INTERNATIONAL TECHNOLOGICAL UNIVERSITY

Spring 2021 - 690: CAPSTONE PROJECT

Web Application for Heart Disease Prediction System

Project Supervisor

Prof. Mamoun Samaha & Prof. George Tannous

Submitted By: Team-13

Rakhi Jha - (89347)
Soni Kumari - (94970)
Sunanda Sharma - (95732)

Table of Contents

1. Overview

In recent times, Heart Disease prediction is one of the most complicated tasks in medical field. In the modern era, approximately one person dies per minute due to heart disease. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. This project makes use of heart disease dataset available in UCI machine learning repository. The Heart Disease Prediction application is an end user support and online consultation project. Here, we propose a web application that allows users to get instant guidance on their heart disease through an intelligent system online. The application is fed with various details and the heart disease associated with those details. The application allows user to share their heart related issues. It then processes user specific details to check for various illness that could be associated with it. Here we use some intelligent data mining techniques to guess the most accurate illness that could be associated with patient's details. Based on result, they can contact doctor accordingly for further treatment. The system allows user to view doctor's details too. The system can be used for free heart disease consulting online.

1.1 Objectives

The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historic heart data set Heart disease prediction system aims to exploit Machine learning techniques on medical data set to assist in the prediction of the heart diseases.

- To evaluate machine learning algorithms and selecting one model based on best performance in predicting heart disease.
- To develop wireless health monitoring system.
- To create a website (UI) for visualize/displaying the prediction results.

1.2 Mission

The main goal of this project is to provide a web application for users to predict heart disease as early stage to avoid disastrous consequences and build heathy society**.** It will help people to know about their heart condition without much spending.

1.3 Management Team

Rakhi Jha (89347) - Master's in Software Engineering
Soni Kumari (94970) - Master's in Computer science
Sunanda Sharma (95732) - Master's in Computer science

1.4 Industry Needs

A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data. These systems are designed to support people to know about their heath condition free of cost and any time of emergency. Some hospitals use decision support systems, but they are largely limited. Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

1.5 Scope & Limitation

Scope

The scope of Machine Learning algorithms are increasing in predicting various diseases. The nature of machine learning algorithm to think like a human being is making this concept so important and versatile. The non-linear tendency of the Cleveland heart disease dataset was exploited for applying Random Forest to get an accuracy of 96.88%. The method of predicting heart diseases using Random Forest with well-set attributes fetches us more accuracy. By the proposed algorithm for heart disease prediction, many lives could be saved in the future and as its completely free of cost. So, this project will help those who cannot afford expensive medical bill to get to know their heart condition.

Limitations

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Machine Learning is autonomous but highly susceptible to errors. Suppose we train an algorithm with data sets small enough to not be inclusive. We end up with biased predictions coming from a biased training set. Massive amount of data for the most accurate prediction is required for this application to better train the ML model.

2. System Development Methodology

2.1 Implementation—Web Application:

Included here is the implementation of web application through feature driven development. Each activity of feature driven development is discussed with artifacts produced during that activity. The web application has been developed using Django framework of python.

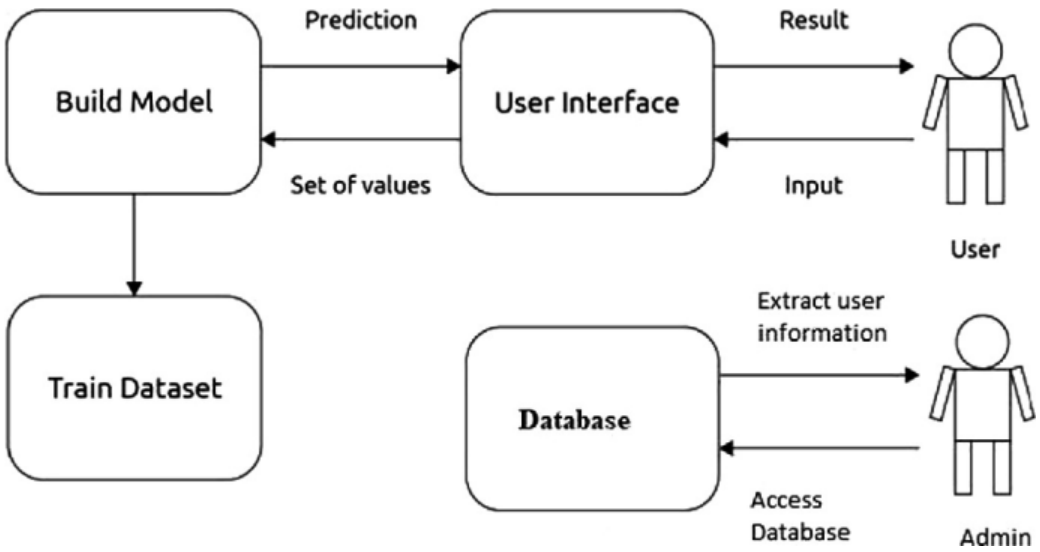
During this activity of feature driven development, software requirement specification document was prepared for capturing the requirements. ER Diagram and requirement specification document was designed. After that, for the completion of this activity, a domain object model was prepared along with the overall application architecture.

2.2 Functional Requirements

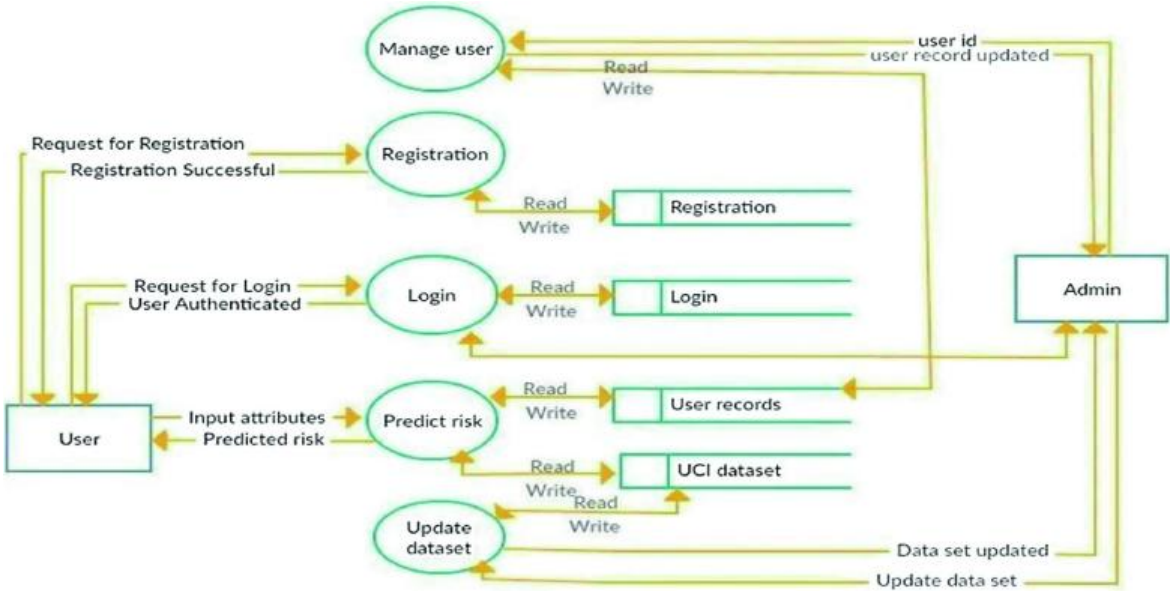
- The system allows users to create an account and login.
- It allows the users to update their profile, medical records and password.
- The system will allow user to fill their medical details into given form.
- The system performs heart disease prediction based on information's given by users.
- This site also contains admin site where it can monitor suspicious users and any potential threat to the system.
- It also stores all the user information in the database which user can modify later.

2.3 Use cases diagrams

Machine Learning System Diagram



2.4 System use-case diagram



2.5 Non-functional Requirements

Performance - The website is responsive and have consistent across different screen sizes and resolutions can also handle multiple user requests at the same time.

Scalability - This application can be easily scaled out or scaled in. Meaning, it can change according to the user's requirement. For example, we can integrate multiple machine learning models to get more specific heart disease problem in future or in addition to the existing application, we can also add IOT devices to take the actual data from the users instead of manually filling them.

Portability - We've used SQLite database for this project which pretty light and easily portable even easy for migrations.

Security - For one thing, Django hides the website's source code. The framework has protection against XSS and CSRF attacks, SQL injections, clickjacking, etc. Django notifies of a number of common security mistakes better than any other frameworks.

Reliability - This application is pretty reliable as we have used machine learning model which has good accuracy for the prediction.

Usability - very easy and user-friendly interface for the users to use.

