**Role of smart health records in health care using machine learning:**

**A PROJECT REPORT**

***in partial fulfilment for the award of degree***

**BACHELOR OF TECHNOLOGY in**

**COMPUTER SCIENCE AND ENGINEERING**

***under the guidance of***

**Dr. Sabyasachi Pramanik**

***Submitted By***

**DEV PROTIM SIKDAR (10300118117)**

**ARUNIMA PRAMANIK (10300118122)**

**ARCHISMAN BANIK (10300118125)**

**APARAJITA JANA (10300118126)**



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**HALDIA INSTITUTE OF TECHNOLY**

**2022**

**TABLE OF CONTENT**

1. **Introduction…………………………………………………………6**
2. **Project Scope……………………………………………………….7**
3. **Requirement Specification…………………………………………8**
   * 1. **Functional requirements**
     2. **Hardware requirements**
     3. **Software requirements**
4. **Modules……………………………………………………………9**
5. **Technology Description………………………………………….11**
6. **Machine Learning Algorithms……………………………….…12**
   1. **Decision Tree algorithm:**
7. **Random Forest algorithm: …………………………………..13**
8. **Conclusion and Result…………………………………………….23**
9. **Advantage…………………………………………………………24**
10. **Future Plan………………………………………………………. 25**
11. **References………………………………………………………….26**

****

**CERTIFICATE OF APPROVAL**

This is to certify that the Final year project entitled " **ROLE OF SMART HEALTH RECORD IN HEALTH CARE USING MACHINE LEARNING**" is submitted by:-

**DEV PROTIM SIKDAR (10300118117)**

**ARUNIMA PRAMANIK (10300118122)**

**ARCHISMAN BANIK (10300118125)**

**APARAJITA JANA (10300118126)**

under the supervision and guidance of Assistant Prof. Mr. Sabyasachi Pramanik of Haldia Institute of Technology, Haldia. We are satisfied with his work, which is being presented for the partial fulfilment of the degree of Bachelor of Technology in Computer Science and Engineering, Haldia Institute of Technology Affiliated to Maulana Abul Kalam Azad University of Technology, Kolkata- 700064

**Research Supervisor**: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Mr. Sabyasachi Pramanik (Asst.Prof)**

**Programme Co-ordinator: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Mrs. Rajrupa Metia (Asst.Prof)**

**Head of the Department: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Prof (Dr.) Subhankar Joardar**

**DECLARATION**

* We hereby declare that the project entitled “Role of smart health records in health care using machine learning” is an authentic record of our work done under the guidance of Dr. Sabyasachi Pramanik at Haldia Institute of Technology affiliated to Maulana Abul Kalam Azad University of Technology, West Bengal, towards the partial requirement for the conferment of degree of Bachelor of Technology in computer science and Engineering batch of 2018-2022.
* We also state that, this submission is our own work and is not the work of any other person. It holds no materials which have been previously published or written by others.
* Any contribution made by others to the research or development of this thesis is explicitly acknowledged in the paper. Any of the above violation will be liable for punitive measures.

Signature of the students

Dev Protim Sikdar (10300118117)

Arunima Pramanik (10300118122)

Archisman Banik (10300118125)

Aparajita Jana (10300118126)

**ACKNOWLDGEMENT**

We are immensely thankful to our mentor and professors who gave us a chance to work on this project. However, it won’t be possible without the kind participation and support of a lot of individuals responsible to this project. We are also grateful to our college for transfer us all the required resources for the project. Wholly, we would like to appreciate everyone involved in this project and assisted us with their recommendation to make the project exceptional.

**ABSTRACT**

The project entitled ‘Machine learning implementation for smart health records’, is an application using which we can maintain the health records of the patients using a digital card which can be used only by the doctor, receptionist and the respected departments in the hospital. We will implement prediction method using the responsive web application and the statistical analysis with the advanced methodologies like machine learning and python to predict the patient’s health condition by analysing the earlier health record.

