“Effective Heart Disease Prediction Using IBM Auto AI Service”

1. Introduction:

1.1 Overview:

This project represents the guided project assigned by Externship Program approved by AICTE in collaboration with IBM and SmartInternz. Cardiovascular diseases are the most common cause of death worldwide over the last few decades in the developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. In this project, we have developed and researched about models for heart disease using Auto-AI service of IBM cloud.

1.2 Purpose:

The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. This project aims to integrate are of Machine learning and cloud computing to provide the ease of development of complicated problem with cloud-based solutions.

1.3 Objective:

The Objective of this project is to

- Work with Watson Studio
- Create a project in Watson Studio
- Use Auto AI experiment to create a model
- Deploy the ML model as a webserver
- Integrating Model and Node-RED Service

- Build an application using Node-RED which takes inputs from the user and showcases the prediction on UI

2. Literature Survey:

The automatic prediction of heart disease is one of the most essential and difficult health problems in the real world. Heart disease affects the functionality of blood vessels, and causes coronary artery infections that weaken the body of the patient, especially adults and old people. The World Health Organization (WHO) has determined that more than 18 million deaths occur every year in the world due to cardiovascular diseases [3].

S. Indhumathi et.al., proposed the Naïve Bayes algorithm for high risk of heart disease prediction. The author followed data cleaning, normalization, data reduction phases in data pre-processing stage. etc. The prediction phase involves classification and prediction of disease types.[1].

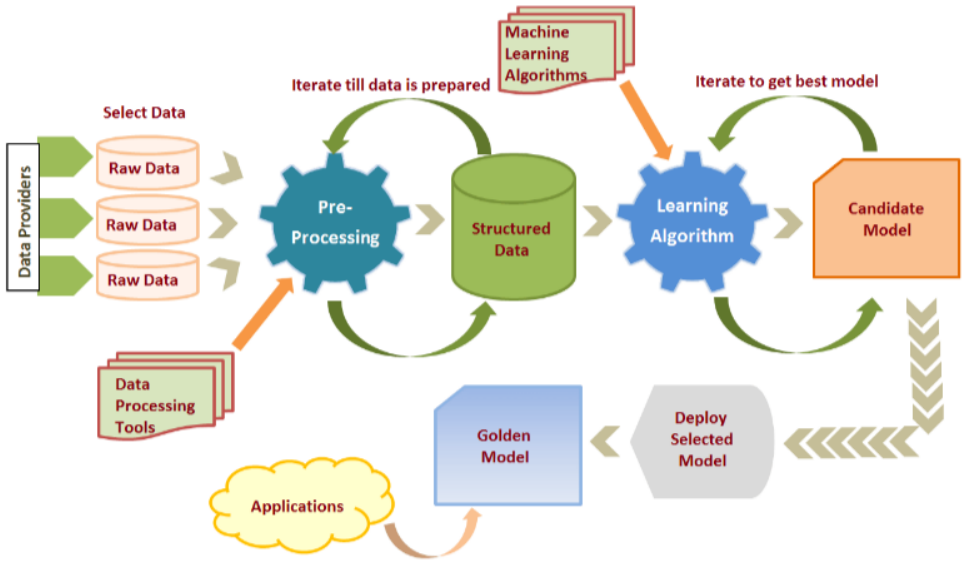
Ashok Kumar Dwivedi compared 6 machine learning techniques for heart disease prediction. These techniques were evaluated based on 8 different performance indices and ROC curve. Out of the 6 machine learning techniques logistic regression provided the promising results with classification accuracy of 85 % , 89% sensitivity and 81% specificity [2].

Vasiliy Andreevich Laptev et.al discussed some AI-based systems whose information and communication infrastructure, data processing and decision-making tools are hosted in a cloud storage service. The author also discussed some existing systems based on IBM Watson service. This study aims to determine the possible modes of AI's functioning, to identify the participants in medical-legal relations, to define the legal personality of AI and circumscribe the scope of its competencies [4].

3. Theoretical Analysis:

3.1 Block Diagram:

The following diagram shows a typical machine learning process.



