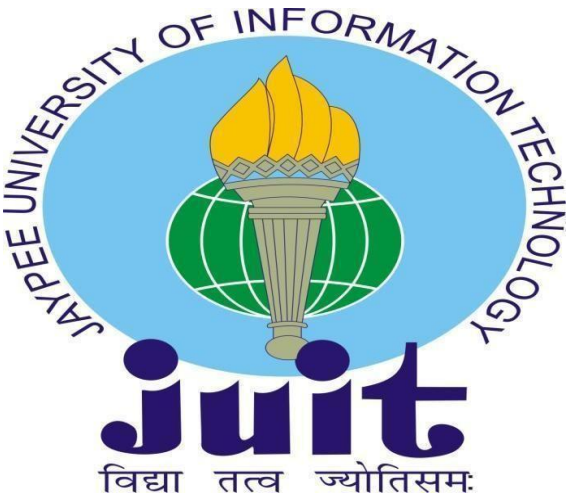
By

Utkarsh Bhatnagar (191297)

Under the supervision of

Dr Rajni Mohana

To



Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology Waknaghat,

Solan-173234, Himachal Pradesh

Candidate's Declaration

I hereby declare that the work presented in this report entitled **"Diabetic Prediction Using DevOps CI/CD Pipeline "** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2015 to December 2015 under the supervision of Dr. Rajni Mohana (Associate Professor (SG)).

I also authenticate that I have carried out the above-mentioned project work under the proficiency stream **Cloud Computing.**
The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Utkarsh Bhatnagar (191297)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr. Rajni Mohana

Associate Professor (SG)

Computer Science & Engineering Department

Dated:18/11/2023

Plagiarism Certificate

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: .....

Type of Document (Tick): PhD Thesis | M.Tech Dissertation/ Report | B.Tech Project Report | Paper

Name: _____ **Department:** _____ **Enrolment No** _____

Contact No. _____ **E-mail.** _____

Name of the Supervisor: _____

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): _____

_____

_____

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages =
- Total No. of Preliminary pages =
- Total No. of pages accommodate bibliography/references =

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at .....(%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

(Signature of Guide/Supervisor) **Signature of HOD**

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

<table>
<tr><th>Copy Received on</th><th>Excluded</th><th>Similarity Index (%)</th><th colspan="2">Generated Plagiarism Report Details (Title, Abstract & Chapters)</th></tr>
<tr><td rowspan="2"></td><td rowspan="4">• All Preliminary Pages
• Bibliography/Images/Quotes
• 14 Words String</td><td rowspan="2"></td><td>Word Counts</td><td></td></tr>
<tr><td rowspan="2">Report Generated on</td><td>Character Counts</td><td></td></tr>
<tr><td rowspan="2"></td><td>Submission ID</td><td>Total Pages Scanned</td><td></td></tr>
<tr><td></td><td>File Size</td><td></td></tr>
</table>

Checked by
Name & Signature **Librarian**

.....

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com

ACKNOWLEDGEMENT

We owe our profound gratitude and indebtedness to our project supervisor **Dr Rajni Mohana**, who took a keen interest and guided us all along in our project work titled
—Diabetic Prediction Using DevOps CI/CD Pipeline, till the completion of our project by providing all the necessary information for developing the project. The project development helped us in research, and we got to know many new things in our domain. We are really thankful to him.

Firstly, I express my heartiest thanks and gratefulness to almighty God for His divine blessing make it possible to complete the project work successfully.

I am really grateful and wish my profound indebtedness to Supervisor Dr.Rajni Mohana Associate Professor, Department of CSE Jaypee University of Information Technology, Waknaghat.

His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to Dr.Rajni Mohana, Department of CSE, for his kind help in finishing my project.

I would also generously welcome each one of those individuals who have helped me straightforwardly or in a roundabout way in making this project a win.

In this unique situation, I might want to thank the various staff individuals, educating and non-instructing, who have developed their convenient help and facilitated my undertaking. Finally, I must acknowledge with due respect the constant support and patients of my parents.

Table Of Content

Table of Figure

ABSTRACT

Diabetes is one of the most common and quickly expanding illnesses in the world (World Health Organisation, WHO). In most nations, it is a very serious pathological condition.

The polygenic disease is a disorder in which your body is unable to produce the necessary amount of internal secretion to regulate the body's level of sugar (National Center for Biotechnology Information, NCBI). In general, factors like female gender, age above 35, and those with excessive weight are associated with a higher chance of developing diabetes.

Diabetes is a condition brought on by a high blood glucose level. Diabetes is a disorder that should not be ignored since it can have major side effects on a person's heart, kidneys, blood pressure, eyes, and other body organs if left untreated.

Diabetes can be managed if it is discovered early. We will perform early diabetes prediction in a human body or patient using a range of machine learning approaches for a greater level of accuracy.. approaches for machine learning By creating models from patient data, you can improve the outcome of your forecast.

A frequent clinical approach used to capture a visualisation of the retina is retinal imaging. The early identification of retinal disorders such hypertension, diabetes, and glaucoma uses the segmentation of blood vessels in retinal pictures.

Hyperglycemia, a metabolic disease brought on by the body's inability to adequately secrete and react to insulin, is a hallmark of diabetes mellitus. Diabetes poses a risk to important body parts like the eyes, kidneys, nerves, heart, and blood arteries and can be fatal if not adequately managed or detected in time. Machine learning has emerged as a promising option for the prediction of diabetes after years of research in computational diagnosis of diabetes. The accuracy rate to date, though, implies that there is still a lot of space for development. With the use of the PIMA Indian dataset and the laboratory for the Medical City Hospital (LMCH) diabetes programme, we propose a machine-learning framework for diabetes prediction and diagnosis in this research.

Today's needs include early detection and diagnosis of this diabetes disease. A significant classification issue is the diagnosis and examination of diabetes data. To build a classifier, it must be practical, valid, and economical

Chapter 1: Introduction

1.1 Introduction

Diabetes is the third biggest cause of death, behind cancer and heart disease. But as machine learning techniques advance, we might be able to address this problem. The goal of data mining and machine learning is to produce an intelligible pattern description.Using machine learning, we will create a diabetic diagnostic tool that can determine whether or not a patient has the disease.

Diabetes is a condition brought on by a high blood glucose level. Diabetes is a disorder that should not be ignored since it can have major side effects on a person's heart, kidneys, blood pressure, eyes, and other body organs if left untreated.

In fact, catching an infection early aids in treating people before it becomes dangerous. The enormous amount of data related to diabetes can be mined using machine learning and data mining techniques to uncover hidden information. With the advancement of technology, it has also been very successful in the field of research medicine.

The aim of this project is to apply machine learning techniques for the identification of patterns that can help with diabetes patient care and medical diagnosis. Effective maintenance of the patient's health depends on the early identification of diabetes. The primary problem with diabetes detection is that manual diagnosis requires a lot of time, money, and effort.

The latter is also more challenging, and the use of machine learning algorithms has facilitated the early identification of diabetes. And create a website where anyone can enter their information and easily able to know whether they have diabetes or not.

In order to eliminate manual labour, our project primarily aims to replace the manual labour performed by hospital staff members with an online and automated system.

The hospitals' present work flow entails gathering patient samples, keeping track of X-ray reports, and manually gathering patient medical reports. Thus, we suggest an online option to streamline the doctor's appointment process for patients and healthcare facilities.

1.2 Problem Statement

Chronic, lifelong diabetes mellitus is brought on by abnormally high blood sugar levels. In the medical industry, classification algorithms are frequently employed to divide data into various classes based on restrictions placed on a single classifier.

According to a 2019 World Health Organization estimate, 463 million people worldwide have diabetes, and there were 1.5 million related fatalities.

With this knowledge, it is simple to assume that a sizeable percentage of diabetes cases are severe and chronic.To identify diseases, several researchers test various machine learning methods more successful.

Patients who may have diabetes need to undergo a battery of tests and examinations in order to appropriately diagnose it. These evaluations could entail pointless or repeated medical procedures, which causes issues and wastes time and resources.

Diabetes has a far higher economic cost than it does directly medical expenditures to the healthcare system since it diminishes quality of life and decreases productivity at work.

Lack of a proper diagnosis method, a lack of funding, and a general lack of information are the key contributors to these negative effects.

Therefore, a significant economic burden could surely be reduced and the patient's ability to manage their diabetes helped if the sickness could be completely avoided through early detection.

Today's needs include early detection and diagnosis of this diabetes disease. A significant classification issue is the diagnosis and examination of diabetes data. To build a classifier, it must be practical, valid, and economical.

After heart disease and cancer, diabetes is the leading cause of death. But with the development of machine learning methods, we may be able to solve this issue. Data mining and machine learning seek to extract knowledge from the dataset's data by x`producing a clear and understandable pattern description

1.3 Objectives

One of the terrible chronic health issues that has preventable repercussions is diabetes. High blood glucose levels brought on by inadequate insulin synthesis would be the main factors. In terms of the number of people with type 2 diabetes, there are roughly 12 million men and 11.5 million women. The best way to improve quality of life is to manage one's own diabetes, which requires more patient assistance and education.

One of the worst illnesses there is is diabetes. Diabetes can be brought on by being overweight, having high blood sugar, and other things. It changes how the hormone insulin works, which makes crabs' metabolisms erratic.

Although technology has advanced and new diabetes therapies have been discovered, most people find the obstacles of self-comprehension to be the most difficult. It necessitates individualised patient self-management, which involves tracking blood glucose levels, upholding a balanced diet, taking medicine, and working out frequently. Self-management behaviours have high non-compliance patterns; this might typically be attributed to the changes that are experienced in the patient's daily life.

The patients are typically driven to reach their goals and adopt their new lifestyle in order to live a long life that enables them to manage their diabetes. This is how they adapt and withstand such changes. Achieving self-management objectives depends heavily on the help, support, and feedback received. The patients can hold to the shared changes and self-awareness in these organisations where peer diabetic people support groups are present, which is a vital resource.

The main goals of this course are to introduce students to the fundamental ideas and methods used in the processing of medical images and to pique their interest in further research and study in the field. to provide computer techniques and algorithms for quantifying and evaluating biological data

1.4 Methodology

The outlined have been used in this.

1. Import the necessary libraries and the dataset related to diabetes.
2. Preprocess the data in step two to fill in any gaps.
3. To divide the perform 80% scaling in step three.

The fourth phase involves selecting a machine learning method like;

Step 1: Using the developing a model classifier machine learning technique.

Step 2: Assess the classifier model using the aforementioned machine learning technique using the test set.Perform a comparison analysis of each classifier's test performance findings.

Step 3: After doing an analysis based on several factors, select the algorithm that performs the best. a possible diagram.

Step 4: Using ada boost an ensemble learning algorithm, and adaBoost classifier is created, using the random forest as the base estimator and with 50 estimators and a learning rate of 1.

We propose an enhanced accuracy .In this model, we utilised a variety of machine learning techniques, such as grouping, regression, and classification. The main goal is to improve accuracy using the .In Figure 1, the suggested framework is displayed.