3. Product Details

3.1 Project Description

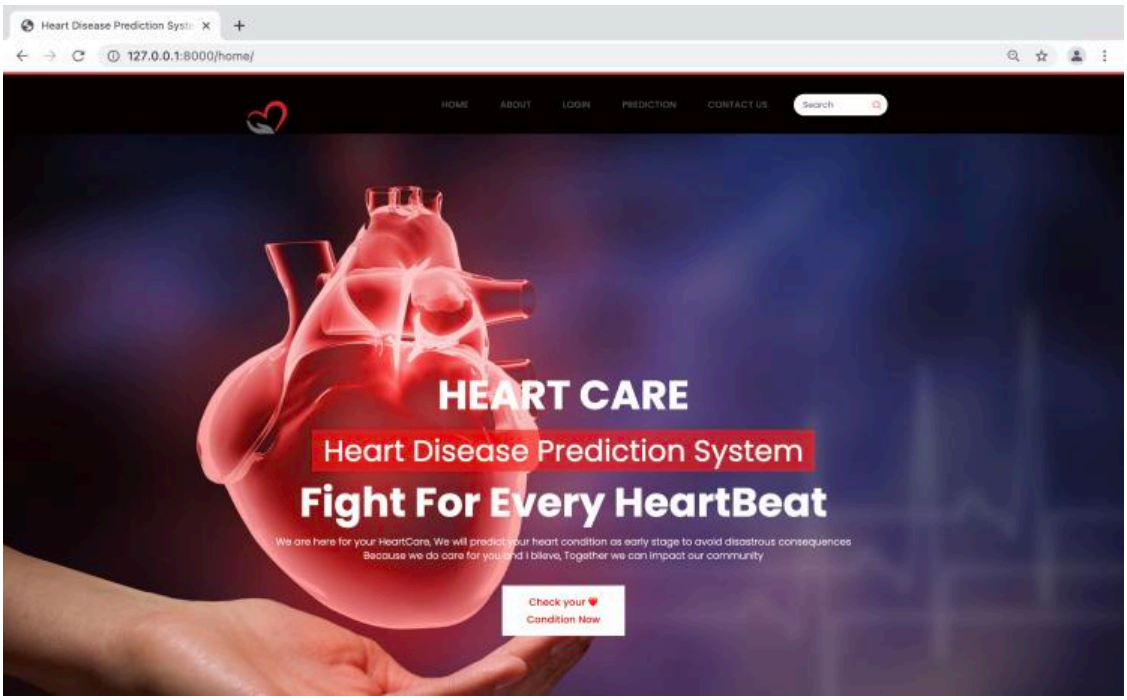
We've developed a web application that allows users to get instant information on their heart condition based on the health record details entered by the users. The application is fed with various details about the heart parameters to predict whether the user have any heart problem or not. Based on results, they may contact doctor accordingly for the further treatment. The system also allows users to register, however it's optional, if someone feels they don't want to share their details they just want to get the prediction then they can do that as well. It allows registered as well as guest to predict their heart condition. The application allows users to register, save and update their medical records for future use.

3.2 Technologies used in the project

- **HTML5** used for development of user interface and the design layout of the pages.
- **Cascading Style Sheet (CSS3)** is used to set the style in web pages that contain HTML elements.
- **JavaScript** is used for all the validation user input validations and to provide some animations in the web site pages.
- **Python3.9** is used all over in the project for backend part. It used to redirect the input from the user to the machine learning model as well as to interact with the database system.
- **SQLite3** database has been used as database management system for the project which will contain all the user details and permissions.
- **Django Framework 3.2** - This is the main component of the system which binds all the other components of system. This provides a safe and strong structure for the application.
- **Machine Learning** is used for data minding, data visualization and mainly to develop an algorithm which will run in the back end of the system, when user requests for the prediction.

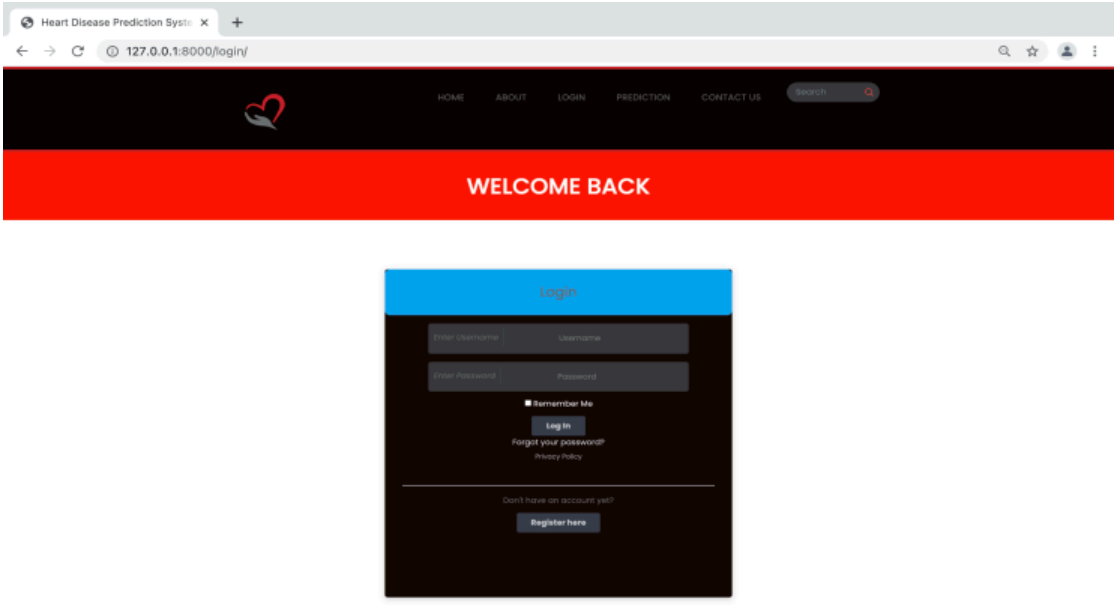
3.3 Important modules and their description

Home Page - Home page is the main page of the application from where all the pages are redirected. It has all the options to access different components of the project.

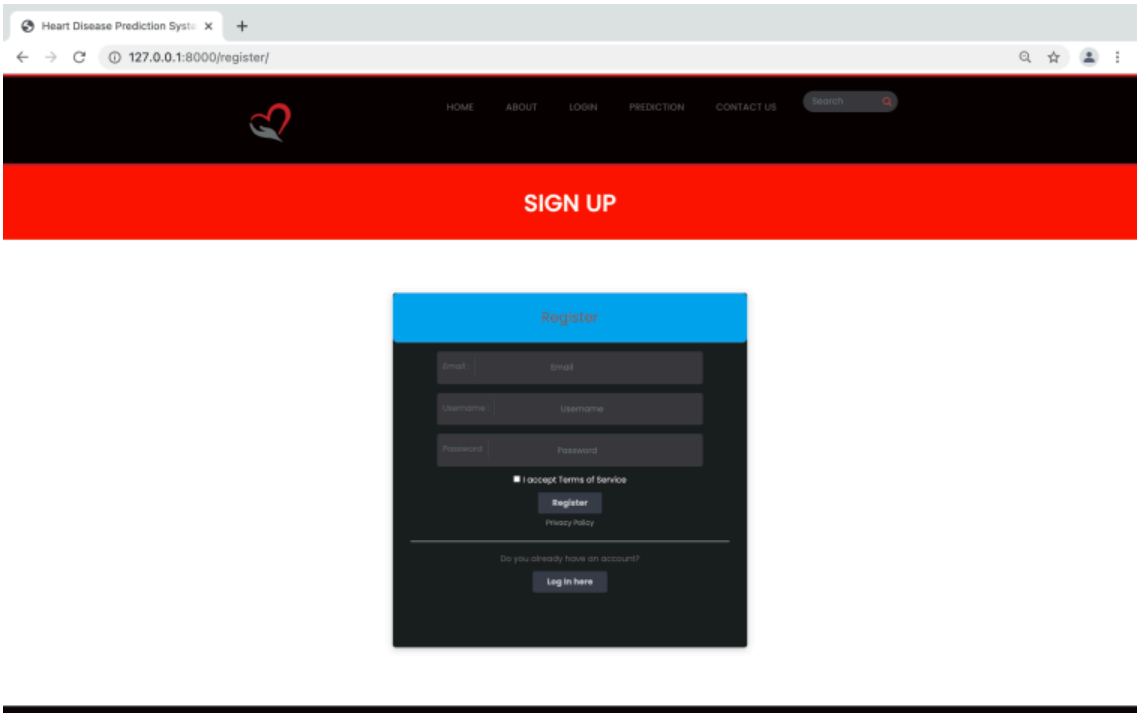


User login, Registration - In this module the application provides interface for registering the user and store the user credentials in the database. When user tries to login with the valid credentials, it'll render the user's dashboard otherwise it'll ask user to register if the authentication fails.