1. **Selecting a data**: The quality of the data that you feed to the machine will determine how accurate your model is. If you have incorrect or outdated data, you will have wrong outcomes or predictions which are not relevant [6].

2. **Pre-processing**:

- Putting together all the data you have and randomizing it.

- Cleaning the data to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc.

- Visualize the data to understand how it is structured and understand the relationship between various variables and classes present.

- Splitting the cleaned data into two sets - a training set and a testing set. The training set is the set your model learns from. A testing set is used to check the accuracy of your model after training [6].

3. **Choose Machine Learning Algorithm:** It is important to choose a model which is relevant to the task at hand. Over the years, scientists and engineers developed various models suited for different tasks like speech recognition, image recognition, prediction, etc. Apart from this, you also must see if your model is suited for numerical or categorical data and choose accordingly [6].

4. **Training the Model:** Training is the most important step in machine learning. In training, you pass the prepared data to your machine learning model to find patterns and make predictions. It results in the model learning from the data so that it can accomplish the task set. Over time, with training, the model gets better at predicting [6].

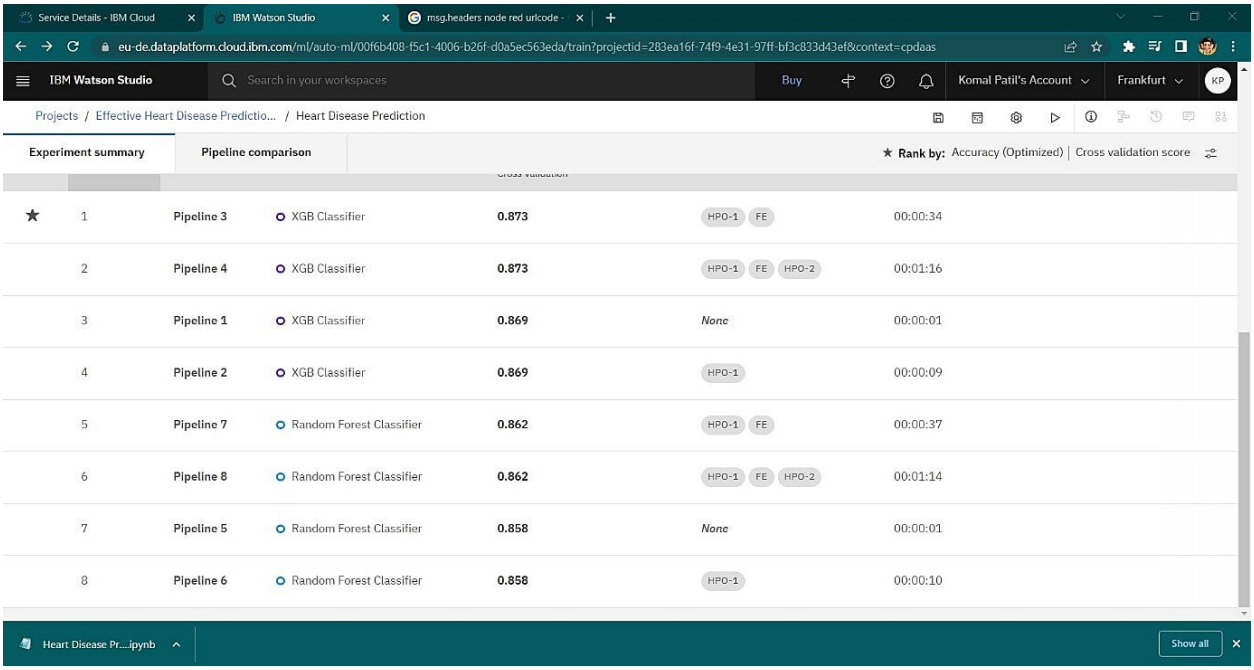
5. **Select Candidate model by evaluation:** After training your model, you must check to see how it is performing. This is done by testing the performance of the model on previously unseen data. The unseen data used is the testing set that you split our data into earlier. If testing was done on the same data which is used for training, you will not get an accurate measure, as the model is already used to the data, and finds the same patterns in it, as it previously did. This will give you disproportionately high accuracy [6].

6. **Making Predictions:** In the end, you can use your model on unseen data to make predictions accurately [6].