Each phase's description is provided;

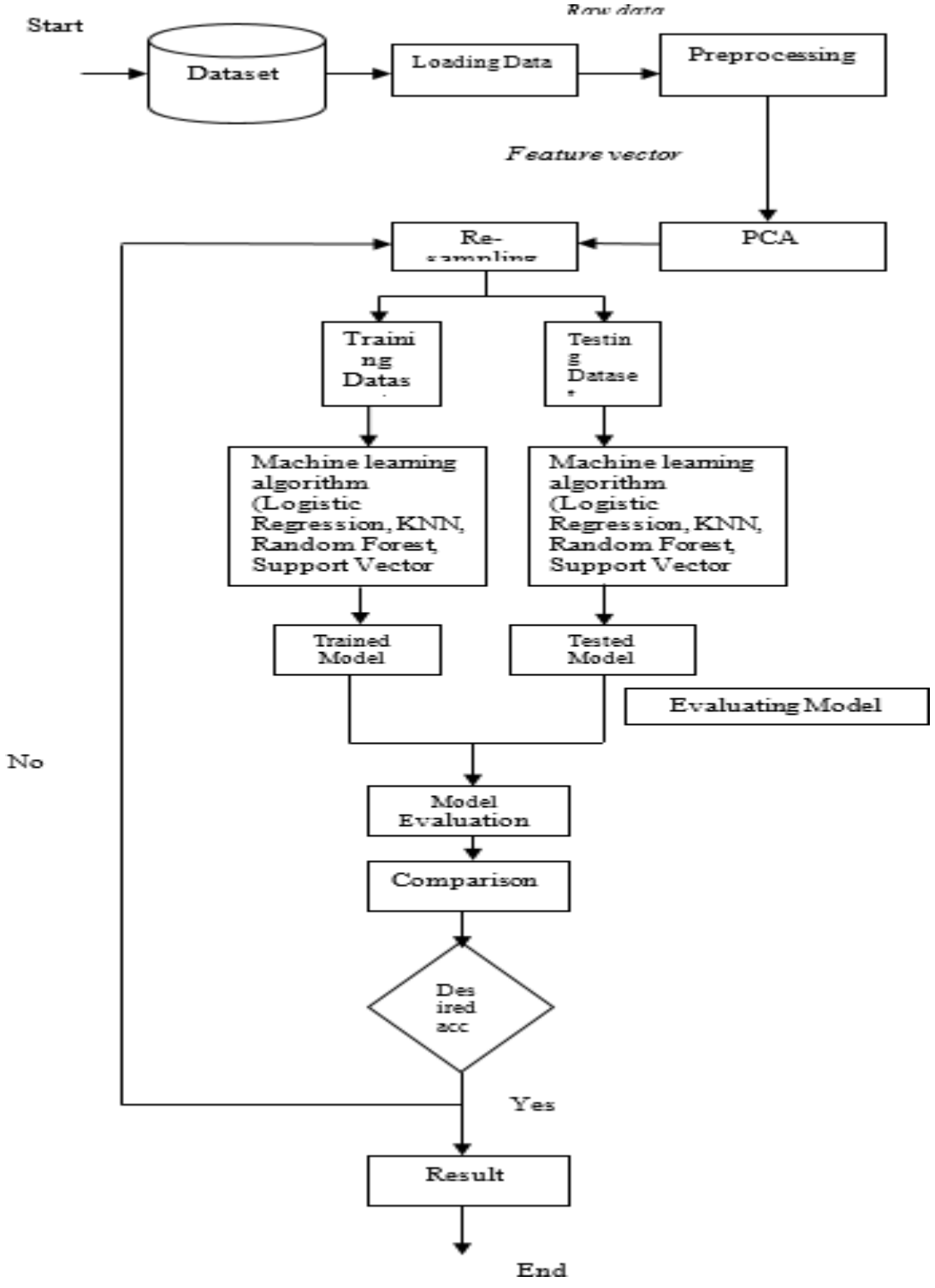


Fig 1.1 Phases of Machine learning

1. **Data selection** is the act of choosing the most pertinent data from a given area in order to derive values that are instructive and aid in learning.

   Eight variables from the PIMA diabetes dataset are utilised to predict diabetes at an

early stage. The UCI repository provided the data for this dataset.

2. **Data pre processing** transforming unorganised data into logical configurations. Data Cleansing, Data Integration, is all included.
3. **Principle Component Analysis for Feature Extraction:** Feature extraction from the dataset to identify the best set of attributes. The PCA's collection of attributes are known as feature vectors.

   We will profit from feature reduction or dimensionality reduction by having less computational.

4. **Re-samplings Filter:** The pre-processed dataset is subjected to the supervised Resample filter. Re-sampling is a set of techniques used to reconstitute training sets and validation sets of your sample data sets. Boot strapping re sampling was used in this study to increase accuracy.

   422 million people globally, especially in low- and middle-income countries, have diabetes. And this number could rise to 490 billion by the year 2030.

   India currently has over 100 million people, however there are 40 million diabetes there. For instance, early management and control of diabetes can prevent death.

   In this study, several diabetes related indicators are used to assess diabetes prediction. To this end, the Pima Indian diabetes database uses a variety of machine learning classification methods and ensembles to predict diabetes..

   An purposefull training method for computers and other machines is called machine learning. Different machine learning algorithms effectively. These findings may aid in the diagnosis of diabetes.

   Many machine learning techniques are capable of making predictions, but selecting the appropriate method can be challenging. Therefore, we use well-known classification and ensemble algorithms on the dataset for this aim to make prediction.

1.4.1 Tools and Libraries

Languages and Libraries Required :

Python 3.10.0: Python is an object-oriented, modular, threaded, and memory-allocation-automatic programming language. A high-level, all-purpose programming language is Python. Its design philosophy places a strong emphasis on code readability through the use of off-side rule-based considerable indentation.
Python is dynamically typed and makes use of garbage collection. Procedural, object-oriented, functional, and structured programming (particularly this one) are just a few of the paradigms it supports. It is frequently described to as a "battery-on-board" language due to its robust standard library.

Its advantages are widely recognised. It has a built-in structure and is open-source, portable, expandable, and simple to use.

Flask: Flask is a Python-based microweb framework.It is regarded as a micro-framework since it doesn't need any particular tools or libraries. It lacks any elements, such as a database abstraction layer, form validation, or other elements, where common operations are carried out by third-party libraries that already exist.

Basic terms in for a flask.

1. **WSGI** Python online application development now adheres to the WSGI standard, or online Server Gateway Interface. A uniform interface between the web server and web applications is described by the WSGI protocol.
2. **Werkzeug** It is a WSGI toolkit that carries out utility operations like requests and response objects. As a result, a web framework may be built on top of it. Werkzeug serves as one of the foundations for the Flask framework.
3. **jinja2** jinja2 is a well-liked Python templating engine. A web templating system combines a template with a particular data source to produce dynamic web pages.

Numpy: Numerical Python is referred to as Numpy. It is an opensource library for the Python programming language, as implied by the name. Large, multi-dimensional matrices and arrays are now supported by Numpy, along with a vast array of cutting-edge mathematical operations that can be used on them. Its goal is to simplify the transformation of challenging functions for you or the

calculation of certain data analysis. The primary benefit of Numpy is its speed. Compared to utilising the built-in Python functions, it is significantly faster. As an illustration, it enables you to quickly determine the mean and median of a dataframe using a single line of code for each. It contains various features including these important ones:

1. A robust multi-dimensional object of array.
2. C/C++ integrating tools for Fortran code.
3. Ability to create random numbers and useful linear algebra.
4. Size is fixed and multi-dimensional.
5. Fast array loops.
6. It uses primitive data types to hold numbers. A primitive data type simply denotes that the information is kept in byte form.

Pandas : Pandas is an open-source Python package that provides you with a very useful collection of tools for conducting data analysis. To improve your machine learning skills, you must learn Pandas. Data science, machine learning, and other applications employ it in addition to data analysis. Simply said, you will require Pandas if it uses data. It can speed up the loading, preparing, merging, joining, reorganising, processing, and changing of data. As previously mentioned, Pandas is a free and open-source package that makes it simple to utilise data structures and data analysis tools for Python. The core of Pandas are Data-Frame objects.

Pandas are highly effective methods of presenting data. This facilitates better understanding and data analysis. Better outcomes for data science initiatives are assisted by simpler data representation. Pandas are extremely powerful animals. They give you access to a wide range of essential tools and guidance that are used to quickly analyse your data. Pandas can be used for a number of tasks, such as filtering your data based on specific criteria or segmenting and splitting the data based on preferences.

.

Sklearn: Scikit-learn (formerly known as scikits.learn and also known as sklearn) is a free machine learning library for the Python programming language. Among the clustering, regression, and classification methods it provides are support-vector machines, random forests, gradient boosting, k-means, and DBSCAN. It is also designed to operate with Python's scientific and numerical libraries, NumPy and SciPy. In addition to substantially utilizing NumPy for high-performance array and linear algebra operations, Scikit-learn is primarily developed in Python.

Here we will discuss some of them.

1. Unlabeled data can be grouped using the sklearn clustering tool.

2. Several supervised learning results can be combined into one prediction using Ensemble feature.
3. Scikit-learn can be used to test the accuracy and the validity of supervised models Unobserved data.
4. You can extract features from text and images using scikit-learn.
5. Factoring, cluster analysis, principal component analysis, and unsupervised neural networks are all included in this collection of algorithms.

Pickle: For the purpose of serialising and deserializing a Python object structure, the pickle module supports binary protocols. Pickling is the process of converting a Python object hierarchy into a byte stream, while unpickling is the process of converting a byte stream (from a binary file or an object that appears to be made of bytes) back to an object hierarchy. However, for the purposes of this article, "pickling" and "unpickling" will be used in place of other terms such "serialisation," "marshalling," "flattening," and 1.

Javascript: JavaScript (JS) is a compact, just-in-time compiled or interpreted computer language with first-class functions. Despite the fact that PHP is most frequently associated with Web page scripting.A few of JavaScript's dynamic features include source-code recovery, object inspection via the for...in and Object utilities, variable parameter lists, function variables, runtime object generation, and object parameter lists. When using toString(), you can access the source text that JavaScript functions keep.

- Asynchronous JavaScript - asynchronous its significance, an it can be used to handle potentially blocking processes, such obtaining resources from a server, in an efficient manner.
- Client-side web APIs - Explores APIs and uses some of the most prevalent APIs you'll encounter when working on development projects.

HTML : HyperText Markup Language, or HTML. Using the markup language, it is used to create web pages. The acronym HTML stands for Hypertext Markup Language. Markup language defines the text document inside the tag that defines the structure of web pages, and hypertext defines the link between the web sites.

Web pages that are shown on the World Wide Web (www) are structured using HTML. It includes Tags and Attributes that are applied to the web pages' design. Additionally, we can use hyperlinks to connect several pages.

CSS: Web pages are styled using CSS (Cascading Style Sheets). CSS is the slang term for Cascading Style Sheets. Utilising this will make the process of producing decent web pages simpler.