Machine learning is a high-level approach for any kinds of health care implementation in this real-world scenario. We are keeping a repository where digital data related to the patients and their treatment is stored. In this article, we impose some machine learning methodologies on our data and will find out the optimal solution for the process of treatment and the good maintenance of the digital records of the patients. The records we maintain will be in sequential order and can be modified later by the admin. All these kinds of information were stored in the form of application, and we explained the procedure of machine learning usage in this application and the prediction models design and implementation.

**INTRODUCTION**

Machine learning is widely used technology in the

field ofscience and technology and the rapid

growth of the requirements of the people in their

daily lifeimproves the utilization of advanced

technologies and they used to rule our daily life.

Perhaps, consider medical industry. We need to

identify the patient health information and in this

kind of situation, we need an application which

can monitor the patient’s health records from the

client side and to the doctor'sside. We need a

proper channel to design application and must

maintain amodel which can identify the

predictions and insight

Machine learning is widely used technology in the

field ofscience and technology and the rapid

growth of the requirements of the people in their

daily lifeimproves the utilization of advanced

technologies and they used to rule our daily life.

Perhaps, consider medical industry. We need to

identify the patient health information and in this

kind of situation, we need an application which

can monitor the patient’s health records from the

client side and to the doctor'sside. We need a

proper channel to design application and must

maintain amodel which can find the

predictions and insight

Machine learning is widely used technology in the field of science and computation. The rapid growth of the requirements of the people in their daily life improves the use of advanced technologies. In this project we consider medical industry. We need to find the patient health information and in this kind of situation, we need an application which can monitor the patient’s health records from the client side and to the medical staff side. We need to design application to keep a model which can find the predictions and insights of the data.

We need to look more into the insights of the data and the forms of data which we are gathering and the importance of keeping every feature in our application. In this project we will discuss about data collection and find the process of setting up a counter for data to get predicted. In this article, we are dealing with machine learning models like decision trees and random forest algorithms which are more likely to be used for different prediction models design and implementations. First, we need to find them using some of the machine learning algorithms and keep the accuracy rate using confusion matrix.

**PROJECT SCOPE**

In this project the following things to be noted down for better research of this concept further. They are as follows:

* **Data Collection**: We need to collect the data from some sources. We need to create a repository which will be useful for keeping the stability in gathering the information from other sources. In collection of information, we may get noisy data and we need to remove them using pre-processing method. This process will maintain the stability and ambiguous information in the repository.
* **Backward Elimination:** It is a nested operator with which we need to conduct the process of eliminating the unwanted data in the dataset and the information related to P and SL values will be generated based on our consideration. Based on that information, if the value of the P is greater than the SL value, then the information can be eliminated from the repository. Because they can give the lowest accuracy of the model which we are using.
* **Machine Learning:** Machine learning is the main background of this prediction process and the data we got from the medical application. This application details can be informed in the later sections and the information we gathered can be used for the machine learning models for better prediction of what is going to happen for the patient in future and what are the main constrains the patients must follow if there are any problems with their health condition.
* **Health Record:** Prime Indian Diabetes dataset and heart disease dataset health record is a digital version of patient’s paper chart. These are real time patient records that are made available instantly and securely to the authorised users. It has patients’ medical history, medication, allergies, reports etc. It provides automate and streamline workflow.

**EXISTING SYSTEM**

There are different systems which are designed in the domain of health care. But the most effective thing in this project is that the large amount of data manipulated, and the performance of machine learning will be high.

The existing system does not have the QR code and data security implementation.

Dataset used in the project is Prime Indian Diabetes dataset and heart disease dataset.

**REQUIREMENT SPECIFICATION**

**Functional Requirement:**

* **Validation of administration**: The person who is accessing the application is validated using user id and password to maintain the security of data access.
* **Registration of new patient**: Each new patient detail is added using this function. Personal details and the other necessary medical information are taken as input from the accessing authority.
* **Fetch existing patient details using QR code**: The existing patient history is being fetched by this function by scanning the QR code that is given to patient.
* **Predict the outcome**: This function runs the algorithm and analyses the present dataset to predict possibility.