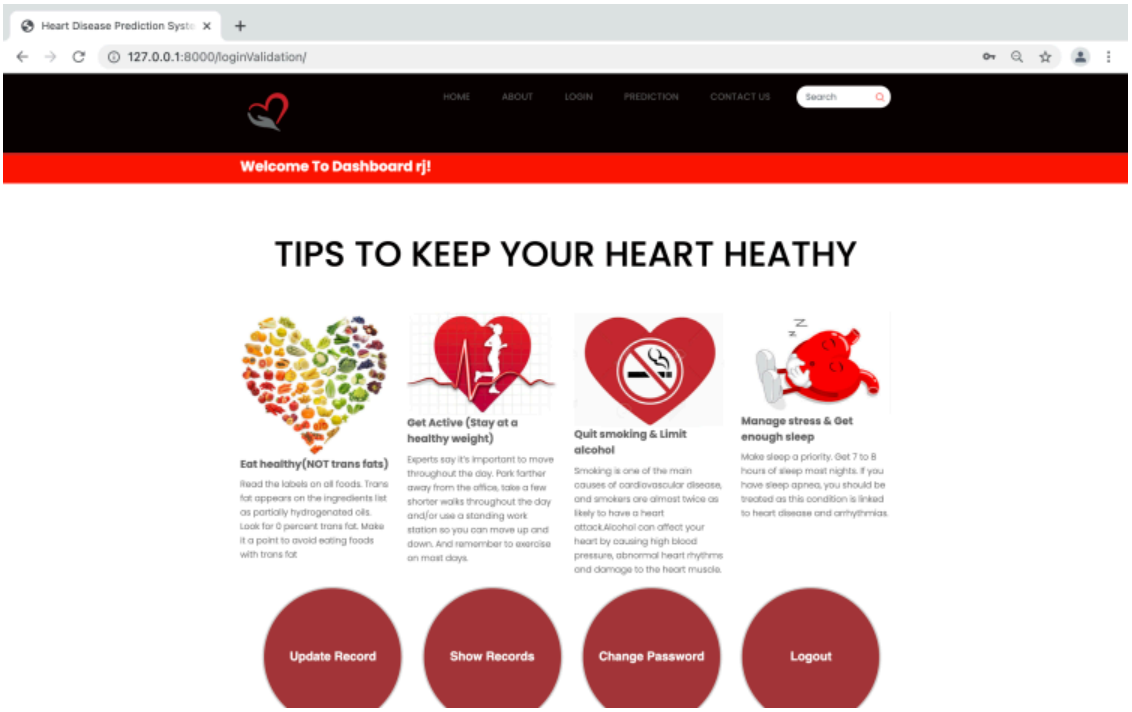
Login Page



Registration Page

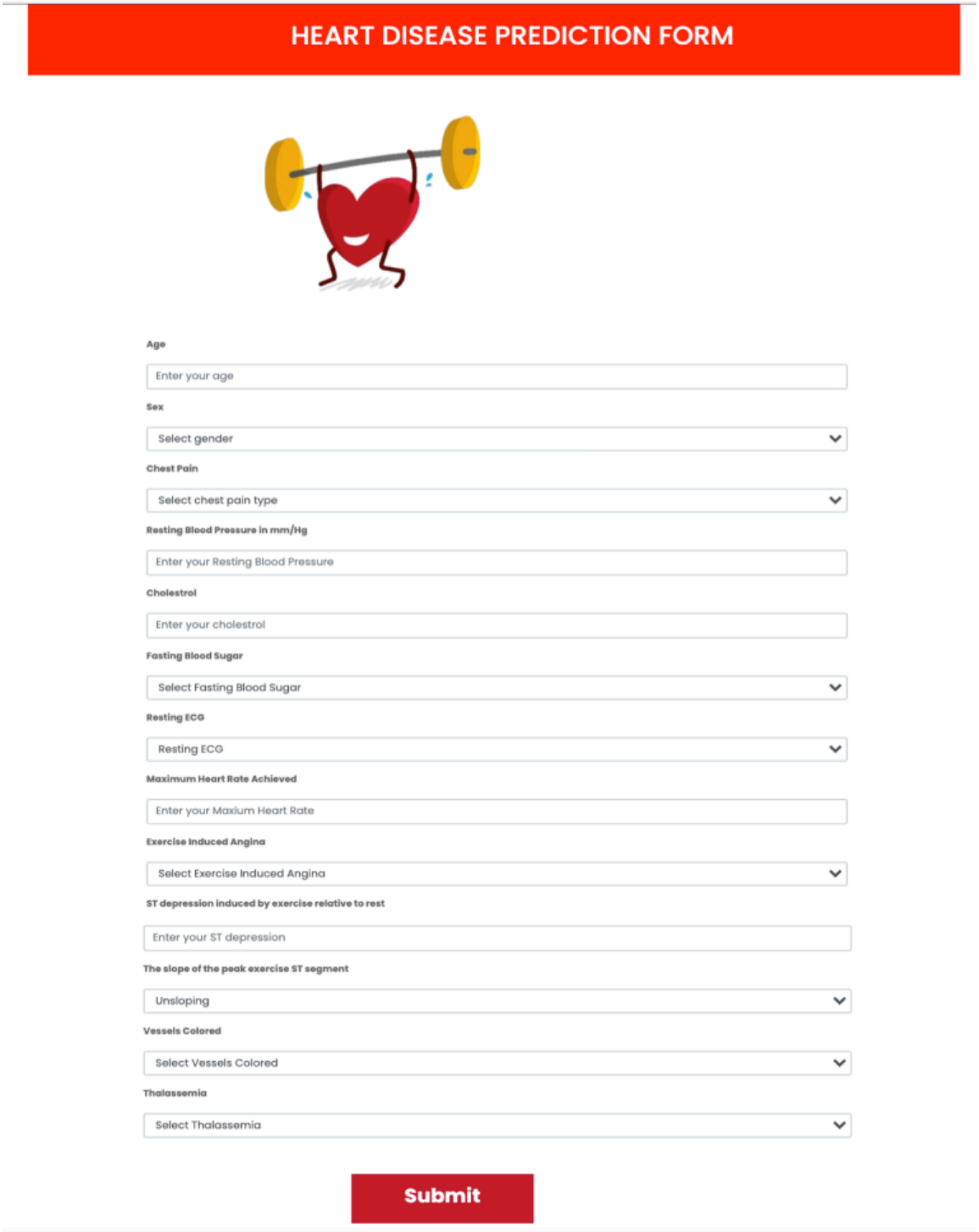


Dashboard Module - This module will have the options for the registered users to update their profile, update and view their previous medical records as well.

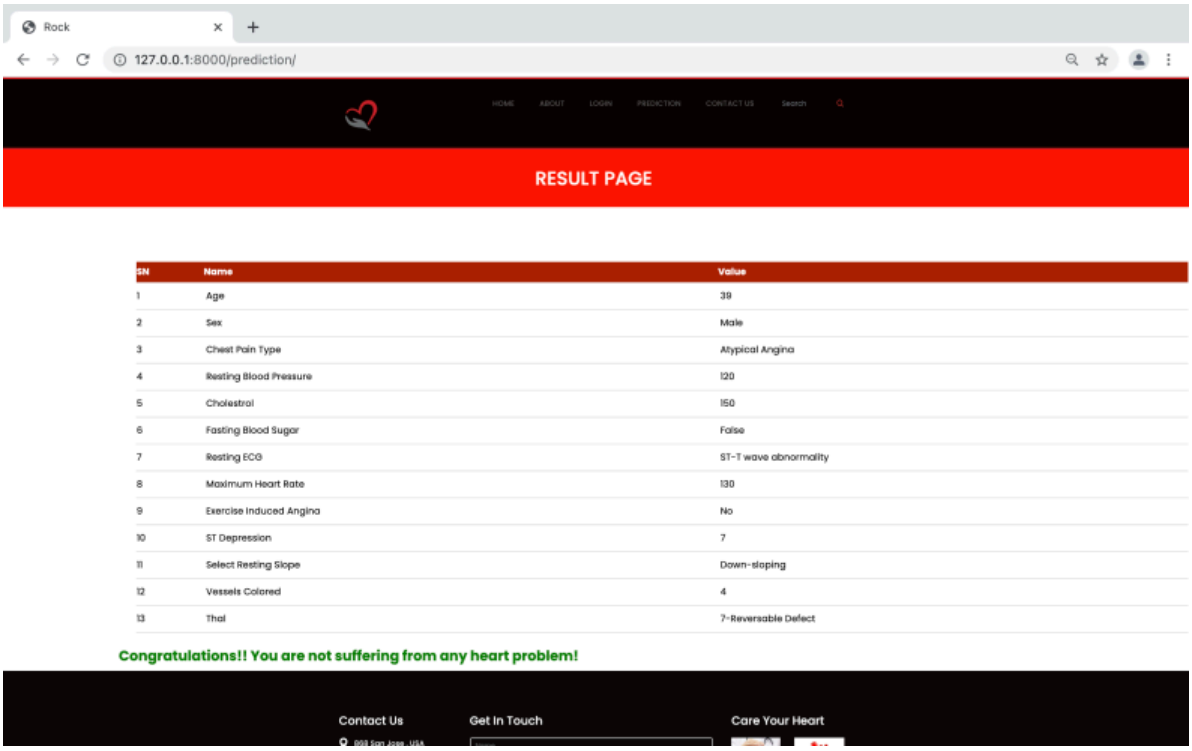


Prediction Module - This module will take all the heart related medical details from the user and save it to the database system. As well as it'll show the prediction to the next page. If the user is not registered, it'll not save the users information, it'll just show the prediction details in the next page.