You can use it to apply styles to websites. More significantly, it makes it possible for you to do this without relying on the HTML that each web page is composed of.

Technical Requirements

Google Collaboratory:

Collaboratory, or “Colab is the abbreviation for a product made by Google Research. Colab is particularly well suited to machine learning, data analysis, and education and allows anyone to create and execute arbitrary python code through the web.

In order to allow machine learning with cloud storage, Google has created Collaboratory, a web IDE for Python. This internal product had a very quiet public release in late 2017 and is poised to significantly alter the field of machine learning, artificial intelligence, and data science work.

Some of the impressive features of Google Colab notebook are :

1. Our notebook can be saved to our individual Google Drive.
2. We can import a specific notebook from Google Drive as well.
3. It offers cost-free cloud solutions.
4. Any notebook from github can be directly imported or published.
5. It is simple to combine with TensorFlow, PyTorch, or OpenCV.
6. From a notebook, we can run commands on the terminal.
7. Without any local setup, we can write and run Python3 codes

Visual Studio Code:

The Electron Framework is used by Microsoft's Visual Studio Code, also known as VS Code, a source-code editor for Windows, Linux, and macOS.Using the Electron Framework, Microsoft produced the source-code editor Visual Studio Code, also known as VS Code, for Windows, Linux, and macOS. Among the features are debugging assistance, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git. By adding extensions, users can change the theme, keyboard shortcuts, preferences, and add functionality. Many distinct programming languages, including C, C#, C++, Fortran, Go, Java, JavaScript, Node.js, Python, and Rust, are

supported by the source-code editor Visual Studio Code.Its base is the Electron framework, which is used to build Node.js web programmes that make use of the Blink layout engine. Visual Studio Code additionally makes use of the Azure DevOps "Monaco" editing component. Users are given the option to open one or more directories, which can then be saved in workspaces for use at a later time in place of a project system. Since it is language-neutral, it can serve as a code editor for any language. It supports many different programming languages, each of which has a unique set of features.

Among the features are assistance with debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git..

1.4.2 Techniques for Machine Learning:

A.Random Forest:

By creating a decision tree during training, the ensemble learning approach for classification and regression creates the class. For preferred trees' propensity to overstuff their preparation set, uncommon choice woods are acceptable.

The values of a random vector sampled separately from every other tree in the forest using the same distribution serve as the basis for the predictor variables for each tree in the forest.

By identifying Random Forests finds a solution to the high variation and high bias problem. They also offer a way for calculating error rates (Out of the Bag error).

Several machine learning models, including linear given sample, may become distorted. The performance or accuracy of the model are unaffected by these extreme or outlier numbers. The RF Algorithm is used to solve and get around this problem.

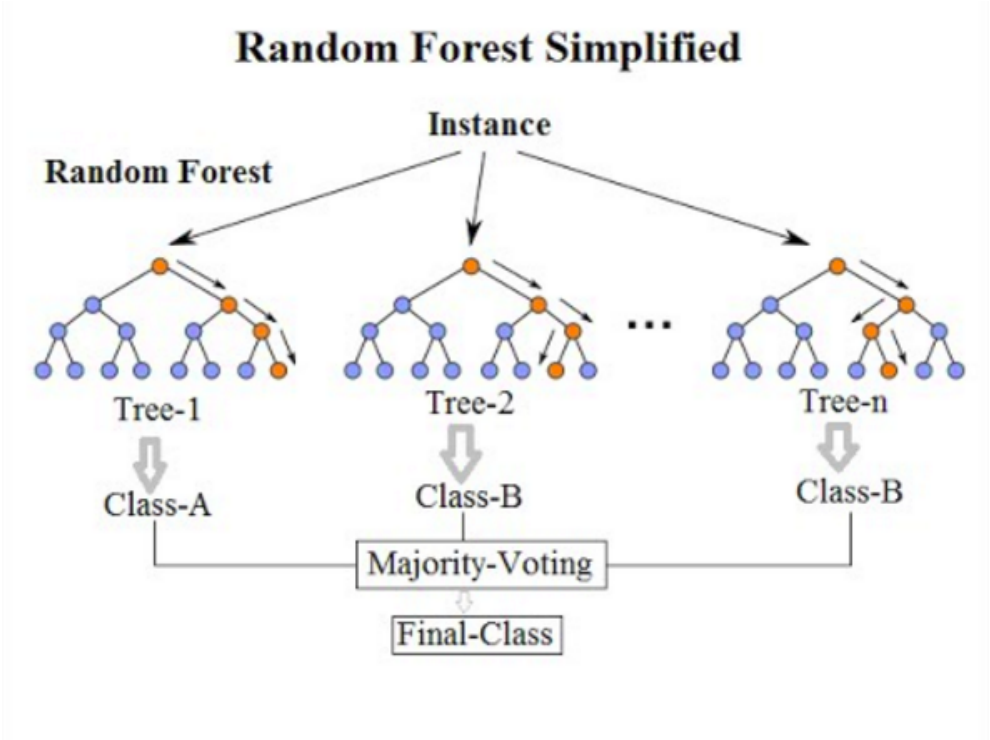


Fig 1.2 Random Forest

B. Support Vector Machine:

The objective of the support vector machine algorithm is to locate a hyperplane in an N-dimensional space (N is the number of features) that categorises the data points precisely.

Finding a hyperplane in an N-dimensional space (where N is the number of features) that accurately classifies the data points is the goal of the support vector machine algorithm.

Decision boundaries known as hyperplanes assist in categorising the data points.The data points that are located on each side of the hyperplane can be assigned to various classes.

Additionally, the amount of features affects how big the hyperplane is.

The goal of the SVM method is to increase the distance between the hyperplane and the data points. Hinged loss is the loss function that aids in maximising the margin.

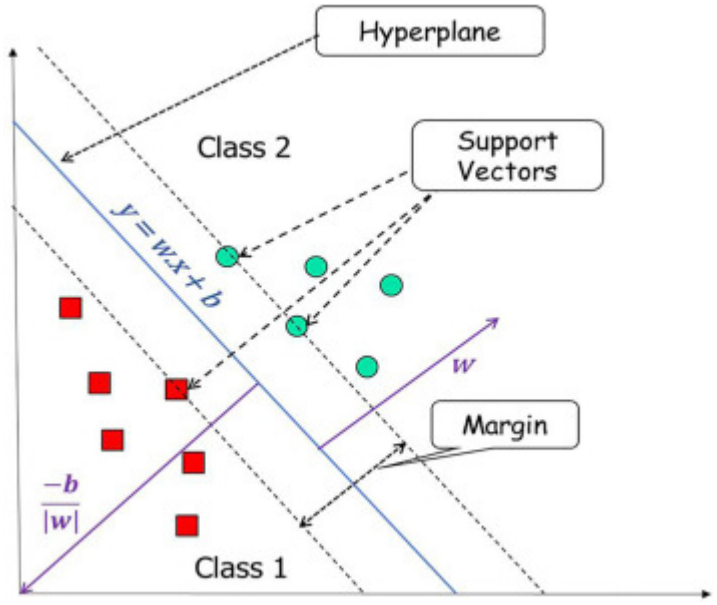


Fig 1.3 Support Vector Machine

C. Decision Tree :

The most effective and well-liked technique for categorization and prediction is the decision tree. Each internal node in a decision tree represents a test on an attribute, each branch a test result, and each leaf node (terminal node) a class label. Decision trees are a sort of tree structure that mimics flowcharts.

A tree can be "learned" from the source data by dividing it into subgroups based on an attribute value test. Repeating this approach for each derived subset is known as recursive partitioning. When the predictions are no longer improved by splitting or when the target variable in the subset at each node equals zero, the recursion comes to an end. Exploratory knowledge discovery can make use of decision tree classifiers since they can be developed without the need for subject-matter expertise or parameter adjustment. High-dimensional data is manageable for decision trees. Decision trees' classifiers typically have good accuracy. A typical inductive strategy to learn about classification is decision tree induction.

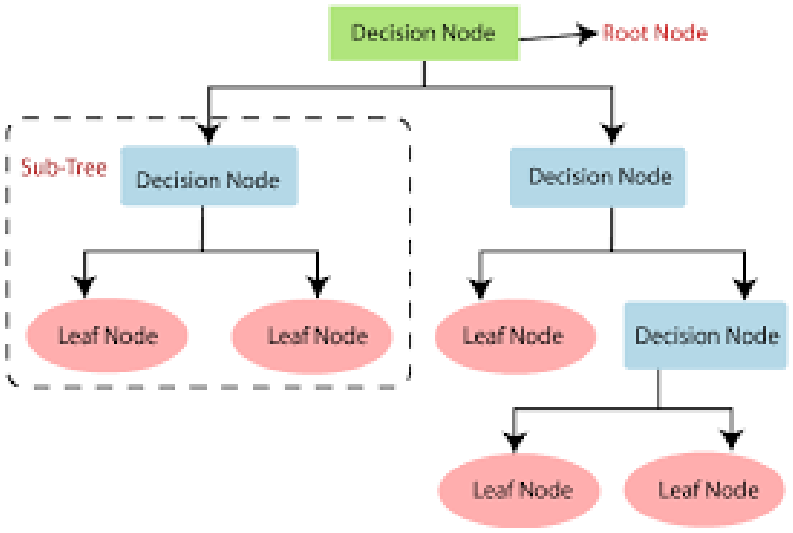


Fig 1.4 Decision tree

D. Ada Boost:

By combining several weak classifiers, the ensemble modelling technique known as "boosting" aims to create a powerful classifier. In order to construct a model, weak models are used in series. Initially, a model is built using the training data set. Then, in an effort to address the shortcomings of the first model, a second model is developed. This process is repeated until either the maximum number of models has been added or the entire training data set has been correctly forecasted.

AdaBoost was the name of the very first binary classification boosting algorithm to be devised. Several "weak classifiers" are blended into a single "strong classifier" using the boosting technique known as "AdaBoost," which stands for "Adaptive Boosting."

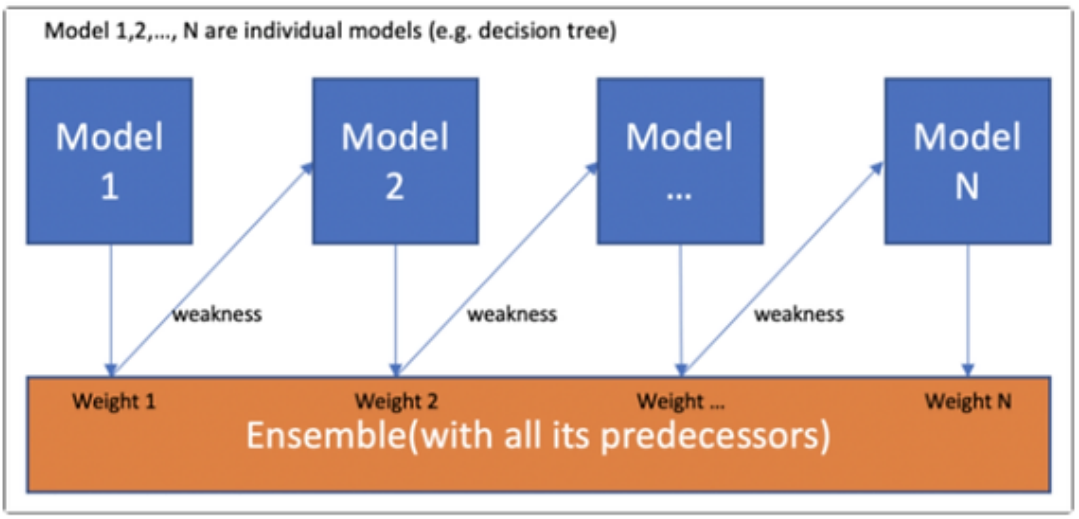


Fig 1.5 AdaBoost Techniques

Automate pipeline with CI/CD:

With the help of a set of procedures, resources, and guiding cultural concepts referred to as "DevOps," software development and IT teams can automate and integrate their processes.
It places a focus on technology automation, cross-team communication, and team empowerment.