**Hardware Requirement:**

* **Processor:** Intel Core i5 2.5GHz or more
* **RAM:** 4GB DDR4 or more
* **HDD:** 512GB or more

**Software Requirement:**

* **OS:** Windows /Linux
* **Python:** Python 3.7
* **Jupyter Lab**
* **Pandas**
* **Google Collaboratory**
* **Joblib**
* **Sci-kit learn**
* **HTML & CSS**
* **Bootstrap**

**MODULES**

This system is currently developed by us. The following are the modules that are developed till now. Other modules are left as the future plan for the next semester.

The working of the modules is as follow:

* **Registration**: Any user that uses the system for first time needs to register to obtain username and password. The user needs to input necessary information. User needs to confirm the details input by them. In case any wrong input is given user is asked to re-enter the information.
* **Login**: Once the user has already register, they can login to the system using this module. The user is asked to input the username and password. Once the user has input the right username and password, they are given the access to the homepage.
* **Forgot password**: In case the authorised user has forgot password, the user is asked to enter the username and according to the user’sname the security question is asked that is stored in data base during the registration. Once the security question is matched, the user is allowed to reset his password.
* **Application Home Page**: Once the user successfully logged in, they are expected to input various personal and medical information. The inputs are taken in the format of web form. The information that anuser is expected to share with the application are:

1. NAME
2. Gender
3. Age
4. Height
5. Weight
6. Contact info
7. Habits
8. Food info
9. Earlier disease information
10. Duration of the suffering
11. Symptoms
12. Lab reports
13. Scanning History
14. Medication History
15. Current Review updates
16. Current scan reports
17. Current lab reports
18. Next update

**Backend prediction module**: In this module of the application, we implement various machine learning algorithm to predict the best fit result for the user’s current data with respect to the data we analysed from Prime Indian Diabetes dataset and heart disease dataset.

In this application currently, we will be focussing on two algorithm methodologies namely: a) Decision tree algorithm and b) Random Forest algorithm. Rest few algorithms with their comparative conclusions will be our scope of exploration in the successive studies.

**TECHNOLOGY DESCRIPTION**

* **Python 3.7:** Python offers concise and readable code. While complex algorithms and versatile workflows stand behind machine learning and AI, Python’s simplicity allows developers to write reliable systems. Developers get to put all their effort into solving an ML problem instead of focusing on the technical nuances of the language.

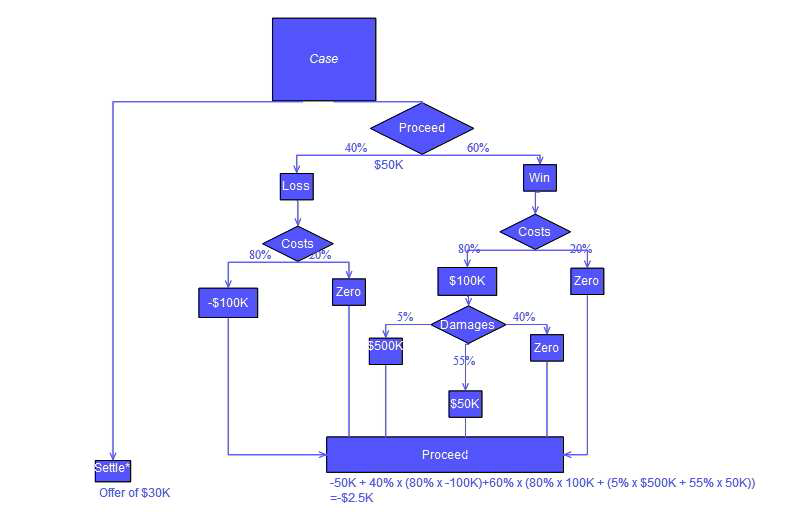
Programmers say that Python is more intuitive than other programming languages. Others point out the many frameworks, libraries, and extensions that simplify the implementation of different functionalities.