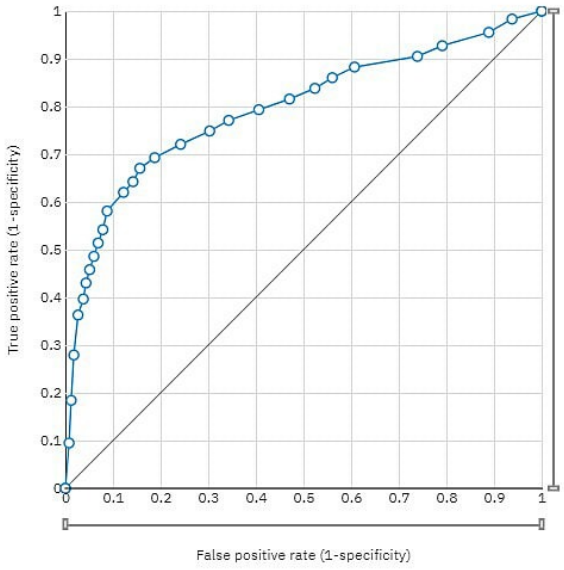
3.2 Hardware/Software Requirements:

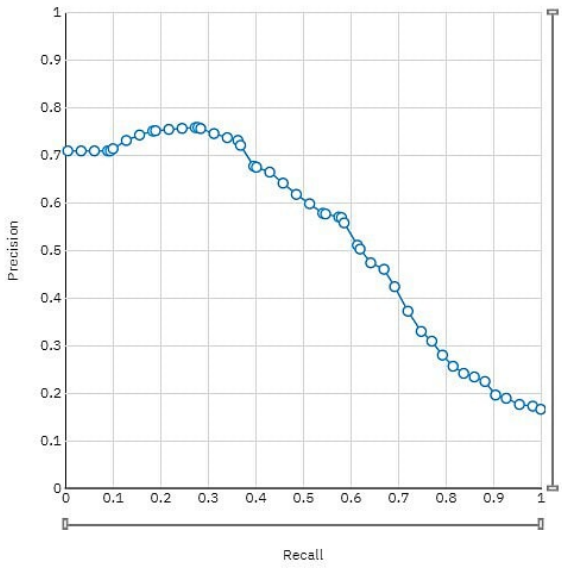
- IBM Watson Studio
- IBM Node-RED service