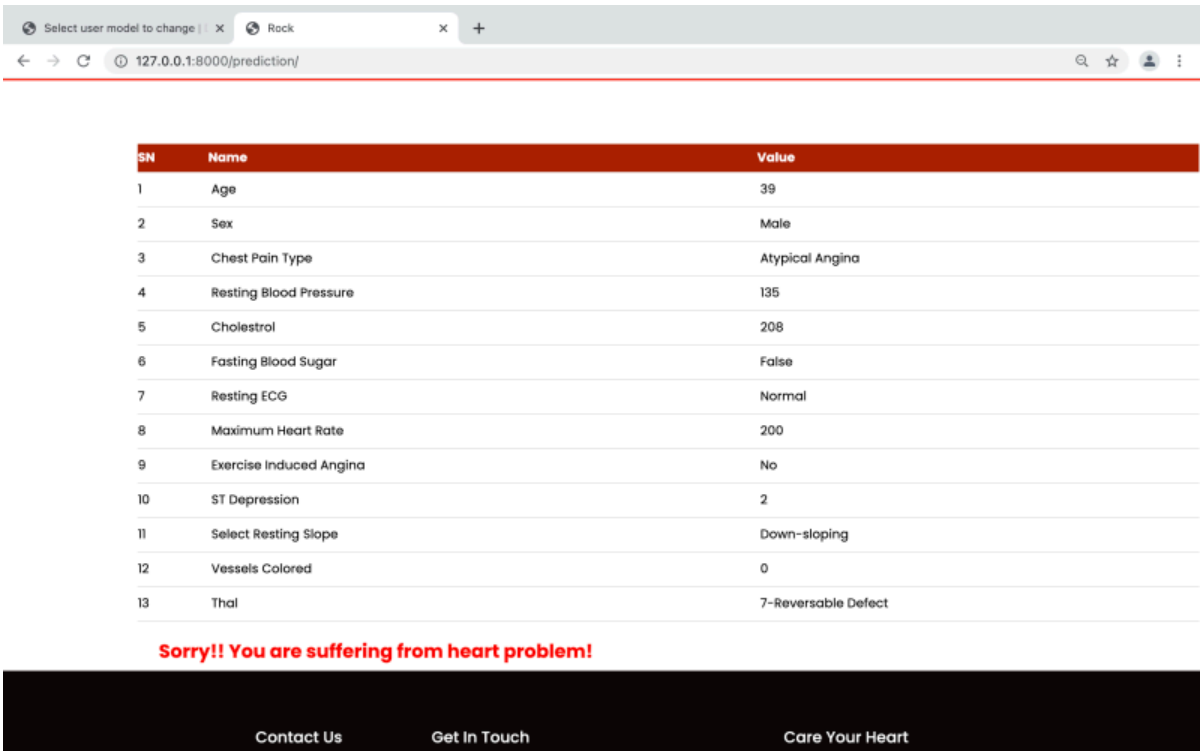
User Input Page



Prediction page who are not suffering from any heart problem



Prediction page for the users who might be suffering from the heart disease



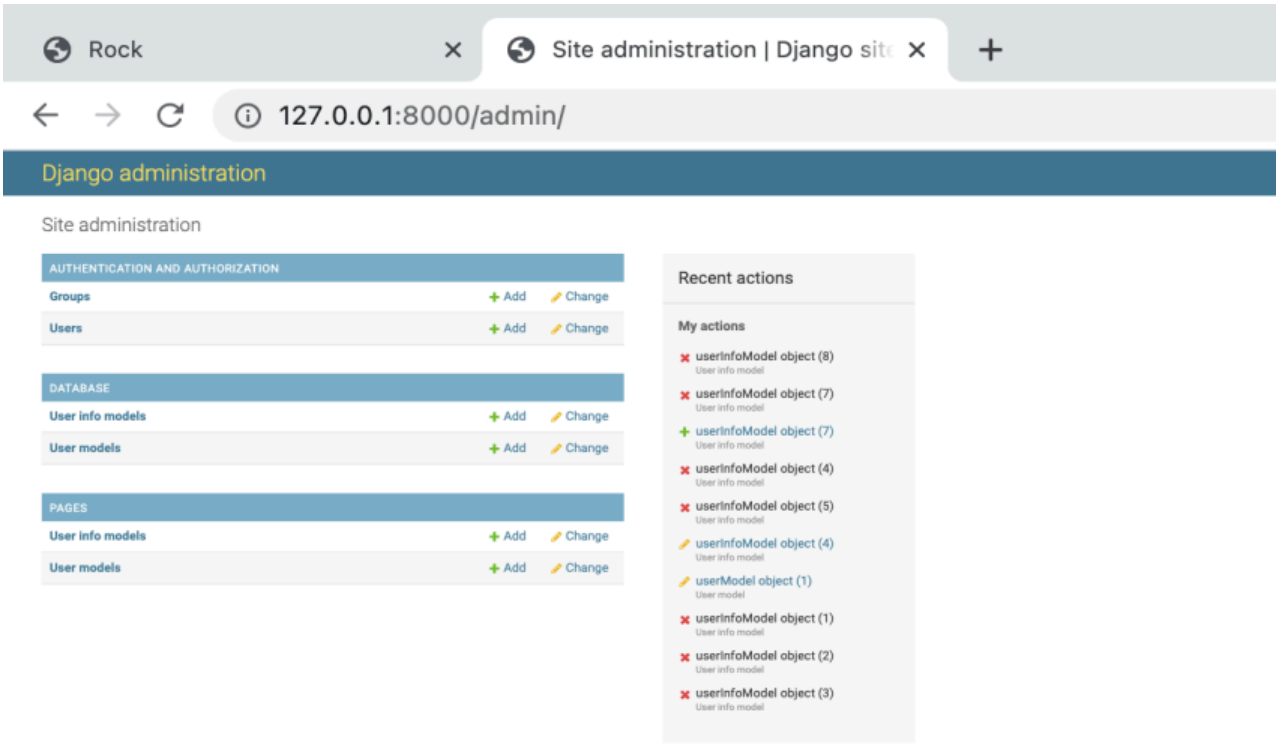
Machine Learning Module - The machine learning model will use the data provided from the user interface and will feed it to the ML model to predict the heart disease condition.

Machine Learning Interaction Module - We have integrated the model using python Django framework. Basically, we've used pickle library of python to create the pickle file of the model and used it in our application.

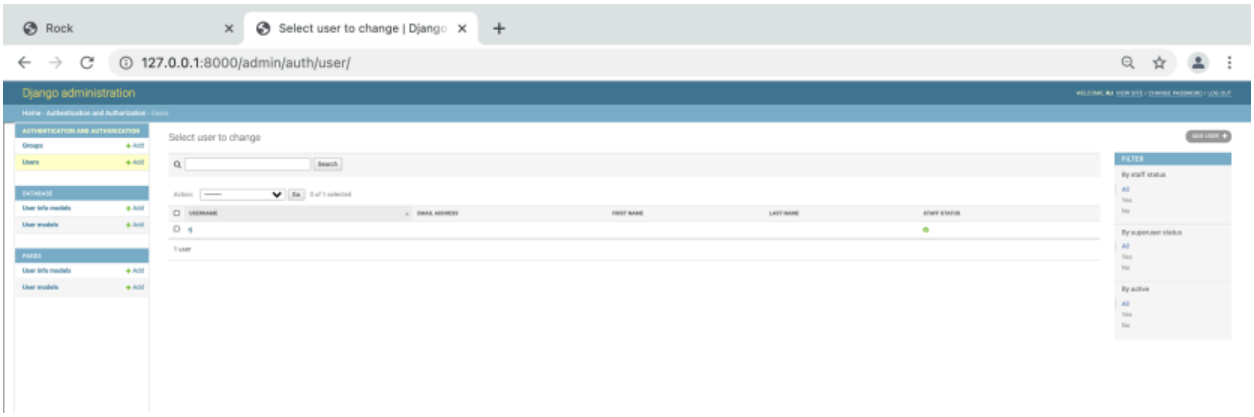
The most important advantage of the Django framework is that, we can build multiple machine learning modules and can integrate with this application in future.

Admin Module - This is also an important aspect of the project, where the admin will monitor the website and have the authorization to add or remove any suspicious users or tackle any potential threat to the system.