* **Jupyter Lab**:Jupyter Lab is a web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design allows for extensions that expand and enrich functionality.
* **Google Collaboratory:** Collab is a Python development environment that runs in the browser using Google Cloud.
* **Pandas:** Pandas is a software library written for the Python programming language for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series

**MACHINE LEARNING ALGORITHM**

* **Decision Tree Algorithm**: A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one of the ways to display an algorithm that only contains conditional control system. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal but are also a popular tool in machine learning

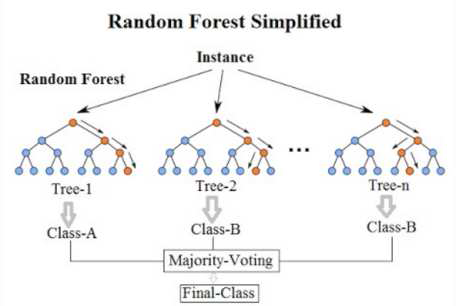
In this scenario we will identify all the possibilities for finding the basic model for the methods. The method consists of the best feature based on the priorities and the priorities will be found by the backward elimination process.



**Random Forest:** Random Forest or random decision forests are an ensemble learning method for classification, regression and other task that runs by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Random decision forest corrects for decision trees habit of fitting to their training set.





Random Forest implementation

**CODE IN USE**

**Dataset Used:**

Heart Disease Dataset

**Heart.py**

|  |
| --- |
|  |
|  | import numpy as np  import pandas as pd |
|  | from sklearn import ensemble |
|  | from sklearn.model\_selection import train\_test\_split |
|  | from sklearn.metrics import accuracy\_score, classification\_report, confusion\_matrix |
|  | import joblib |
|  |  |
|  | df = pd.read\_csv("../data/heart.csv") |
|  |  |
|  | categorical\_val = [] |
|  | continous\_val = [] |
|  | for column in df.columns: |
|  | if len(df[column].unique()) <= 10: |
|  | categorical\_val.append(column) |
|  | else: |
|  | continous\_val.append(column) |
|  |  |
|  | categorical\_val.remove('target') |
|  | dataset = pd.get\_dummies(df, columns = categorical\_val) |
|  |  |
|  | cols = ['cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'exang'] |
|  | X = df[cols] |
|  | y = dataset.target |
|  |  |
|  | X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42) |
|  |  |
|  | print('Shape training set: X:{}, y:{}'.format(X\_train.shape, y\_train.shape)) |
|  | print('Shape test set: X:{}, y:{}'.format(X\_test.shape, y\_test.shape)) |
|  |  |
|  | model = ensemble.RandomForestClassifier() |
|  | model.fit(X\_train, y\_train) |
|  | y\_pred = model.predict(X\_test) |
|  | print('Accuracy : {}'.format(accuracy\_score(y\_test, y\_pred))) |
|  |  |
|  | clf\_report = classification\_report(y\_test, y\_pred) |
|  | print('Classification report') |
|  | print("---------------------") |
|  | print(clf\_report) |
|  | print("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_") |
|  |  |
|  | joblib.dump(model,r"C:\Users\Archi\Desktop\HealthApp\Indivisual\_Deployment\Heart\_API\heart\_model.pkl")  **Open in colab** **Heart\_EDA.ipynb** import pandas as pd  import seaborn as sns  import numpy as np  import matplotlib.pyplot as plt  import warnings  warnings.simplefilter(action='ignore', category=FutureWarning)  df = pd.read\_csv("heart.csv")  df.describe().T  df.plot(kind='density', subplots=True, layout=(7,2), sharex=False , figsize =(18,18))  plt.show()      df.hist(figsize = (10,10), color = "#5F9EA0")  plt.show() |

correlation = df.corr()

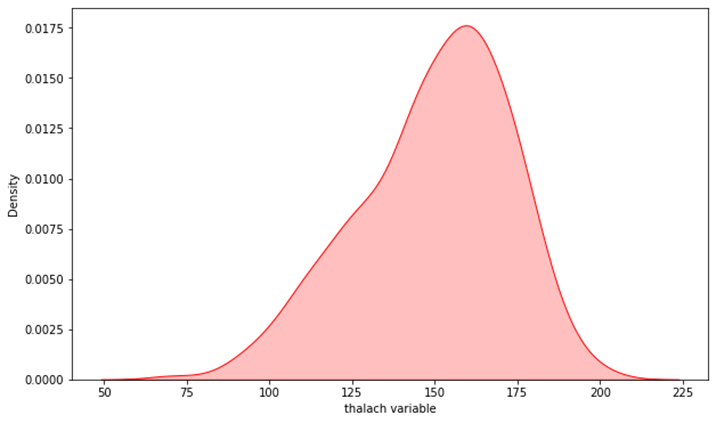
f, ax **=** plt**.**subplots(figsize**=**(10,6))

x **=** df['thalach']

x **=** pd**.**Series(x, name**=**"thalach variable")

ax **=** sns**.**kdeplot(x, shade**=True**, color**=**'r')