4. Experimental Investigation:

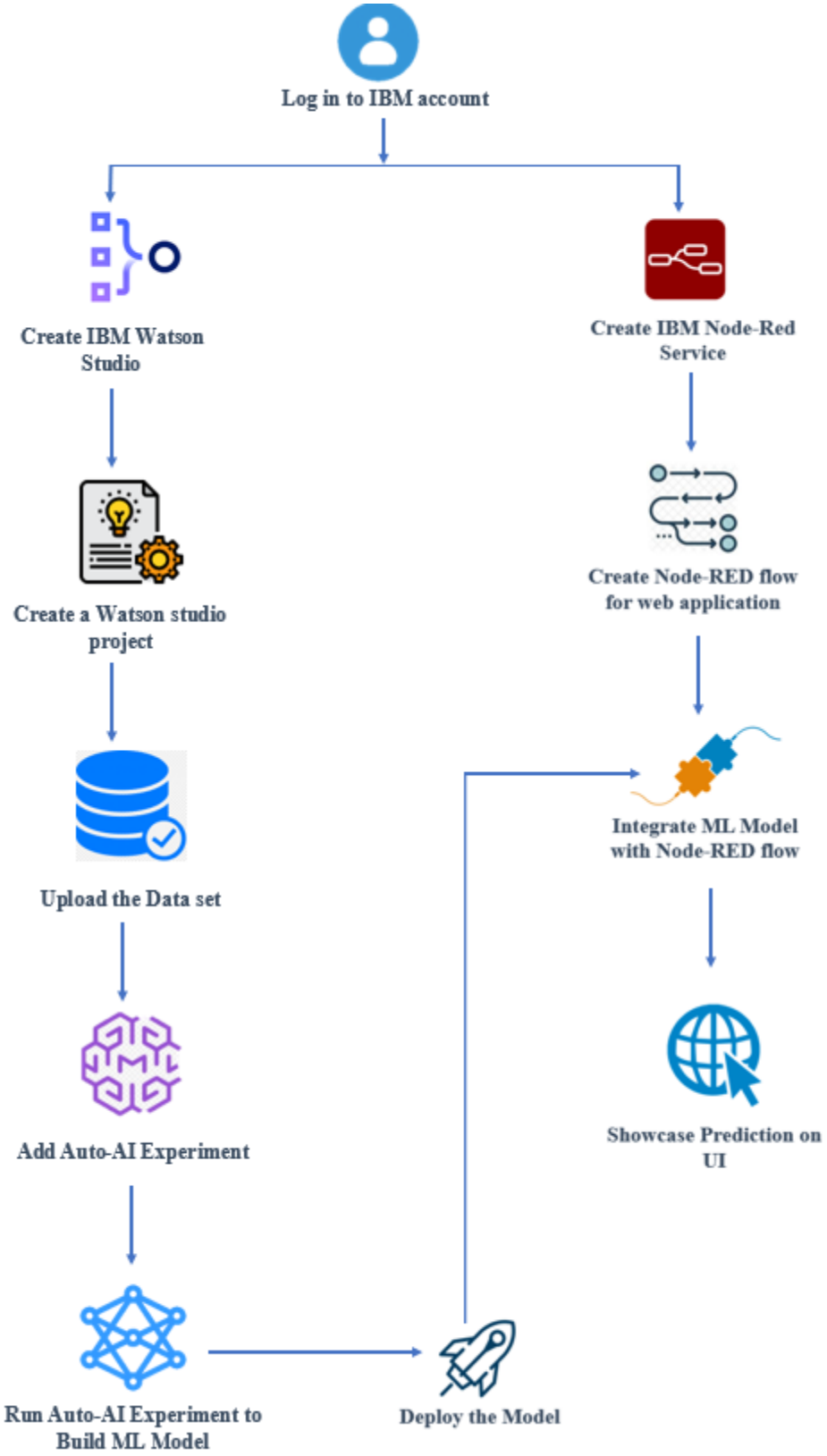


It has been observed that, XGB Classifier outperformed as compare to Random Forest Classifier with accuracy of 87.3% within response time of 34 sec and hence it has been choose for deployment. Random Forest on the other hand gave its best accuracy at 86.2%.
The ROC, precision and recall of XGB classifier is given as follows.