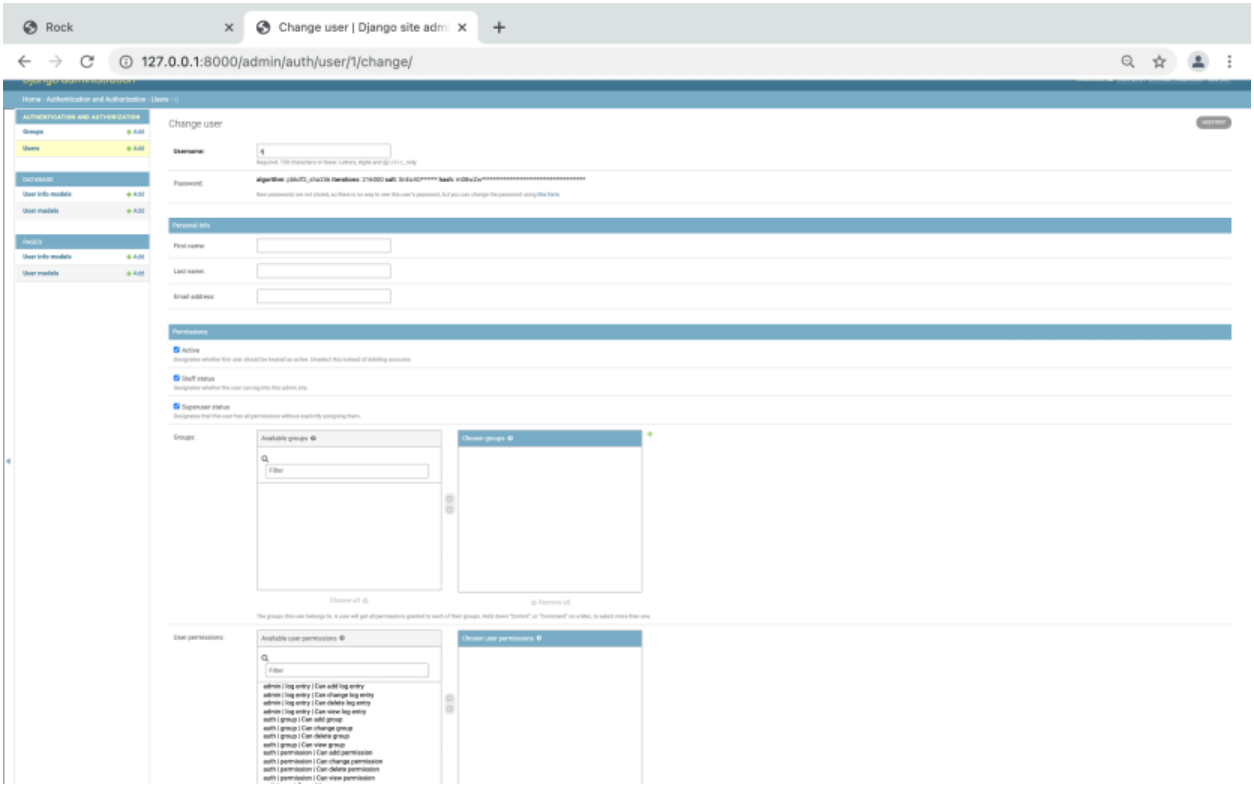
Admin page



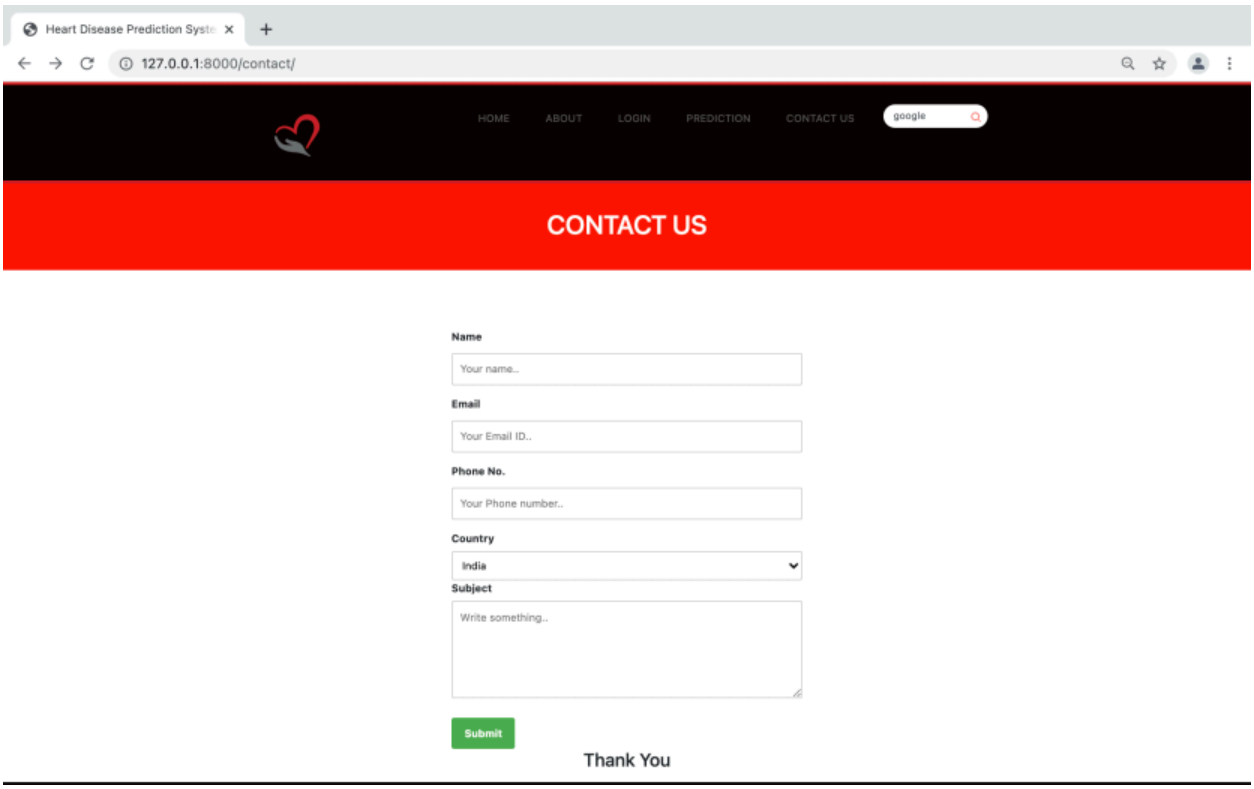
Admin user login page



Admin user detail page



Contact Module - In this module the user will be provided a form to fill for any of their queries which will be sent to the administrator for any difficulty in login access or any other enquiries.



4. **Machine Learning Mechanism:**

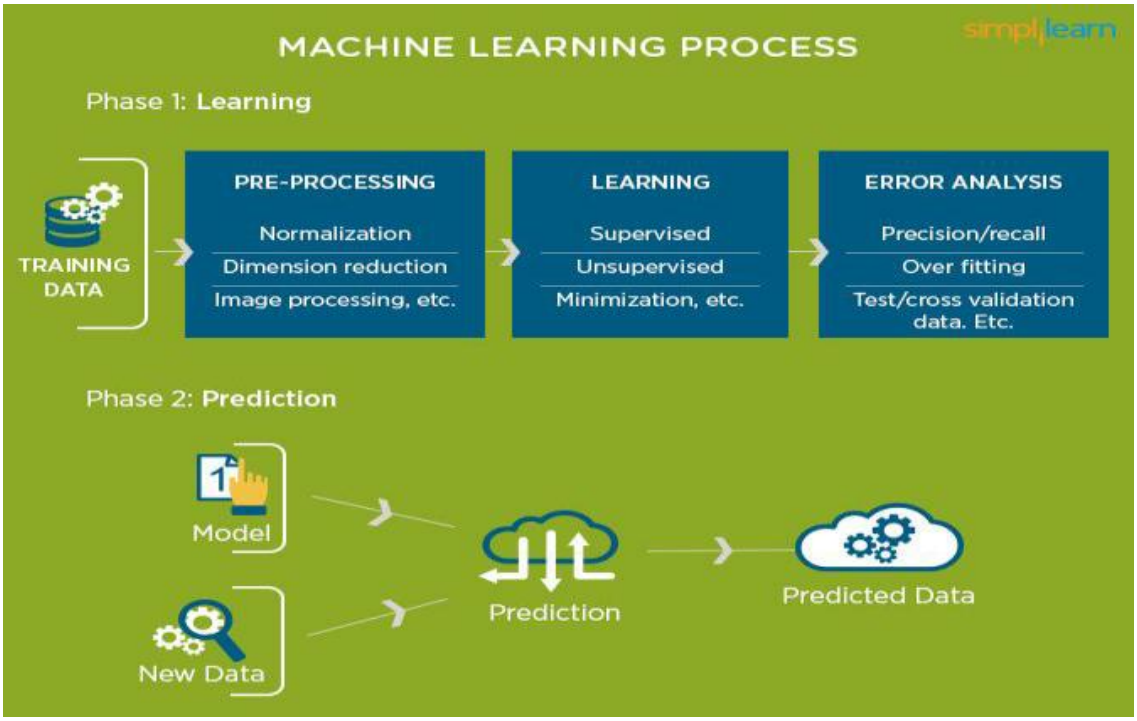
Machine Learning is a core sub-area of Artificial Intelligence (AI). ML applications learn from experience (well data) like humans without direct programming. When exposed to new data, these applications learn, grow, change, and develop by themselves. In other words, with Machine Learning, computers find insightful information without being told where to look. Instead, they do this by leveraging algorithms that learn from data in an iterative process.

At a high level, Machine Learning is the ability to adapt to new data independently and through iterations. Basically, applications learn from previous computations and transactions and use "pattern recognition" to produce reliable and informed results.