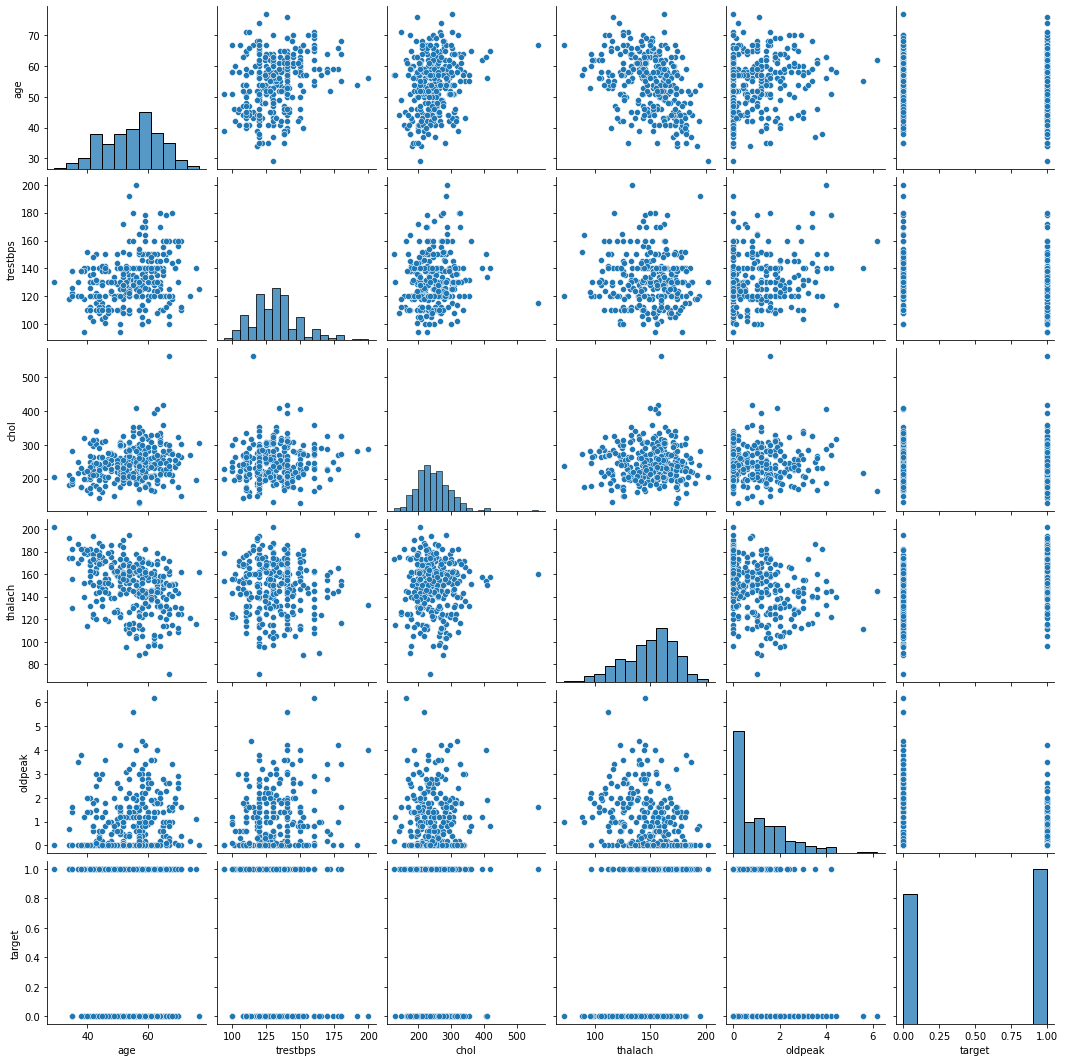
plt**.**show()