A continuous integration and continuous deployment (CI/CD) pipeline is required for the delivery of a new version of software. CI/CD pipelines seek to improve software delivery throughout the whole software development life cycle through automation.

Automating CI/CD during the development, testing, production, and monitoring phases of the software development lifecycle enables organisations to deliver higher quality code more quickly. Every stage of a CI/CD pipeline can be completed manually, however automation is where the full value of CI/CD pipelines is realised.

Making it possible to automatically test and deploy software there are the two DevOps components, such as:

1) Constant integration - is the process of automatically incorporating code changes into software projects. As builds and tests are run, it enables developers to often merge code changes into a single repository. This shortens the time it takes to validate and deliver new software updates while also helping DevOps teams address defects more quickly.

2) Continuous deployment - By automatically distributing code updates to a testing/production environment, continuous delivery enhances continuous integration. The automated builds, tests, and deployments are coordinated as part of a single release workflow, which is known as a continuous delivery pipeline.

3) Continuous feedback - DevOps teams ought to assess each release and produce reports to enhance subsequent releases. Teams may enhance their procedures and utilise client feedback to improve the upcoming release by collecting ongoing feedback.

Data scientists are helped with feature engineering, model design, and hyper parameters by the automated CI/CD. Data scientists may develop deployment environments, submit code to Github, build models, and more.

CI/CD methodology:

Continuous Integration/Continuous Delivery (CI/CD) has always been, and still is, the purview of DevOps professionals.

Code pushes to the repository, builds are triggered, builds are tested, and builds are deployed to the production environment make up a typical CI/CD pipeline.

Building CI/CD pipelines is entirely customised based on the needs and requirements, and it may involve numerous phases, jobs, and other intricate details.

Creating a CI/CD pipeline typically involves the following step:

1. **Code**: into the repository
2. **Build**: Build is started and then made available in a test environment.
3. **Test**: There are automated tests run.
4. **Deploy**: Stage and production environments get code deployments.

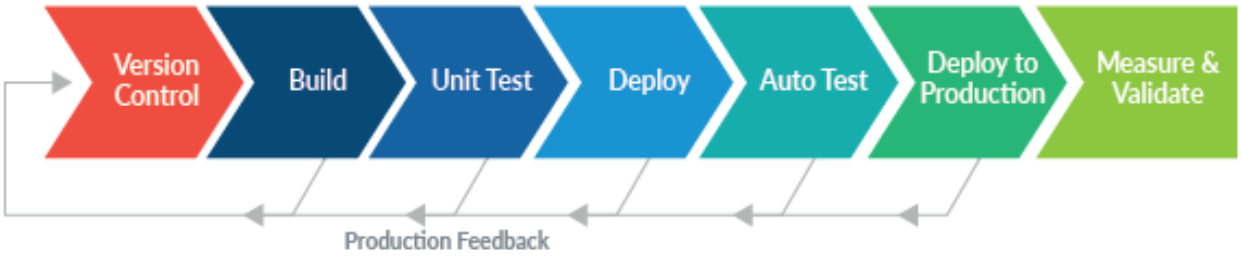


Fig 1.6 DevOps deployment phases

Steps to build a CI/CD pipeline :

1. Source code management tools like Git, Github, etc.
2. When an addition or deletion happens, the repository is automatically updated and deployed in a production environment.

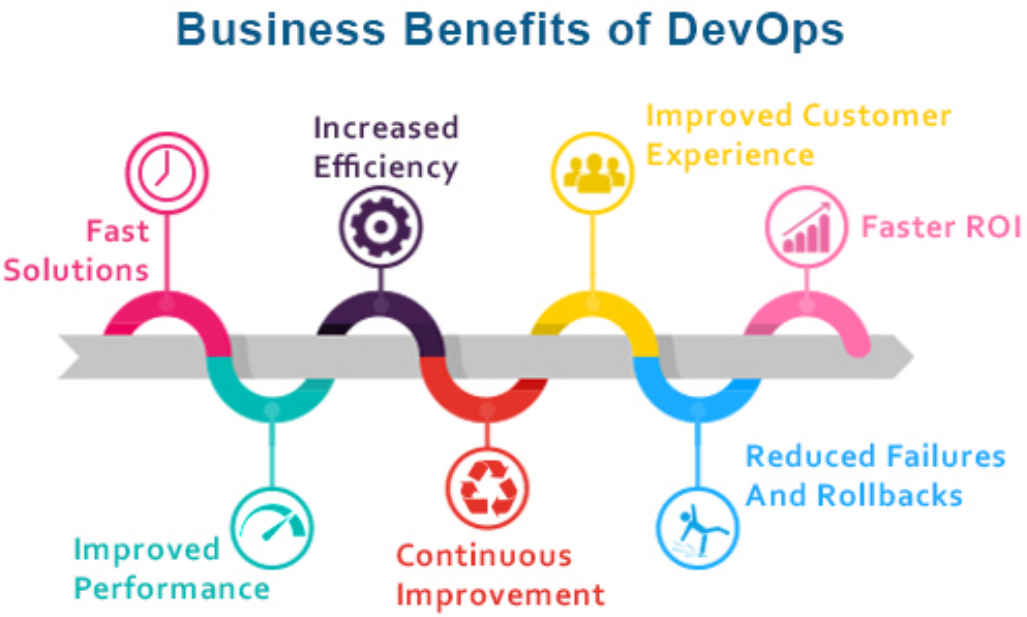
3. Put the updated code into action.

4. Check and evaluate the code.

5. Create the container image.

6. The repository is pushed with the container image.

Continuous delivery: Website testing is the result of the Continuous Integration pipeline outputs implemented . One of the most well-known open source tools for continuous deployment is called github Action. It is developed in Java and integrates with SCM solutions like Github.



The components of the Continuous Deployments:-

1. Production environment: The website code must first be deployed in the environment where it will be tested. The requisite decencies are already present in this context.

2. registry storage: Following testing in a real-world setting, the code must be kept in the model registry.

3. Automated trigger: the trigger will instantly go off, starting a fresh build and pushing the output of the updated code into the production environment.

4. Performance monitoring: Numerous elements are taken into account when measuring performance, such as the effectiveness and the quantity of resources used.

Chapter 2 Literature Survey

2.1 OBJECTIVE

Diabetes is a disease that affects how effectively your body uses food as fuel. Most of the food you consume is converted into glucose, sometimes known as sugar, and then released into your bloodstream. Your pancreas releases insulin when your blood sugar levels increase. Diabetes can elevate blood sugar levels, raising the risk of catastrophic side effects including heart attack and stroke, if it is not carefully and regularly managed. The management of diabetes depends on early diabetes diagnosis.

Multiple tests on a patient make it difficult for doctors to maintain track of various variables during the diagnosis process, which can lead to erroneous results and make detection very difficult. Because of the majority of technological improvements, notably AI calculations, the medical services industry can predict illnesses quickly and accurately. Machine learning promises to increase the accuracy of diagnosing and perceiving diseases. Here, the computers learn and grow in intellect, which gives them the capacity to think. A variety of machine learning techniques are used to categorise the data sets.

2.2 PREVIOUS PAPER

- The study used a dataset with 9 columns and 768 rows, which included 500 non-diabetics and 268 diabetics. One outcome goal variable and a number of medical predictor variables are combined by the study's authors. Predictor variables include the patient's BMI, insulin level, age, number of prior pregnancies, and other factors. All patients are girls aged 21 and older who are of Pima Indian ancestry. Over the course of this inquiry, no experimental testing was done on the dataset in question.
- In order to compare the application of Artificial Neural Networks and Bayesian Networks, two machine learning methods—diabetes and cardiovascular diseases are classified, the research proposes a conceptual framework. Based on the knowledge of researchers from articles [3, 4] that examined machine learning methods but in different fields of research,

- Six characteristics—diabetes result, pregnancy, blood sugar, blood pressure, skin thickness, BMI, and age—from 203 females between the ages of 18 and 77 make up this dataset. The GlucoLeader Enhance blood sugar metre was used in this study to measure blood glucose levels. OMRON

HEM-7156T and digital LCD body fat calliper equipment were used to measure the participants' blood pressure and skin thickness, respectively.

- The authors of the study proposed using the Random Forest algorithms, Decision Tree Algorithm and SVM, to represent actual Diabetes Mellitus prediction. A comparison of these algorithms was done using their respective Metric Measures, Accuracy, Precision, Sensitivity. Based on their accuracy metrics, the Random Forest Algorithm and Support Vector Machine are equally good as a consequence of the research work.Decision trees offered good accuracy in most studies.y

- The training set for the supervised classifier machine learning methods described in this paper[18] was created by removing attributes with little or no relevance to predicting diabetes. Only the traits that were scored highest, given more weight, and more likely to predict the onset of diabetes were taken into consideration when this was done using the chi-squared test. It was observed that the Neural Networks method produced the highest accurate results on this training set. Additionally, it employs the prediction and classification method. This increases the disease prediction's accuracy.

- For both classification and regression applications, Random Forest algorithms are a type of ensemble learning method that is extensively used. The accuracy level is greater when compared to other algorithms. The outcomes showed that the forecasting system can accurately, successfully, and most importantly, swiftly predict the diabetic disease. For predicting diabetes, the suggested model produces the best results.

- The best hyperplane between the two classes must be discovered by the SVM algorithm in order to correctly categorise all the data points. The hyperplane that increases the margin between the two classes is the ideal one. Support Vectors are the data points, also known as vectors, that are closest to the hyperplane, giving the algorithm the name Support Vector Machines.

- The training dataset's Support Vectors include the most crucial data points. The dividing hyperplane would move to a different location if these data points were eliminated from the training dataset. Additionally, they are the hardest to categorise data points.Depending on which side of that border the data lies, the SVM algorithm assigns a diabetic result of 1/0.