4.1 **How Machine learning Works?**

The Machine Learning process starts with inputting training data into the selected algorithm. Training data being known or unknown data to develop the final Machine Learning algorithm. The type of training data input does impact the algorithm, and that concept will be covered further momentarily.
To test whether this algorithm works correctly, new input data is fed into the Machine Learning algorithm. The prediction and results are then checked.

If the prediction is not as expected, the algorithm is re-trained multiple numbers of times until the desired output is found. This enables the Machine Learning algorithm to continually learn on its own and produce the most optimal answer that will gradually increase in accuracy over time.



4.2 Dataset Overview:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

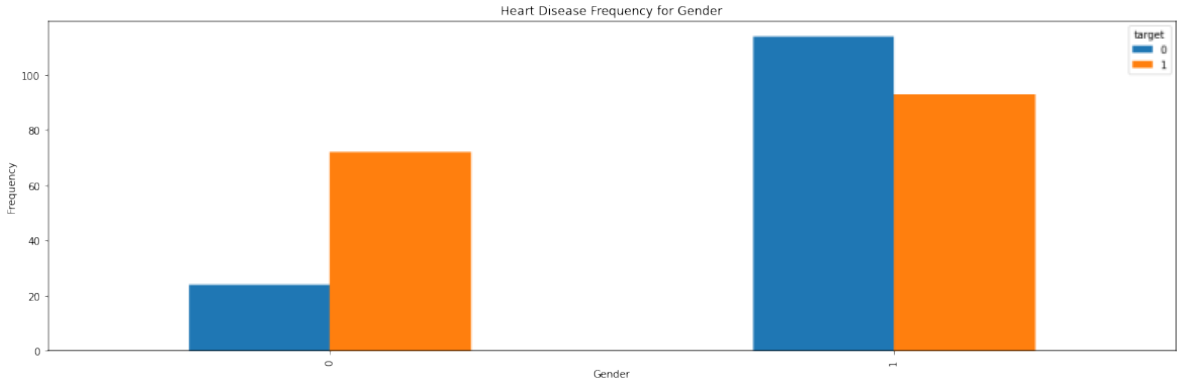
Dataset downloaded from UCI Machine Learning Repository containing information of various individual's medical records
https://archive.ics.uci.edu/ml/datasets/Heart+Disease

This dataset consists of 14 columns/features as described below:

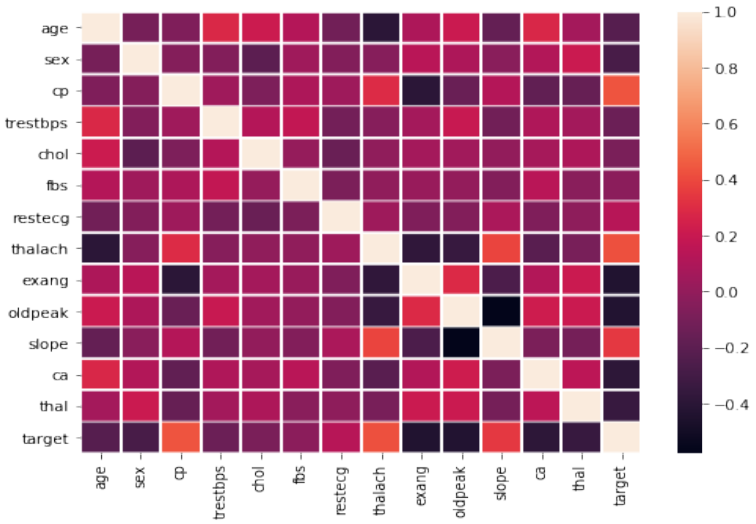
- **Age**: displays the age of the individual.
- **Sex:** displays the gender of the individual using the following format:
  1 = male , 0 = female.
- **cp:** Chest-pain type displays the type of chest-pain experienced by the individual using the following format:
  1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptotic
- **trestbps:** The person's resting blood pressure (mm Hg)
- **chol:** The person's cholesterol measurement in mg/dl
- **fbs:** (Fasting Blood Sugar), the person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
- **restecg:** resting electrocardiographic results, having values:
  0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy
- **thalach:** The person's maximum heart rate achieved
- **exang:** Exercise induced angina (1 = yes; 0 = no)
- **oldpeak:** It's the ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot.)
- **slope:** the slope of the peak exercise ST segment: 0 : down sloping, 1 : flat, 2 : upsloping
- **ca:** The number of major vessels (0–3) colored by fluoroscopy
- **thal:** A blood disorder called thalassemia Value 0: NULL (dropped from the dataset previously
  Value 1: fixed defect (no blood flow in some part of the heart)
  Value 2: normal blood flow
  Value 3: reversible defect (a blood flow is observed but it is not normal)
- **target:** Heart Disease (1 = yes, 0 = no)

4.3 Data Visualization:

Heart Disease Frequency with 'Sex'(Gender) and 'Target' variables:



Correlation Matrix with heatmap:

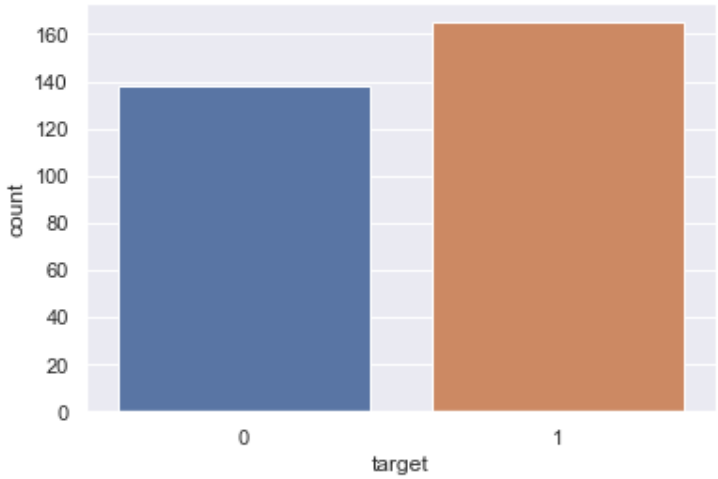


The maximum correlation is shown by the following features with the target variable :

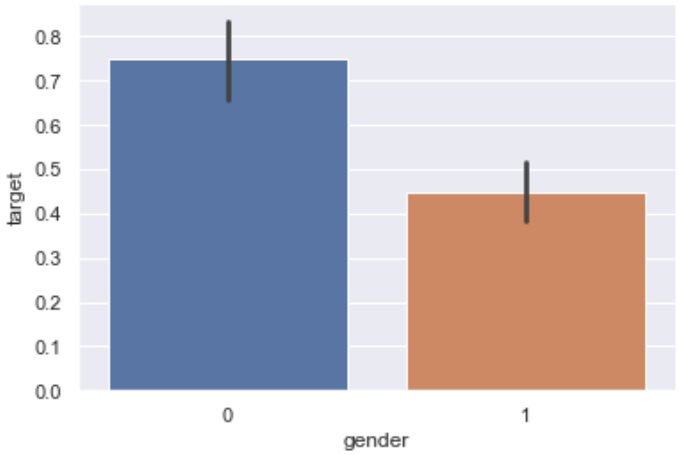
exang (Exercise induced angina)**:** 0.436757
cp (Chest-pain)**:** 0.433798
oldpeak (ST depression induced by exercise relative to rest): 0.430696
thalach (blood disorder called thalassemia): 0.421741

Correlation indicates how the features are related to each other or to the target variable. The correlation may be positive (increase in one value of the feature increases the value of the target variable) or negative (increase in one value of the feature decreases the value of the target variable)
Heatmap makes it easy to classify the features are most relevant to the target variable

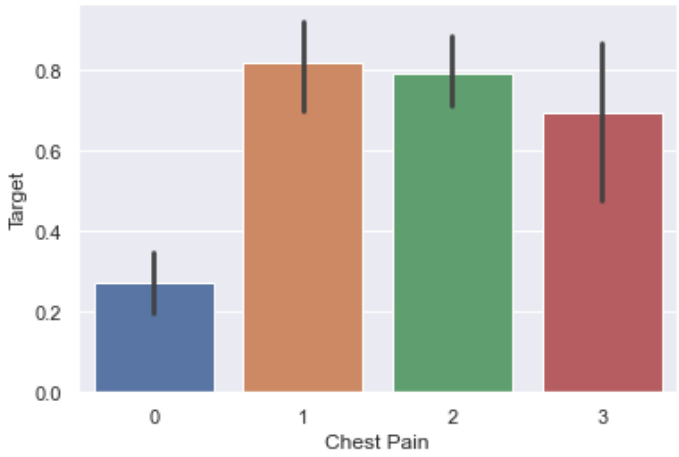
Count plot for 'target' variable:



