PROJECT REPORT

1. INTRODUCTION

1.1 OVERVIEW

The project title is Effective Heart Disease Prediction using IBM Auto AI Service. In present world, Cardiovascular diseases are the main cause of death. Heart failure is a common event caused by Cardiovascular diseases. In this project, we are given a dataset which contains 9 features. They are:

1. AVGHEARTBEATSPERMIN
2. PALPITATIONSPERDAY
3. CHOLESTEROL
4. BMI
5. AGE
6. SEX
7. FAMILY HISTORY
8. SMOKERLAST5YRS
9. EXERCISEMINPERWEEK

Heart failure is very hard to detect early, but with the help of a National Institutes of Health (NIH) grant, a team of scientists at IBM Research partnered with scientists from S utter Health and clinical experts from Geisinger Health System to study and predict heart failure based on hidden clues in Electronic Health Records (EHRs). Over the last three years, using the latest advances in artificial intelligence (AI) like natural language processing, machine learning and big data analytics, the team trained models to identify heart failure one to two years earlier than a typical

diagnosis today. This research uncovered important insights about the practical trade-offs and types of data needed to train models, and developed new application methods that could allow future models to be more easily adopted.

Today, doctors will typically document signs and symptoms of heart failure in the patient record and also order diagnostic tests for heart failure. Despite best efforts, a patient is usually diagnosed with heart failure after an acute event that involves a hospitalization where the disease has advanced with irreversible and progressive organ damage.

Our team's focus was investigating if we could use the data contained in EHR systems to detect and predict a patient's risk of heart failure one or more years before a typical clinical diagnosis.

1.2 PURPOSE

using above features we can predict mortality by heart failure.
Today, doctors will typically document signs and symptoms of heart failure in the patient record and also order diagnostic tests for heart failure. Despite best efforts, a patient is usually diagnosed with heart failure after an acute event that involves a hospitalization where the disease has advanced with irreversible and progressive organ damage.

Our team's focus was investigating if we could use the data contained in EHR systems to detect and predict a patient's risk of heart failure one or more years before a typical clinical diagnosis.

2. LITERATURE SURVEY

2.1 Existing problem

There are number of works are done in predicting heart disease using various machine learning algorithm.

Sonam Nikhr has built up the paper titled as Prediction of Heart Disease using Machine Learning Algorithms. He used Naive Bayes and Decision tree classifier algorithms on same data sets, and the result uncovers that Decision tree beats over Bayesian classification system.

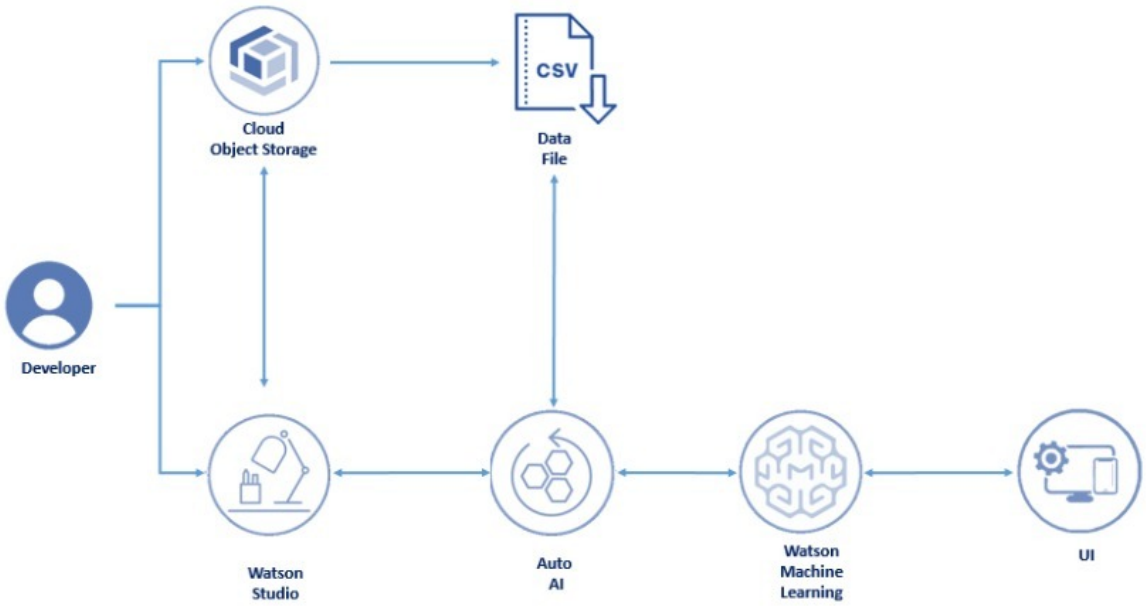
A second aim was to determine if we could more accurately predict heart failure by combining unstructured data from doctors' notes with structured EHR data. To do this, we applied machine learning methods to build predictive models that took into account a mix of variables. Our findings showed us that other data types that are routinely collected in EHRs (such as disease diagnoses, medication prescriptions and lab tests) when combined with FHFSS could be more helpful predictors of a patient's onset of heart failure.