- In order to increase accuracy, For the type 2 diabetes mellitus dataset, D. Jeevanandhini, E. Gokul Raj, V. Dinesh Kumar, and N. Sasipriyaa [12] performed a performance analysis..Here, they used eight key attributes to compare the four prediction models.According to this study's findings, the Support Vector Machine (SVM) classifier performs better than the other three classifiers, with an accuracy of 77.82%. On Pima datasets, Dr. K. Thangadurai and N. Nandhin[13] utilised a variety of data mining algorithms. It has been discovered that the genetic algorithm performs better than five data mining algorithms. In this paper[49], we compared a few different classification algorithms using the Matlab tool for analysis.Following a comparison, we found that neural network methods are more precise and had a lower error rate.

- The findings of this study indicate that, in contrast to the earlier findings, the modified spline SSVM is efficient in detecting the diagnosis of diabetes. The research of Drs. B.L. Shivkumar and S. Thiyagarajan suggests a useful machine learning technique to categorise type dm patients. This machine learning approach for classification will find the ideal hyperplane that splits the various classes.

- Users will have access to more accurate and useful information overall thanks to the trending technologies and potent algorithms. Machine learning models may be trained to recognise the patterns and characteristics of diabetic cases on the parameter of insulin,blood pressure,pregnancy rate,glucose,skin thickness,bmi,diabetic pedigree function, and age.

- To accurately identify diabetes in patients who may have it, a battery of tests and examinations must be completed. Diabetes has a far higher economic cost than it does directly medical expenditures to the healthcare system since it diminishes quality of life and decreases productivity at work. Through data analysis and visualisation, we were also able to draw some conclusions from the data. We were able to accurately predict whether the patients in the dataset have diabetes

or not by creating a machine learning model (random forest, best one).

Generally speaking, the main objective of hate speech identification by machine learning is to ensure that users can make informed decisions based on real and reliable assessments.

1. **Obtaining data** The initial step is to collect information from dataset, website such as kaggle and other, online platforms , and websites were used for collection of raw data. These data sources are able to provide information that is labelled and includes both diabetic and non diabetic patient's data. It is essential to ensure that the data reliably and diversely covers various categories,symptoms, and thresh hold values.

2. **Data pre-processing**: After the data have been gathered, they must be pre-processed to eliminate extraneous zero values and useless data. Skin thinning and For example, the Body Mass Index (BMI) cannot be zero. The zero value has been replaced by the comparable mean value. The training and test datasets were split into two groups using holdout validation: In training, 80% of the data were used, and in testing, 20%.Using the pre-processed data, a numerical image resembling a bag is created.

3. **Feature selection:** The performance of the model must be improved, and the input's dimensionality must be decreased, by selecting the appropriate features. The characteristics that are most important for determining whether cases are diabetic or not are chosen in this step...

4. **Model selection:** It can be classified using machine learning models such random forests, decision trees and support vector machines. The data's characteristics and the essential performance requirements have an impact on the model selection. Either an unsupervised learning method, where the model finds patterns in the data without any labels, or a supervised learning strategy, where labelled data is utilised to train the model, can be used to teach the models.

5. **Model evaluation:** In order to ensure that the model is generalizable, it is crucial to assess the model's performance after training using metrics such as insulin, blood pressure, pregnancy rate, glucose, skin thickness, bmi, and diabetic pedigree function.

6. **Model Compression**: Out of all the models utilised, the best one is selected.

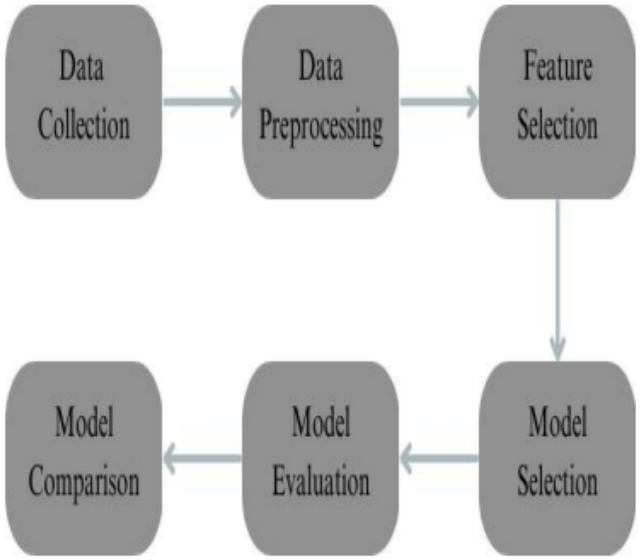


Fig 2.1 Phases of model training

7.**Website Deployment** :The model was made cloud-based using CI/CD pipeline, tried pushing binary values, and predicted the results on the PIMA Dataset using machine learning algorithm pipelines were used for the project's final deployment, presenting the Frontend for projecting results using Flask Technology and hosting of the same will be done on RENDER and implementation of CI/CD pipeline through Gitlab..

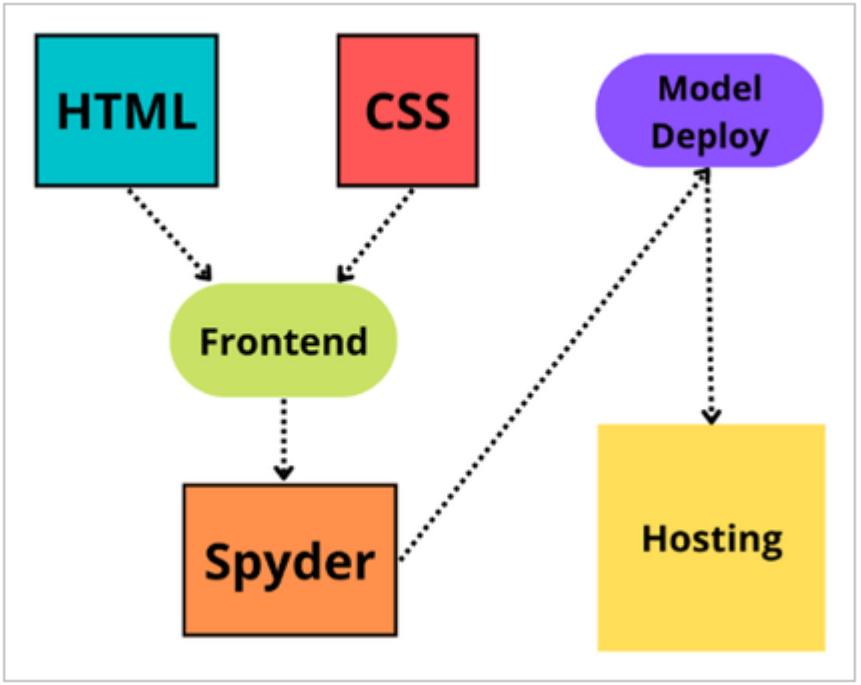


Fig 2.2 Deployment setup

Data Set:

On Kaggle, you may access the dataset that was initially taken from the Pima Indians Diabetes Database. It comprises of one target variable and a number of medical analyst variables. Predicting whether the patient has diabetes or not is the dataset's main goalThe outcome is the only dependent variable in the dataset; all other factors are independent. The patient's BMI, insulin level, age, and other factors are independent variables ,EDA of the same is mentioned in the given fig:

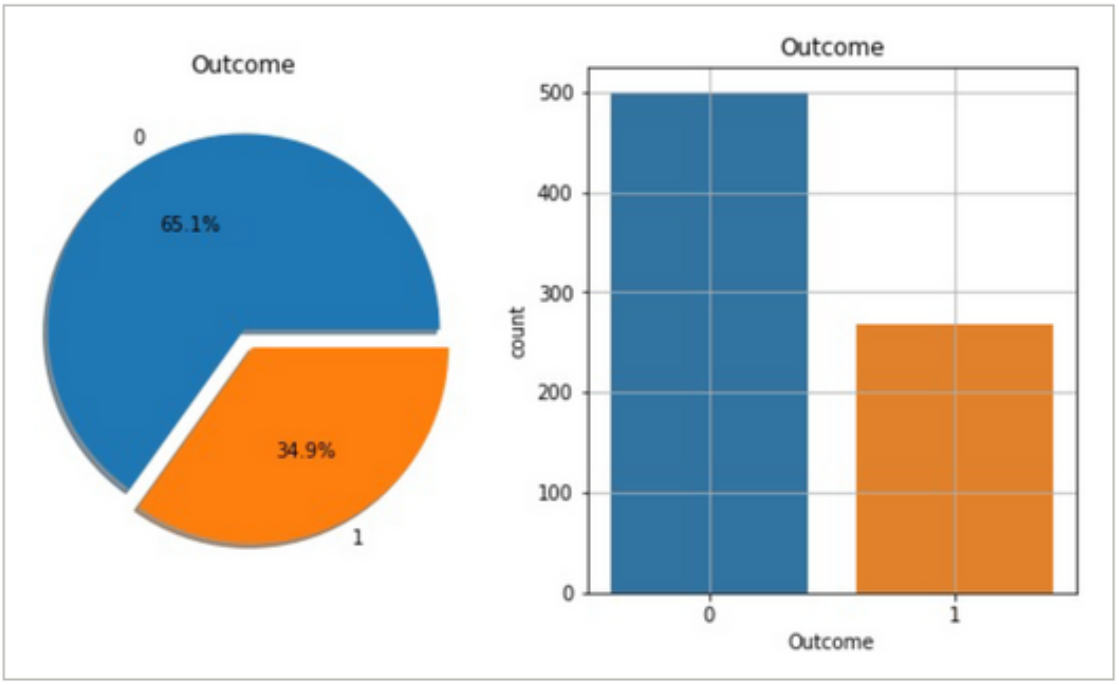


Fig 2.3 patient's BMI, insulin level, age, and other factors are independent variables ,EDA

Chapter 3- System Development

3.1 Software Requirements:

- Python
- Python Crypto Package
- Flask Web Development Micro-framework
- VS Code
- PyCharm Community Edition
- AWS (Amazon Web Services)
- Git / GitHub

3.2 Hardware Requirements:

- 4 GB or more of random access memory.
- Central Processing Unit : Processor running at 2.4 GHz or higher
- GNU/Linux is the operating system (OS) (Ubuntu & Manjaro)

3.3 Analytics:

The fundamental goal of this project is to offer a secure ,accurate and feasible solution to the normal people to predict diabetes and prevent early stage of disease.

it is challenging for medical professionals to diagnose diabetes mellitus early and accurately, especially in the disease's early stages. By employing artificial intelligence and machine learning techniques as a guide, they may grasp the fundamentals of this condition and reduce their effort. Numerous experiments have been done to automatically predict diabetes using machine learning and ensemble methodologies.

This study offers a distinctive dataset of 203 samples for diabetes mellitus, which is a noteworthy contribution. This private dataset, referred to in this research as the "RTML dataset," was collected from female Rownak Textile Mills Ltd. employees in Dhaka, Bangladesh. Six characteristics—pregnancy, glucose, blood pressure, skin thickness, BMI, age, and the end result of diabetes—were gathered from 203 individuals.

The design of the suggested autonomous diabetes prediction system and the use of several machine learning algorithms are part of our planned effort. Prior to analysis, the dataset was collected and preprocessed to address concerns with unbalanced classes, replace null occurrences with mean values, etc. The dataset was then split into the training set and test set using the holdout validation procedure. Then, using a number of classification methods, the best classification algorithm for this dataset was found. The recommended website is then updated with the most accurate prediction model.