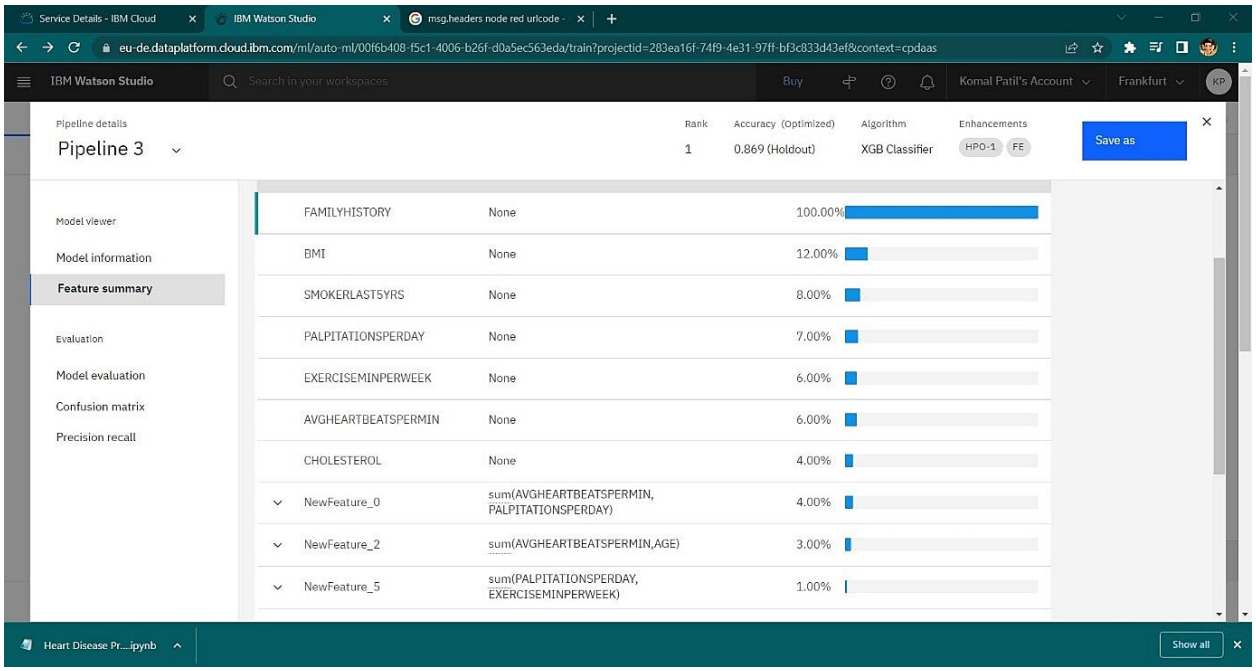


5. **Flowchart**: The workflow of the project is given below

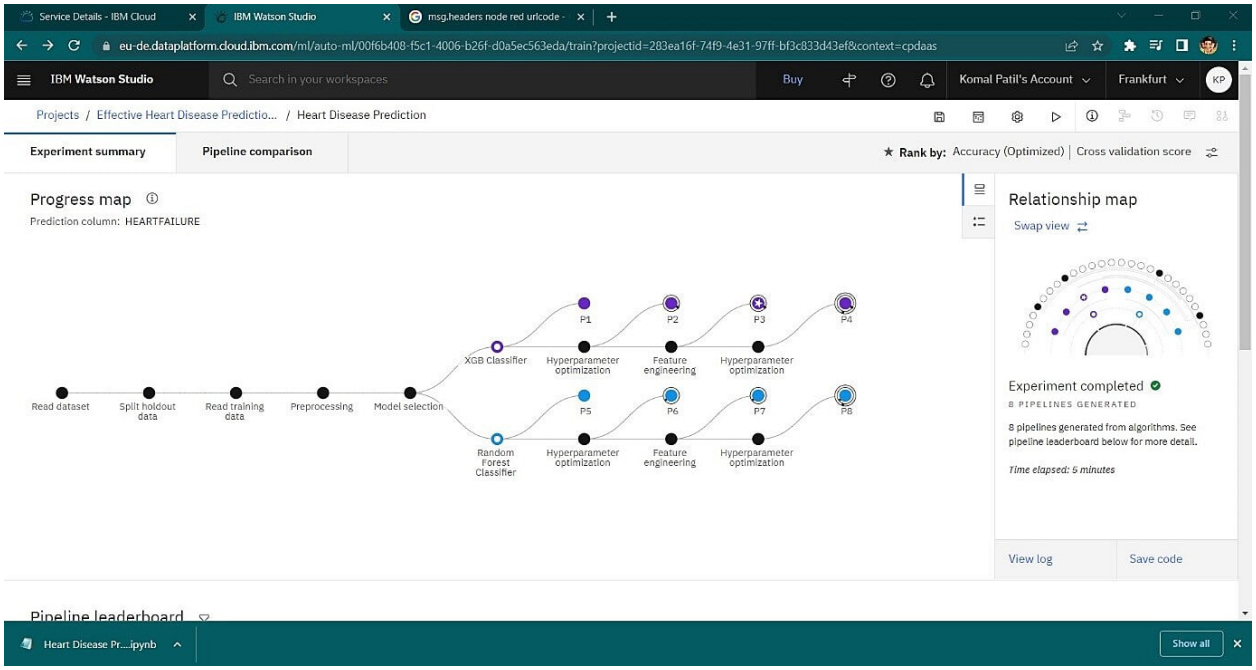


6. **Results:**

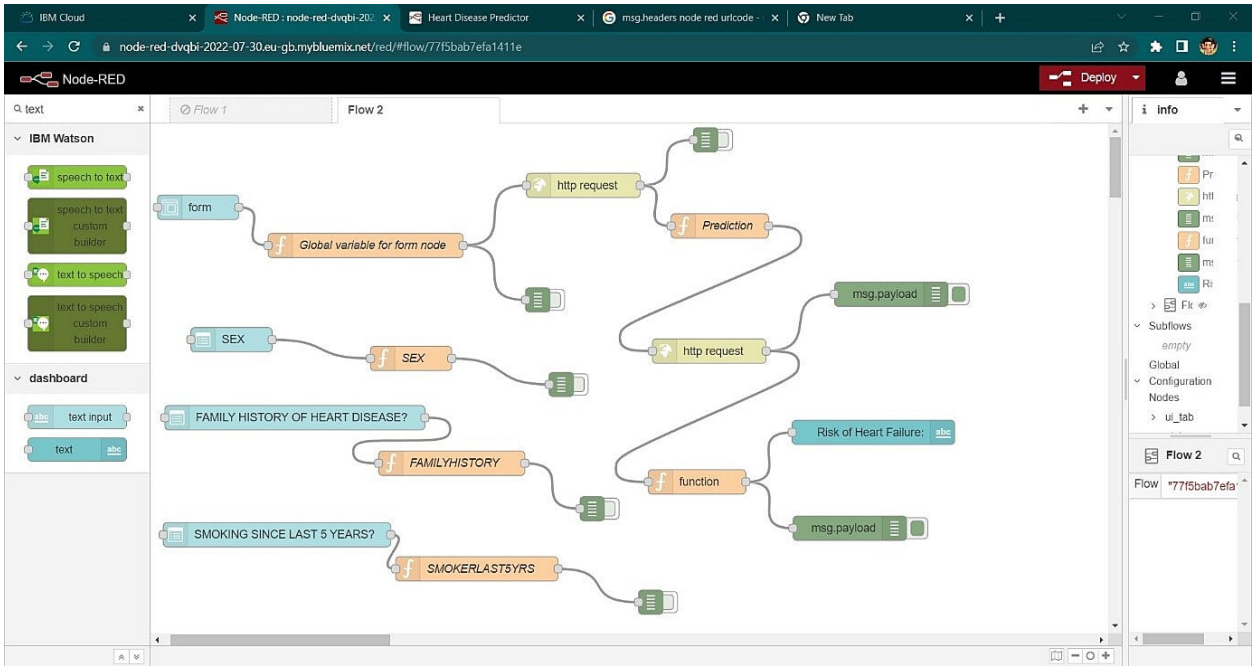
The feature summary of the dataset is given below