num\_var **=** ['age', 'trestbps', 'chol', 'thalach', 'oldpeak', 'target' ]

sns**.**pairplot(df[num\_var], kind**=**'scatter', diag\_kind**=**'hist')

plt**.**show()



|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

**Dataset Used:**

Pima Indian Diabetes Dataset

**Diabetes.py**

|  |
| --- |
|  |
|  | import numpy as np  import pandas as pd |
|  | from sklearn import ensemble |
|  | from sklearn.model\_selection import train\_test\_split |
|  | from sklearn.metrics import accuracy\_score, classification\_report, confusion\_matrix |
|  | import joblib |
|  |  |
|  | df = pd.read\_csv(r"C:\Users\Archi\Desktop\HealthApp\data\diabetes.csv") |
|  |  |
|  | df['Glucose'].replace(0, np.nan, inplace=True) |
|  | df['BloodPressure'].replace(0, np.nan, inplace=True) |
|  | df['SkinThickness'].replace(0, np.nan, inplace=True) |
|  | df['Insulin'].replace(0, np.nan, inplace=True) |
|  | df['BMI'].replace(0, np.nan, inplace=True) |
|  | df.drop(['SkinThickness' ,'Insulin'], axis=1, inplace=True) |
|  | df = df.dropna(axis=0) |
|  |  |
|  | cols = ['Pregnancies', 'Glucose', 'BloodPressure', 'BMI', |
|  | 'DiabetesPedigreeFunction', 'Age', 'Outcome'] |
|  |  |
|  | df\_outcome\_1 = df[df['Outcome'] == 1].copy() |
|  | i = len(df\_outcome\_1) |
|  | df\_outcome\_0 = df[df['Outcome'] == 0].sample(i, random\_state=1) |
|  | df\_balanced = df\_outcome\_0.append(df\_outcome\_1) |
|  | df\_balanced = df[cols] |
|  |  |
|  |  |
|  | y = df\_balanced['Outcome'] |
|  | X = df\_balanced.drop('Outcome', axis=1) |
|  | X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2 |
|  | , stratify=y, random\_state=1) |
|  | print('Shape training set: X:{}, y:{}'.format(X\_train.shape, y\_train.shape)) |
|  | print('Shape test set: X:{}, y:{}'.format(X\_test.shape, y\_test.shape)) |
|  |  |
|  | model = ensemble.RandomForestClassifier() |
|  | model.fit(X\_train, y\_train) |
|  | y\_pred = model.predict(X\_test) |
|  | print('Accuracy : {}'.format(accuracy\_score(y\_test, y\_pred))) |
|  |  |
|  | clf\_report = classification\_report(y\_test, y\_pred) |
|  | print('Classification report') |
|  | print("---------------------") |
|  | print(clf\_report) |
|  | print("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_") |
|  |  |
|  | joblib.dump(model,r"C:\Users\Archi\Desktop\HealthApp\Indivisual\_Deployment\Diabetes\_API\diabetes\_model.pkl")  **Open in colab**  **Diabetes\_EDA.ipynb**  import pandas as pd  import numpy as np  import matplotlib.pyplot as plt  import seaborn as sns  df = pd.read\_csv("diabetes.csv")  df.describe().T  df.hist(figsize = (10,10), color = "#5F9EA0")  plt.show() |