3.3.1. Computational:

After carefully examining the dataset we have divide proposed work into 7 steps as follows:

1. Obtaining data The initial step is to collect information from dataset.

2. Data pre-processing: Data must be pre-processed to weed out extraneous information after it has been obtained.

3. Feature selection: Choosing the right features is essential for improving the model's performance and reducing the dimensionality of the input.

4. Model selection:Selecting models marks a significant step as it contributes to the overall accuracy and also help to improve the result.

5. Model evaluation: To make sure the model is generalizable, it is critical to assess the model's performance using important measures.

6. Model Compression: Out of all the models utilised, the best one is selected.

7.Website Deployment :Finally the model is deployed to see the outcome

3.4 System Architecture

One of the serious problems in the modern world is diabetes. It is a chronic condition that can lead to numerous health issues. It is a category of syndromes that cause excessive blood sugar levels. Diabetes-related chronic hyperglycemia has been linked to long-term damage, organ decline, and organ failure, especially in the eyes, kidneys, nerves, heart, and veins. Machine learning is now being used more and more in the medical field. Building a model that can precisely forecast a patient's chance of developing diabetes is the study's goal.

This paper proposes a novel architecture for predicting diabetes patients using K-means clustering and support vector machines (SVM). The retrieved K-means features are then classified using the SVM classifier. The Pima Indians Diabetes Database is used to evaluate this strategy on a publically available dataset. The dataset used showed a 98.7% accuracy rate.

Mathmatical:

Confusion matrix presented Below, the performance of the suggested strategy has been evaluated:

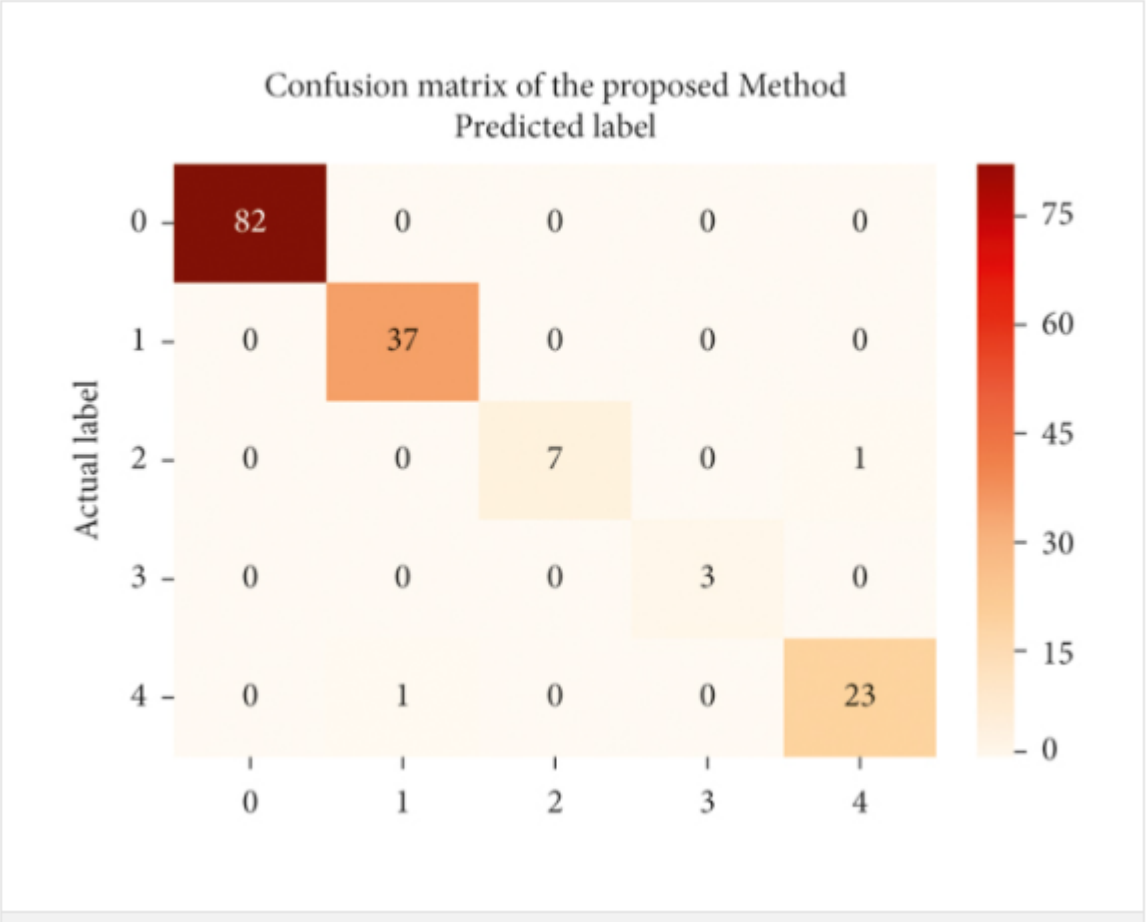


Fig 3.1 Confusion Matrix

Following are the 4 possible outcomes of the confusion matrix: true positive (TP), true negative (TN), false positive (FP), and false negative (FN):

Next, we analyse the proposed model using the following metrics.

Accuracy is measured as the ratio of the total number of correct predictions to the total number of test cases for any given model, and it can be calculated as follows:

Accuracy=TP+TNTP+FP+FN+TN

Precision is the ratio of correctly predicted positive outcomes to all positive outcomes.

Prediction = TP/TP+TN

Recall: Total positive predictions vs. actual positive values is known as recall

Recall = TP/TP+FN

F1-score: F1-score takes precision and recall into account and can be described as follows:

F1_score = 2*(precion*recallprecion+recall

One of the best datasets for studying machine learning methods for diabetes prediction is the Pima Indian Diabetes Dataset (UCI Machine Learning Repository, 1998). based on diagnostic markers such as pregnancy, blood pressure, skin thickness, and insulin, diabetes pedigree, and blood sugar levels.

3.5 Issue Analysis:

To accurately identify diabetes in patients who may have it, a battery of tests and examinations must be completed. These evaluations could include pointless or repeated medical procedures, which causes issues and wastes time and resources. Diabetes has a far higher economic cost than it does directly medical expenditures to the healthcare system since it diminishes quality of life and decreases productivity at work. Lack of a proper diagnosis method, a lack of funding, and a general lack of information are the key contributors to these negative effects. Therefore, a significant economic burden could surely be reduced and the patient's ability to manage their diabetes helped if the sickness could be completely avoided by early identification.

3.6 Solution :

The "Diabetic Prediction" model that has been proposed is a completely computerised method for processing the dataset. Any company or person can use the suggested system because it is an internet tool. We discovered that the Support Vector Machine (SVM) model performs well at predicting diabetes. We constructed the NN model with various hidden layers and epochs, and we found that the NN with two hidden layers offered 88.6% accuracy.One of the most widely utilised model types in VS nowadays is the SVM. Where comparisons are made, it is discovered that they perform well. In our study, we created a system that can accurately predict diabetes. The PIMA DATASET was used to preprocess the data. In the PIMA dataset, we employed one output feature (outcome) and five input characteristics (glucose, BMI, insulin, pregnancy, and age). To predict diabetes, we tested the effectiveness of three different machine learning algorithms, including RF, DT, and SVM, using a variety of metrics. For several metrics like accuracy, precision, recall, and F-measure, all models produce positive outcomes. Every model offered accuracy of at least 70%. For both train/test split approaches, SVM provided accuracy between 77% and 78%.

3.6.1 Design of Problem Statement :

This research work aims to analyze the Diabetes dataset, design, and implement a Diabetes prediction and recommendation system utilizing machine learning classification techniques.And Deploymet of the model using CI/CD pipeline and technology such as flask

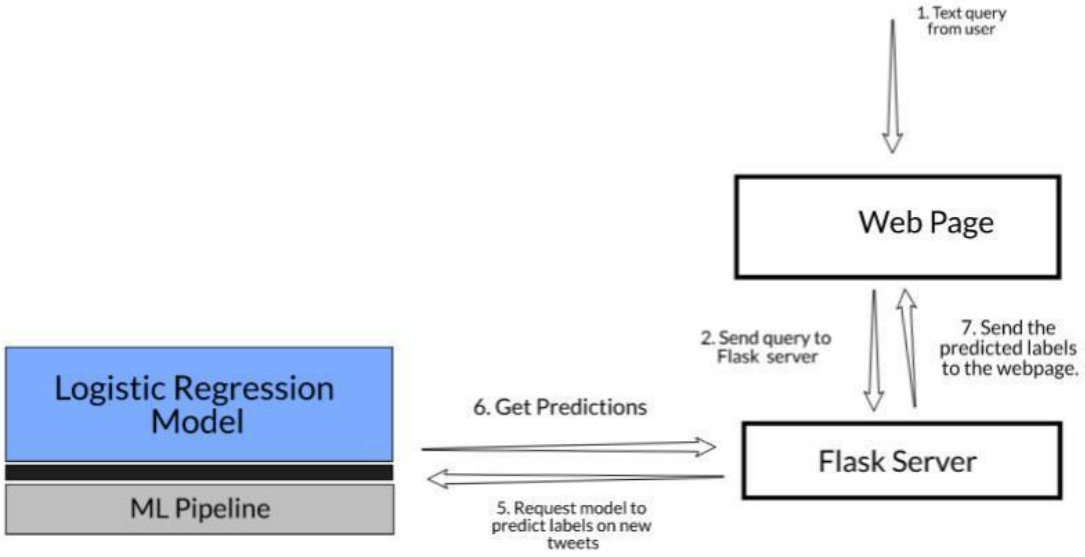


Fig 3.2 Deployment Using Flask

The framework used to run local servers and for web page routing is called Flask. When implementing our machine learning model, we will utilise it so that a typical user may input data or values and obtain the required outcome.

The Flask website increases the user-friendliness of the system, earlier we have to enter commands in google collabs or Jupiter Notebook to get the output of the file and if anyone wants to use it, he needs to have a somewhat understanding of Python and machine learning library so that he can use it and understand the given output by the model.

But by using flask to create a user-friendly website so that anyone without any knowledge of the technology(python, machine learning) can easily use the website and get the output in a more admissible way so that he can easily understand it.

Objective of creating such an application is so that anybody with can check their diabetes and get to know wheater they are prone to get diabetes or not in future if they keep living in a particular lifestyle.

And Checking diabetes becomes as easy as shopping for items from an e-commerce website.

3.6.2 Machine learning automates pipeline with CI/CD:

The main goal of this strategy is to continuously train and test the machine learning model using a pipeline. The automatic model retrain in machine learning is possible with the two elements of DevOps such as:-

1) Constant integration:The build and unit testing phases of the software release process are referred to as continuous integration. Every time a revision is committed, an automatic build and test are started.
2)
3) Continuous deployment CI/CD automation aids data scientists with model architecture, and hyperparameter implementation. This includes generating models, pushing code to Github, creating deployment environments, and many other tasks.