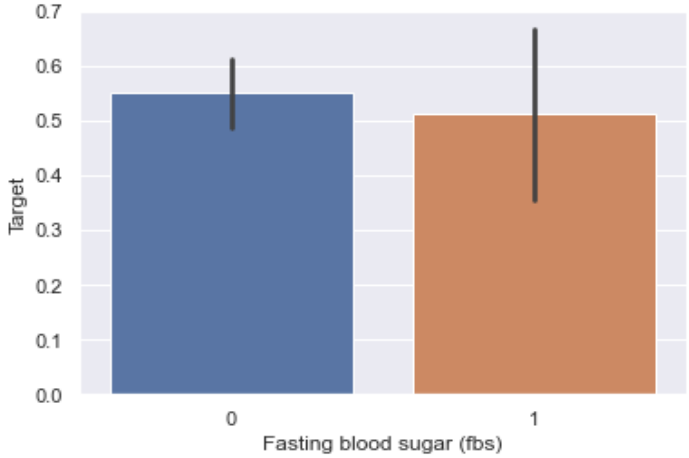
Bar Plot for 'Sex'(gender) and 'target' columns:



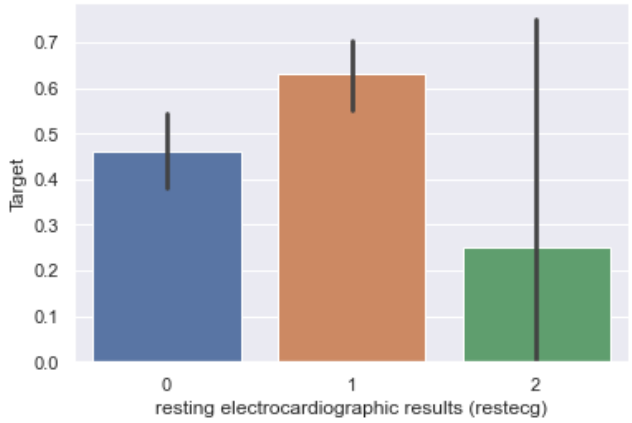
Bar Plot for 'cp'(chest pain) and 'target' columns: -



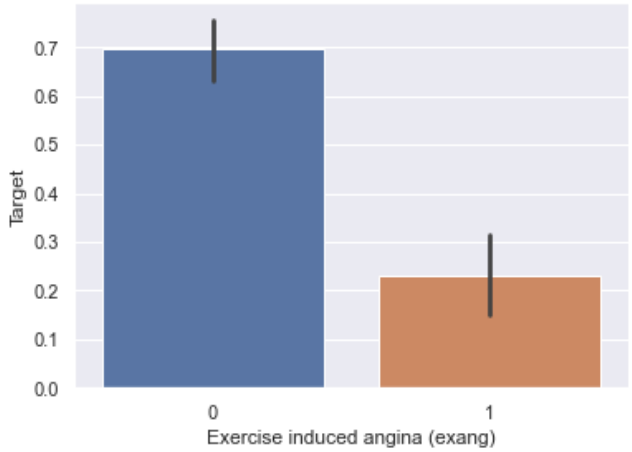
Bar Plot for 'fbs'(fasting blood sugar) and 'target' columns:



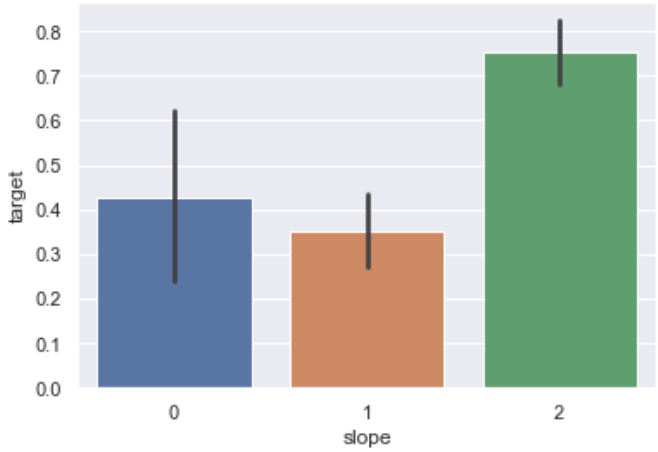
Bar Plot for 'restecg'(resting electrocardiographic results) and 'target' columns:



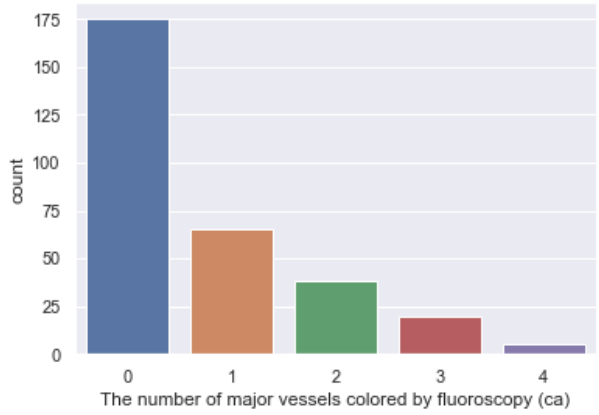
Bar Plot for 'exang'(Exercise induced angina) and 'target' columns:



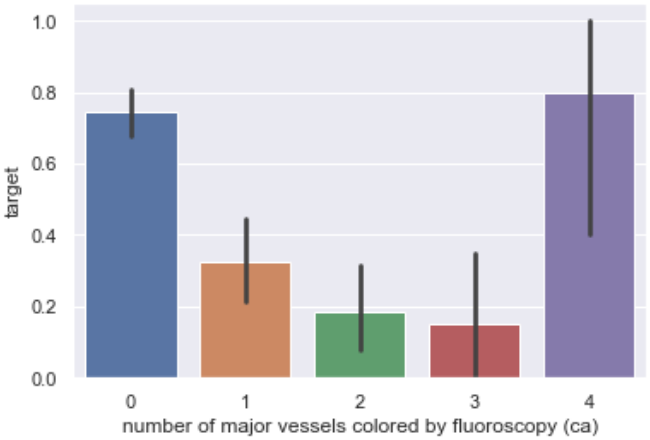
Bar Plot between 'slope' and 'target' columns:



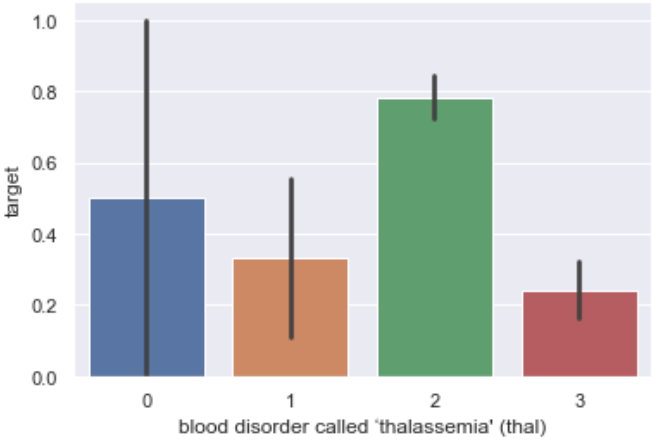
Count Plot for 'ca'(number of major vessels colored by fluoroscopy):



Bar Plot for 'ca'(number of major vessels colored by fluoroscopy) and 'target' columns:



Bar Plot of 'thal'(blood disorder called 'thalassemia') and 'target' columns:



4.4 **Random Forest Algorithm:**

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks).

This algorithm creates the forest with a number of trees. In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high the accuracy results.

"Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction."

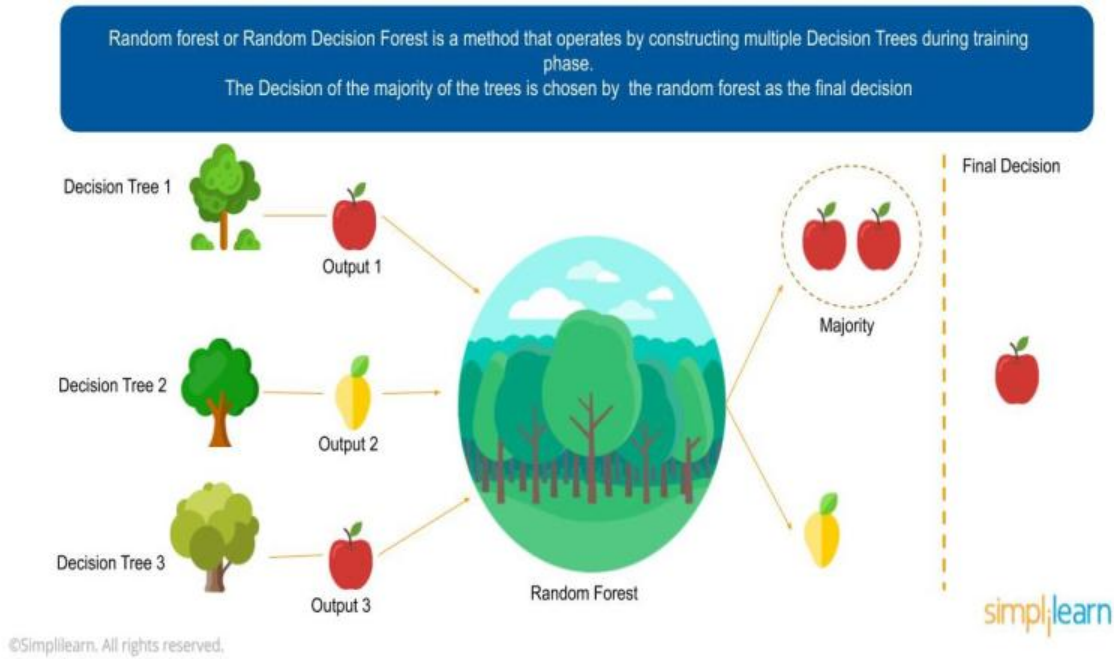
Random Forest is a learning method that operates by constructing multiple decision trees. The final decision is made based on the majority of the trees and is chosen by the random forest.