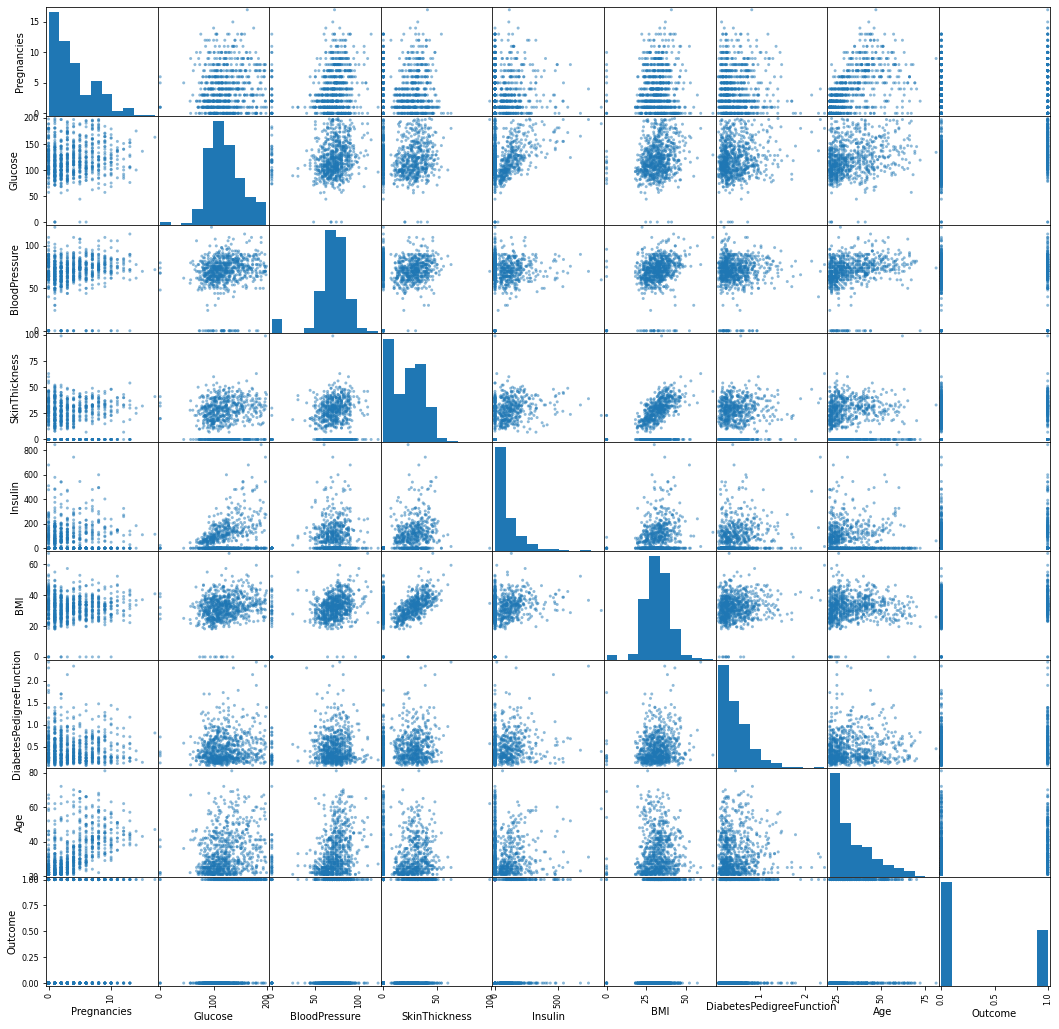
**from** matplotlib **import** pyplot

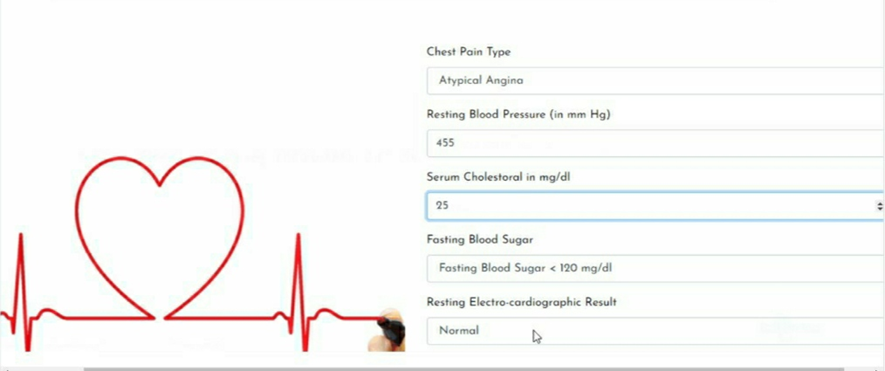
**from** pandas **import** read\_csv

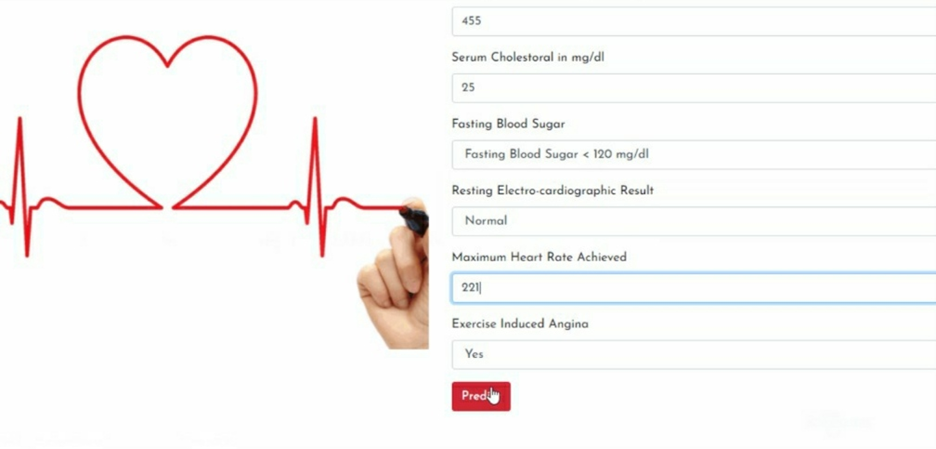
**from** pandas.plotting **import** scatter\_matrix

scatter\_matrix(df ,figsize **=** (18,18))

pyplot**.**show()





 **CONCLUSION AND RESULT**

The random forest methodology will give the highest accuracy with the decision boundary design and implementation using the multiple decision tree. The decision tree is used to identify the feature and predict a specific value and the random forest is the combination of more decision tree and we may get a variable decision boundary in the random forest. So based on requirement we identified that pure random forest could get the accurate result of identifying the patient health information and decision trees will stand next to RF in the accuracy. Machine learning enabled individuals in tackling challenges and shortcomings that else would have been burdensome in a prudent manner. Machine learning tools aid to bring out insights on data to analyse patterns and to build models to make predictions. Having Machine learning methodologies being enforced in the health domain benefits for processing immense amounts of data beyond the scope of human ability, vivid predictions to be formulated with machine learning models and aid for physicians for diagnosis in an efficacious manner. All those tedious and time-consuming processes can be hastened to save both time and labour. Our project entitled in the name ‘The Health Prediction system’ aids decide the possible disease in reference to symptoms. However, the challenges are still unsolved. Models are prone to overfitting that may end up in wrong predictions. Diagnosis cannot be done merely considering symptoms, there exist numerous factors concerned about the patient that can lead to diseases. They include lifestyle, gender, hereditary etc. Advancements must be brought in models to predict the disease based on factors other than symptom which aids doctors to rely on these models for efficient disease predictions