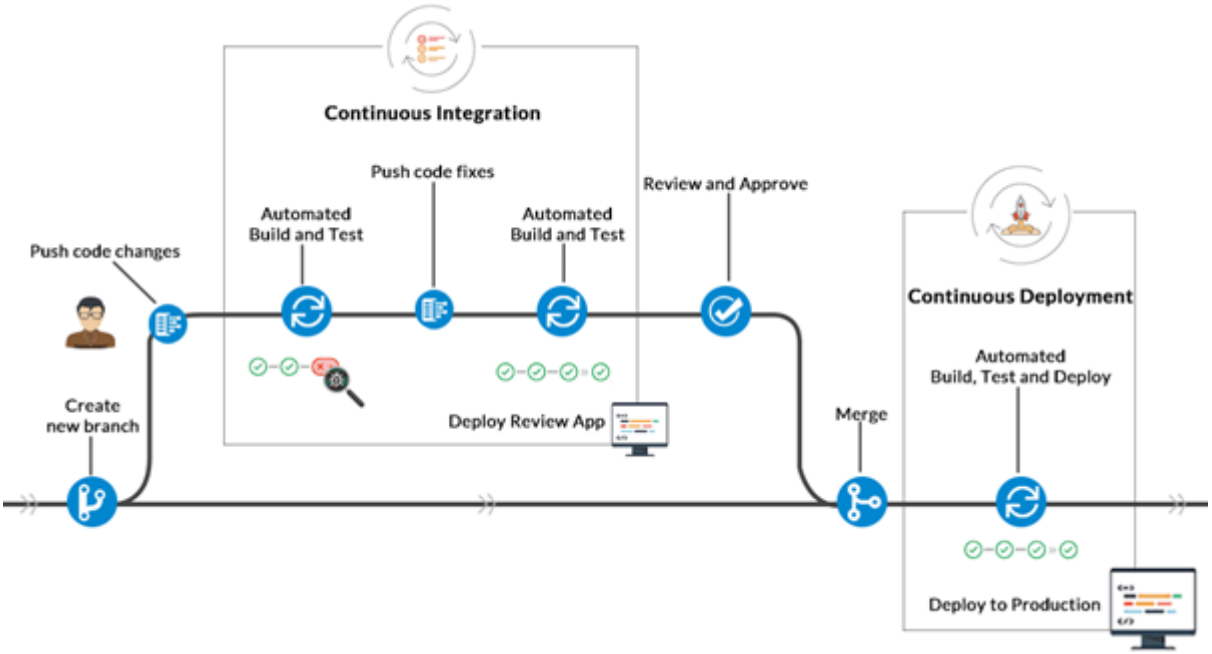


Fig 3.3 CI/CD pipeline deployment

Chapter 4 -Experiment and Results

4.1 Experiments and result

After utilising all of these patient records, we were able to create the best machine learning model, random forest, which reliably predicts whether or not the patients in the dataset have diabetes. We were also able to get some insights from the data through data analysis and visualisation.

The specifics of the dataset utilized inThis section presents the findings of this inquiry. Various classification algorithms and recommended techniques are used to calculate the results. The details are as follows.

All the parameters of the used dataset are as follows:

1. Pregnancies: Represents how many number of times woman was pregnant.
2. Glucose: Plasma glucose levels surpassed 2 hours while undergoing an oral glucose tolerance test.
3. Blood Pressure: Diastolic heart rate (mm Hg).
4. Thickness: Tells the triceps skinfold thickness (mm).
5. Insulin: It tells the 2-hour serum insulin (mu U/ml).
6. BMI: It describes the body mass index of a person.
7. Diabetes Pedigree Function (DPF): DPF is a function that scores the chances of diabetes based on family history.
8. Age: the age of the person (in years).
9. Outcome: This parameter represents the class variable. 0 means nondiabetic, and 1 means diabetic.

Table 2

First five records in the Pima Indians Diabetes Dataset.

Sr. no.	Pregnancies	Glucose	BP	ST	Insulin	BMI	DPF	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

4.2 Details of experiments performed:

For this project, it was first essential to install a Jupyter Notebook in order to implement the code. This project needs certain algorithms and useful diabetes-related datasets that may be downloaded. Machine learning requires 2 TB of hard drive space and 512 GB of SSD in addition to RAM that is 128 GB DDR4 2133 MHz.

A specialised standard library called NumPy supports multi-dimensional array objects, matrices, and various components. These arrays are also subjected to a substantial number of intricate mathematical processes.

Pandas are essential to the project's success. It is a library that is employed in both data cleaning and analysis. Pandas include a number of functions that can be used to modify, organise, visualise, and analyse data. Data can be imported into Pandas from many different file kinds, including SQL, comma-delimited values, Microsoft Excel, JSON, etc. Numerous data manipulation techniques are available, including capabilities for reshaping, combining, selecting, cleaning, and wrangling data.

Data may be plotted using the Python computer language's Matplotlib package. A Python extension for numerical mathematics is called NumPy. For embedding plots in programmes created using multipurpose GUI toolkits like wxPython, Tkinter, Qt, or GTK+, it offers an object-oriented API. Pyplot includes command-style graphs from Matplotlib.

As with MATLAB, it enables matplotlib to function. Every plot function alters a map, among other things by creating a plotting region in the figure, placing some lines there, and labelling the plot. The following need for the project is Seaborn. It is a matplotlib-based Python module for producing statistical visuals that is closely connected with Panda's data structures.

This research work aims to analyze the Diabetes dataset, design, and implement a Diabetes prediction and recommendation system utilizing machine learning classification techniques. The specific objectives of this project work are:

(i) To review existing literature along the area of diabetes diagnosis and prediction.

(ii) Design and Develop a model using machine learning techniques.

(iii) To analyze the Diabetes dataset and use Support Vector Machine and Random forest algorithms to develop a prediction engine.

(iv) To identify and discuss the benefits of the designed system along with effective applications.

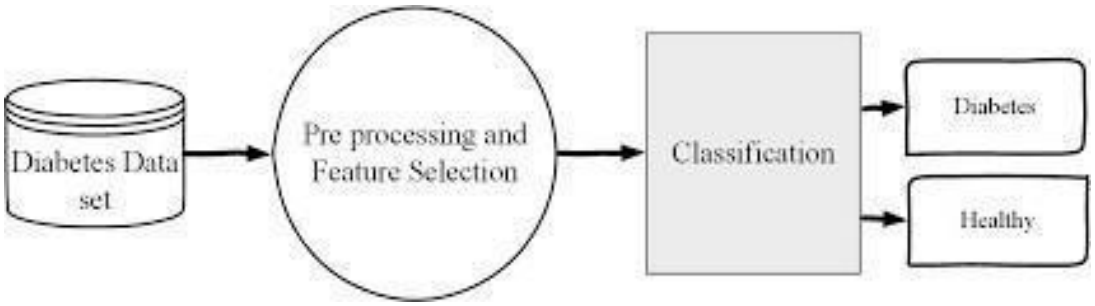


Fig 4.1 Overview of full system deployed

4.3 Method Used:

If diabetes is not consistently and carefully treated, blood sugar levels might rise, increasing the risk of serious side effects including heart attack and stroke. As a result, we decide to forecast using machine learning in Python

Steps involved

1. Installing the Required Libraries
2. Importing the Dataset in the local storage
3. Data Cleaning
4. Exploratory Data Analysis
5. Predicting Unseen data
6. Implementing of Machine Learning Models (supervised or unsupervised)
7. Feature Engineering
8. Deploying the machine learning model on the website.

The framework used to run local servers and for web page routing is called Flask. When implementing our machine learning model, we will utilise it so that a typical user may input data or values and obtain the required outcome.

The Flask website increases the user-friendliness of the system, earlier we have to enter commands in google collabs or Jupiter Notebook to get the output of the file and if anyone wants to use it, he needs to have a somewhat understanding of Python and machine learning library so that he can use it and understand the given output by the model.

But by using Flask to create a user-friendly website so that anyone without any knowledge of the technology(python, machine learning) can easily use the website and get the output in a more admissible way so that he can easily understand it.

Structuring your web application

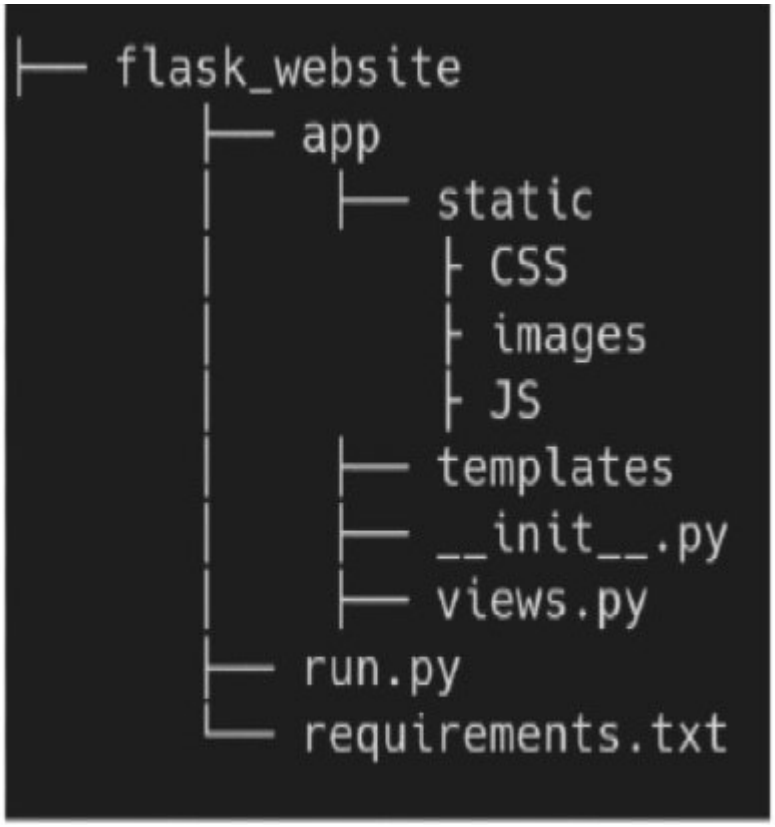


Fig 4.2 Structure of Flask file

The first step in constructing your web application is to make the "**app"** folder, the "**run.py**" file, and the "**requirement.txt**" file.

We may define our Flask application as a package that can be imported in any area of the app that we would need thanks to the flexibility of the "app" folder structure.

Essentially, the run.py file will act as a pointer to Flask, alerting it of the location of the App and causing it to launch. The "requirement.txt" file, which is the last, contains a list of every package used in the project.

The Run.py file

This file acts as a pointer to Flask, letting it know that our application exists and instructing it to execute the application. Although this appears complicated, it will soon make sense. We begin by importing the app module from the app folder we generated before.

Running the Flask server

Now, when you launch the flask server on your terminal with the command python server.py, your development server is hosted locally. Now, copy and paste the local URL you were given when the server first started.

4.5 Result on various stages

Table shows the result of the proposed approach to the Pima Indians Diabetes Dataset. The accuracy of 83.7% is recorded using theAdaBoost Technique, the accuracy of 76.62% is recorded using only the Decision Tree classification algorithm and 78.57% using Random Forest.