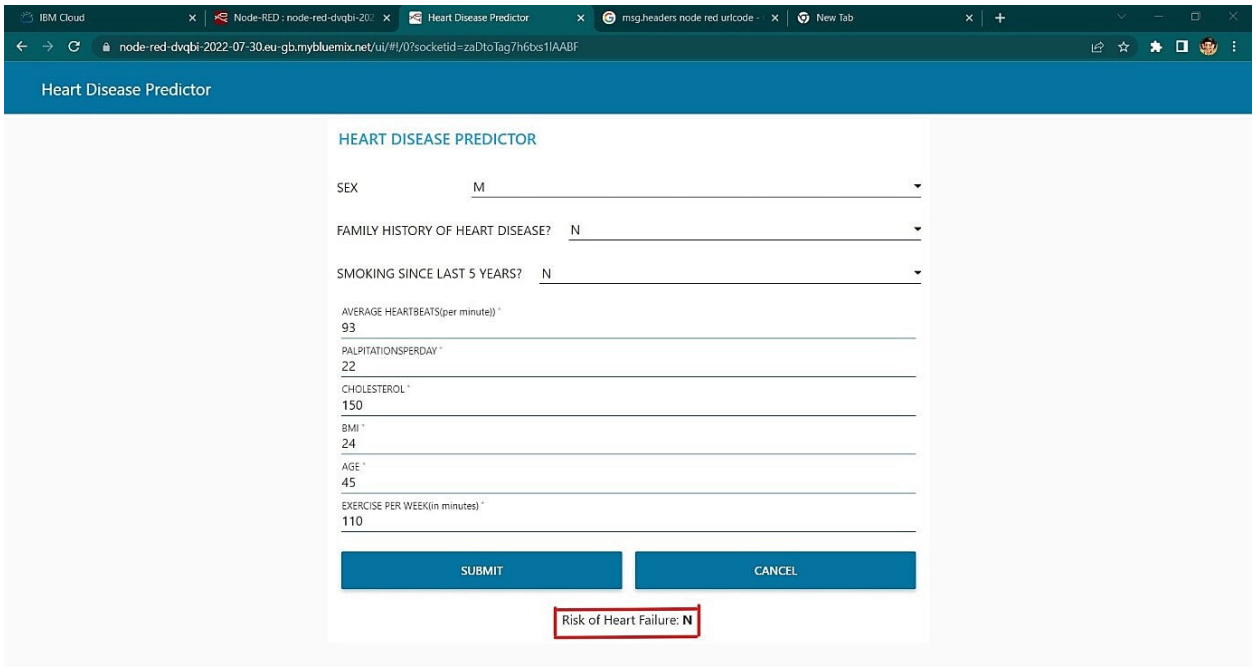
According to the feature summary, family history of heart disease is the most significant attribute for classification followed by BMI, Smoking habit, Palpitation per day, Exercise per week (in min), average heartbeat and cholesterol.