**ADVANTAGES**

* The application is secure to use since it allows only registered user to log in
* The application predicts the future health hazard of the user.
* The application might be used for predicting the estimated expense of a customer by a health insurance company.
* It uses QR code technology which allows contactless and fast communication between patient and admin

**FUTURE SCOPE**

The modules that are left untouched in this application yet will be covered in the next semester.

We will work on the modules of QR code generation and embedding the same.

We will evaluate the dataset with few other machine learning algorithms to analyse and compare the result for better and confident outcome.

We will work on the backend database handling code for seamless user experiences.

**REFERENCES**

1. K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna, ”Performance comparison of three data mining techniques for predicting kidney disease survivability”, International Journal of Advances in Engineering & Technology, Mar. 2014.

2. D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big data in healthcare: using analytics to identify and manage high-risk and high-cost patients,” Health Affairs, vol. 33, no. 7, pp. 1123–1131, 2014.

3. Disease prediction methodology: D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782

4. Boshra Brahmi, Mirsaeid Hosseini Shirvani, “Prediction and Diagnosis of Heart Disease by Data Mining Techniques”, Journals of Multidisciplinary Engineering Science and Technology, vol.2, 2 February 2015, pp.164- 168.

5. Streamlit official documentation: <https://docs.streamlit.io/en/stable/>

6. Decision tree and random forest exploration: https://www.khanacademy.org/computing/computer-science/informationtheory/info-t heory/pi/decision-tree-exploration