Model Name	**Accuracy in Percentage**
Random Forest Accuracy	0.7857142857142857
SVM Accuracy	0.8181818181818182
Decision Tree Accuracy	0.7662337662337663
AdaBoost Accuracy	0.8376623376623377

Deployment of the same is shown in the figure.

```
# Creating the AdaBoost classifier
abc = AdaBoostClassifier(base_estimator=rf, n_estimators=50, learning_rate=1, random_state=0)

# Training the AdaBoost model
abc.fit(X_train, y_train)

# Saving the model using pickle with .pkl extension
filename = 'adaboost_model.pkl'
pickle.dump(abc, open(filename, 'wb'))

# Making predictions on the testing set using the AdaBoost model
y_pred = abc.predict(X_test)

# Evaluating the AdaBoost model
accuracy = accuracy_score(y_test, y_pred)
print("AdaBoost Accuracy:", accuracy)
```

```
Random Forest Accuracy: 0.7857142857142857
SVM Accuracy: 0.8181818181818182
Decision Tree Accuracy: 0.7662337662337663
/usr/local/lib/python3.10/dist-packages/sklearn/ensemble/_base.py:166: FutureWarning: `base_es
  warnings.warn(
AdaBoost Accuracy: 0.8376623376623377
```

Fig 4.3 Machine learning model accuracy

Running the Flask server

Your development server is now hosted locally when you launch the flask server on your terminal using the command python3 app.py. Now, copy and paste the local URL you were given when the server first started.

```
(env) utkarsh@utkarsh-ubuntu:~/Downloads/WebDev/Flask/FlaskIntroduction$ python3 app.py
 * Serving Flask app 'app'
 * Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment. Use a pro
 * Running on all addresses (0.0.0.0)
 * Running on http://127.0.0.1:5000
 * Running on http://192.168.29.230:5000
Press CTRL+C to quit
127.0.0.1 - - [08/May/2023 14:39:23] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [08/May/2023 14:39:23] "GET /favicon.ico HTTP/1.1" 404 -
```

Fig 4.4 Deploying Flask Server

File Structure to setup Flask application

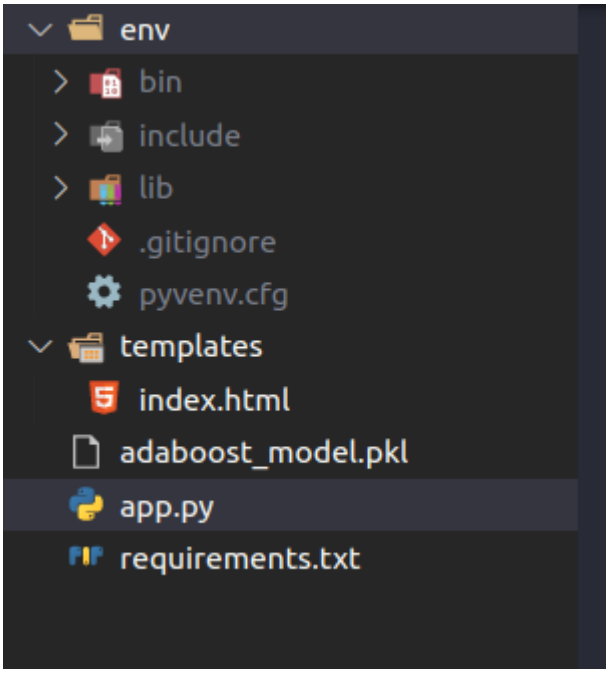


Fig 4.5 Flask Environment Setup

Website after deployment

Website which accepts information about as form and spend a put request to the server.

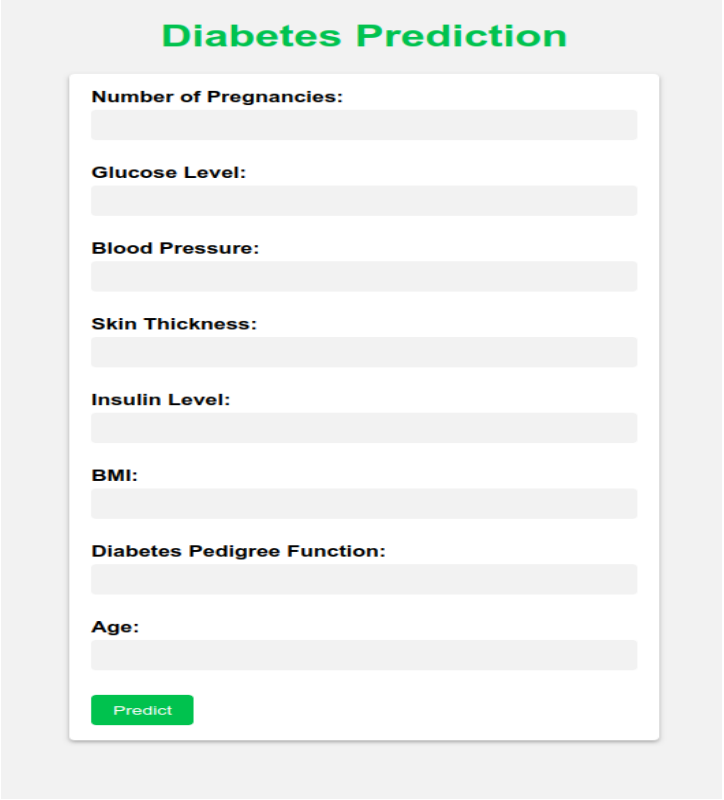


Fig 4.6 Deployed Website

Chatpter 5 - Performance Analysis

5.1 Conclusion

Life expectancy and quality of life may be impacted by diabetes. Long-term reductions in risk and complications from numerous diseases can be achieved by earlier detection of this chronic condition. A machine learning-based automatic diabetes prediction system has been put forth in this study. Both the open-source Pima Indian dataset and the female Bangladeshi patient private dataset were used in this investigation.

The problem of imbalanced class problems has been addressed by the use of SMOTE and ADASYN preprocessing techniques. For several machine learning and ensemble algorithms, this study presented performance metrics, including accuracy. With an accuracy rate of 81%, the AdaBoost classifier had the best performance when using the Random Forest as base classifier. The adaptability of the suggested prediction system has then been demonstrated by using the domain adaption technique.

The Flask framework has been implemented to build a website which uses the trained mode to forecast diabetes. Future directions for this research include, for example, including more private data from a personal of patients to improve outcomes. Combining fuzzy logic methods with machine learning models and using optimisation strategies is another development of this work.

Due to the application of cutting-edge computational techniques and the accessibility of a sizable number of epidemiological and genetic diabetes risk datasets, machine learning has the potential to fundamentally alter the capacity to forecast the risk of developing diabetes.

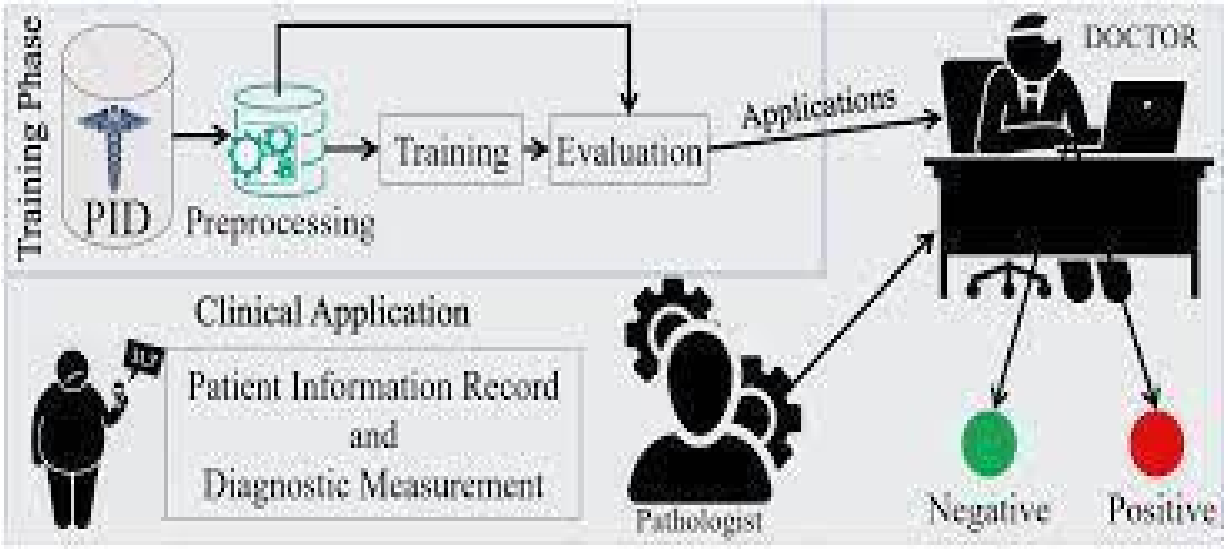


Fig 5.1 Deployment of full Application

5.2 Future Scope:

But by using Flask to create a user-friendly website so that anyone without any knowledge of the technology(python, machine learning) can easily use the website and get the output in a more admissible way so that he can easily understand it.

Objective of creating such an application is so that anybody with can check their diabetes and get to know wheater they are prone to get diabetes or not in future if they keep living in a particular lifestyle.
And Checking diabetes becomes as easy as shopping for items from an e-commerce website.

Future Plans The "KNN algorithm" is the classification method used by the suggested system to detect diabetes; other data science classification algorithms include Naive Bayes, SVM, Decision Tree, ID3, and others; we can add more algorithms in the future and compare them to find the most efficient algorithm.

A treatment module can be added where doctors can upload patient treatment information and patients can read it. Healthcare workers struggle to locate healthcare data and conduct studies on them due to a lack of tools and resources. However, we can work around this by using ML to analyse real-time data, which will enhance modelling predictions. Healthcare services have improved overall.

References

1. *Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importances for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.*
2. *K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.*
3. *Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Perfor- mance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 Feb- ruary, 2019.*
4. *Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineer- ing Research and Application, Vol. 8, Issue 1, (Part -II) Janu- ary 2018, pp.-09-13*
5. *Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.*
6. *Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabe- tes Disease Prediction Using Data Mining ".International Con- ference on Innovations in Information, Embedded and Com- munication Systems (ICIIECS), 2017.*
7. *Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.*
8. *A.K., Dewangan, and P., Agrawal, Classification of Diabetes Mellitus Using Machine Learning Techniques, International Journal of Engineering and Applied Sciences, vol. 2, 2015.*
9. *Priyanka Indoria, Yogesh Kumar Rathore. A survey: Detection and Prediction of diabetics using machine learning techniques. IJERT, 2018.*
10. *Khaleel, M.A., Pradhan, S.K., G.N Dash. A Survey of Data Mining Techniques on Medical Data for finding frequent diseases.IJARCSSE, 2013.*
11. *K. Vembandasamy, R. Sasipriya, E. Deepa. Heart Diseases Detection using Naïve Bayes Algorithm. IJISET, 2015*