In the given pipeline, Pipeline3(XGB Classifier) gave the highest accuracy and chosen for deployment.
The following Node-RED flow has been designed for UI application; it is integrated with the trained model to provide the predictions.



The final Web application of the project is given in the following figure



7. Advantages and Disadvantages:

Advantages:

- Auto-AI enables you to build the machine learning model without any coding experience or lengthy training process.
- Because it automates tasks, we can experiment and produce results more systematically. Comparative analysis of different ML techniques has become much easier and more efficient.
- As all the services are cloud based there is no specific hardware requirement, all the services are accessible through browser.

Disadvantages:

- Customization is not possible**:** even though Auto-AI provides some customizations with hyperparameter tuning and feature selection its difficult to modify the core algorithms as per our convenience.
- Cost: though the services are free for some period the cost incurred after that is much more, one can implement the same model free of cost if uses local machine and resources.

8. Application:

The application can be used to avoid the potential cardiac arrest based on the parameters provided to it. One can track their state of wellness and if a risk is identified then can take precaution or preventive measures well before time.

9. Conclusion:

Through this guided project, we have found the risk of having heart disease can be a serious problem. We have observed the different attributes that can increase the chances of a heart disease. This study also describes the use of IBM Watson Studio, Auto-AI and Node-RED service to build and compare different machine learning models.

10. Future Scope:

The existing application can be integrated with the smart devices such as smart band/ activity tracker, daily activity data can be recorded and be given to the model. Based on the daily

activities of the user it can suggest the set of activities or diet to avoid the potential risk.

11. Bibliography:

[1] Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, “A Data mining approach for prediction of heart disease using neural network’s”, IJCET, Volume 3, Issue 3, 2012, pp. 30-40.

[2] Ashok Kumar Dwivedi “Performance evaluation of different machine learning techniques for prediction of heart disease”, © Springer, Computer Applications and Mathematics, 2016.

[3] H. Ahmed, E.M.G. Younis, A. Hendawi, A.A. Ali, Heart disease identification from patients’ social posts, machine learning solution on Spark, Futur. Gener. Comput. Syst. (2019).

[4] Vasiliy Andreevich Laptev, Inna Vladimirovna Ershova and aria Rinatovna Feyzrakhmanova, “Medical Applications of Artificial Intelligence (Legal Aspects and Future Prospects)”, Laws 2022, 11(1), 3; https://doi.org/10.3390/laws11010003.

[5] Calhas, David, Enrique Romero, and Rui Henriques. 2020. “On the use of pairwise distance learning for brain signal classification with limited observations”. Artificial Intelligence in Medicine 105: 101852.

[6]https://www.simplilearn.com/tutorials/machine-learning-tutorial/machine-learning-steps.

[7]https://dataplatform.cloud.ibm.com/docs/content/svc-welcome/wsl.html

[8]https://developer.ibm.com/learningpaths/get-started-node-red/

Appendix

Source Code: **https://github.com/smartinternz02/SI-GuidedProject-85558-1658919695**