One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. With a few exceptions a random-forest classifier has all the hyperparameters of a decision-tree classifier and also all the hyperparameters of a bagging classifier, to control the ensemble itself.

The random-forest algorithm brings extra randomness into the model, when it is growing the trees. Instead of searching for the best feature while splitting a node, it searches for the best feature among a random subset of features. This process creates a wide diversity, which generally results in a better model. Therefore, when you are growing a tree in random forest, only a random subset of the features is considered for splitting a node. You can even make trees more random, by using random thresholds on top of it, for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

Why Random Forest algorithm?

1. The main reason for choosing random forest algorithm is that it can be used for both classification and regression tasks.
2. Overfitting is one critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier won't overfit the model.
3. Classifier of Random Forest can handle missing values
4. Random Forest classifier can be modeled for categorical values.



4.5 Random forest prediction pseudocode:

To perform prediction using the trained random forest algorithm uses the below pseudocode.

1. Takes the **test features** and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the **votes** for each predicted target.
3. Consider the **high voted** predicted target as the **final prediction** from the random forest algorithm.

```
score_random_forest = round(accuracy_score(y_pred_rand_classifier,y_test)*100,2)

print("The accuracy score achieved using Random Forest is: " + str(score_random_forest)+ " %")
```

```
The accuracy score achieved using Random Forest is: 96.77 %
```

New Patient Health Prediction with the Random Forest Classifier:

```
patient1 = rand_classifier.predict([[36, 0, 1, 135, 208, 0, 0, 171, 0, 1.5, 2, 0, 2]])
patient1
```

```
array([1])
```

```
patient2 = rand_classifier.predict([[22, 1, 0, 98, 130, 1, 1, 111, 1, 0.5, 1, 1, 1]])
patient2
```

```
array([0])
```

The following steps to perform suitable Machine learning technique:

- Defines a problem statement.
- Classifying the problem into ML problems.
- Selecting perfect ML algorithms based on their type of problems.
- Collecting and cleaning the data.
- Training a model from data.
- Test the Model from test data
- Evaluate a model from its accuracy

5. Technological advantages

The technologies that we've used in the project has some great advantages over the existing technologies in the current market.

5.1 Django Framework

It is written in Python and Python is amazing, clean, easy to learn, and one of the most taught programming languages. Django is written in Python.
The tech market is overflowed with frameworks, but Django is a good place to start as it has the nicest documentation and tutorials in software development.

- Django provides a wide range of features and functionalities including:
- Magical ORM(Object-Relational-Model)
- Multi-site and multi-language support
- MVC (Model/View/Control) layout
- AJAX support
- Free API
- URL routing
- Easy Database Migrations
- Session handling
- HTTP libraries and templating libraries

The administration interface provided by Django is one of the coolest things. It's truly simple to create and it's really one of the key advantages when using the framework.

Django is Immensely Scalable - One of the nicest advantages of Django is that it can handle traffic and mobile app API usage of more than 400 million+ users helping maximize scalability and minimize web hosting costs.

Best Security - For one thing, Django hides your website's source code. The framework has protection against XSS and CSRF attacks, SQL injections, clickjacking.

Future - it's growing every day and one can be guaranteed that the Django functionalities selected now will continue to operate well together in the future.

5.2 SQLite database

- SQLite is an open-source project available in the public domain.
- Django has in build functions which makes internal queries to interact with the database system using in build Django functions.
- SQLite is a server-less database and is self-contained.
- The SQLite directly stores info in a single file, which makes it easily portable, and speed is also faster than any other database system.

5.3 Python

- **Improved Productivity -** Python is a very productive language. Due to the simplicity of Python, developers can focus on solving the problem.

- **Free and Open-Source -** Python comes under the **OSI approved** open-source license. This makes it **free** to **use** and **distribute**.

- **Vast Libraries Support -** we don't have to depend on external libraries. Because Python has large set of internal libraries.

- **Portability -** In many languages like C/C++, you need to change your **code** to run the program on different platforms. That is not the same with Python. You only write once and run it anywhere.

5.4 Machine Learning Advantages

- **Easily identifies trends and patterns -** Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans.

- **No human intervention needed (automation) -** Giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus software; they learn to filter new threats as they are recognized.

- **Handling multi-dimensional and multi-variety data -** Machine Learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

6. Development Status

- We have done research about heart disease prediction system and Created Business Plan for the project.
- We use Trello App to keep trace on our project development time frame.
- Decided the best suited framework for the project which is Django
- Created database with Django which will store user information and medical details for user.
- Designed and create complete web Application where user can login and check their heart condition.
- Finalized the dataset and train the Machine Learning Algorithm which is random forest classifier, and we achieved the accuracy of 96.88%.
- Test and deployed machine learning model with the front end and deploy the project after complete testing.
- Created **Git Repository** for version control and to streamline our project development within team.

https://github.com/iamsonisingh/Heart-disease-prediction-Web-Application

7. Keys to Success

Heart disease has been identified as one of the largest causes of death even in developed countries. One of the reasons for fatality due to heart disease is due to the fact that the risks are either not identified, or they are identified only at a later stage. However, machine learning techniques can be useful for overcoming this problem and to predict risk at an early stage. We will test our dataset with few ML technique and apply the most suitable one for best accurate result.

- Getting the accurate data for the prediction system is a challenge. So, probably getting the accurate dataset can make this project better.
- Since this project is a non-profit application, it will encourage more people to use this application.

8. Competitive Advantages

In this application we are trying to use all the latest technologies like Machine Learning, Python and Python Framework Django. The advantage of these technologies is that they are open source and have vast community of support groups available online. Python is a very versatile programming language which is used efficiently for Machine learning project, IOT (Internet of Things) applications as well as general programming language. In future, if we want to scale out our project, then we can easily do that using Python.
One more important point about this technology is we are using Django framework, which goes well with Python and machine learning projects. This framework can have multiple apps inside one app, means we can install multiple Machine learning algorithms in one application. It treats every algorithm as different module of the project.

9. Business Strategy

Heart disease and stroke are leading causes of death and disability and the most expensive medical conditions for businesses. As of now we are creating heart disease perdition system website for free use. Our mission is to promote health awareness. Collaboration facilitates the sharing of knowledge, talents and resources for the betterment of all those involved. By cultivating a strong network of health advocates across the state, the Medical College of Wisconsin brings new discoveries that have the power to impact each and every one of us. Together, we will continue to develop collaborative research that advances the health and vitality of our own community.

Strength through community-based partnerships: A healthier community begins with strong networks of health advocates, community clinics and partnering healthcare and heath awareness. By developing this website, we will try to build a heathy community.

Together, we can impact our community.

10. Financial Goals

As of now, we are planning to make this project a non-profit application. This project built purely to help people & community to get better health care with minimal cost. This will encourage more people to use this application and get to know their heart condition and get aware of false alarm of heart attack. At later stages of the project, we can add more **advanced features like –**

- Instead of entering the medical info manually we can add sensors to take the input from wearable devices.
- We can provide list of cardiologists to get appointment if there is any case of emergency and even will link with google map so they can choose the nearest one.
- We can provide the hearth care tips and daily nutrition advise.
- We can contact best cardiologists and get advice on daily exercise and yoga tips which can be available on our website for register user.

After advancing the above features and accuracy to the most, we will plan to launch the mobile application with android and iOS and then we might think of getting profit out of it by asking user for monthly & yearly subscription.

11. Conclusion

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree and Logistic Regression, Random Forest for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 96.77% for prediction of heart-disease. In future the work can be enhanced by developing a web application based on the Random Forest algorithm as well as using a larger dataset as compared to the one used in this analysis which will help to provide better results and help health professionals in predicting the heart disease effectively and efficiently.

12. References:

- https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c
- https://www.djangoproject.com/start/
- https://towardsdatascience.com/understanding-random-forest-58381e0602d2
- https://www.guru99.com/execute-javascript-online.html
- https://machinelearningmastery.com/machine-learning-in-python-step-by-step/
- https://towardsdatascience.com/creating-a-machine-learning-based-web-application-using-django-5444e0053a09
- https://www.youtube.com/watch?v=F5mRW0jo-U4
- https://developer.mozilla.org/en-US/docs/Learn/Server-side/Django/Authentication
- https://towardsdatascience.com/machine-learning-books-you-should-read-in-2020-344b44d9a11e
- https://github.com/iamsonisingh/heart-disease-prediction
- https://www.cdc.gov/heartdisease/risk_factors.htm
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- https://www.amazon.com/Introduction-Machine-Learning-Python-Scientists/dp/1449369413
- https://towardsdatascience.com/creating-a-machine-learning-based-web-application-using-django-5444e0053a09