2.2 Proposed solution

For this I used few machine learning models among them the model with best accuracy is chosen to predict the heart failure of a person. we used 8 pipelines among those 8 pipelines the one with best accuracy is saved and deployed.

3. THEORITICAL ANALYSIS

3.1 Block diagram



3.2 Hardware and software designing

Services Used:

IBM Watson Studio

IBM Watson Machine Learning

Node-RED

IBM Cloud Object Storage

4. EXPERIMENTAL INVESTIGATIONS

For the project the steps followed are-

1. Create an IBM account and then creating IBM Watson Service
2. Create an Node-red App service.
3. In Watson project we insert the dataset that is useful for prediction of heart disease and add Auto AI Experiment to the project.

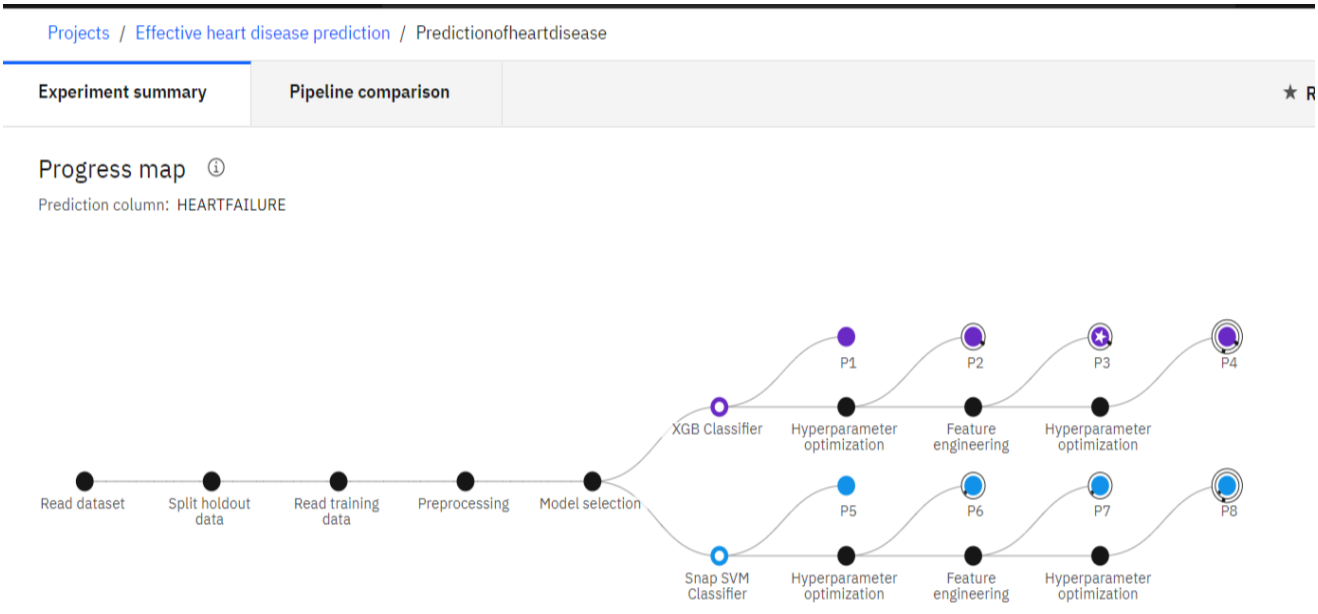
4. After adding Auto AI Experiment, we run it. While running I observed there are 8 pipelines like XGB Classifier, Snap SVM Classifier, Logistic regression and other machine learning models. Among them the one with best accuracy is chosen. In this, XGB Classifier Algorithm with Pipeline 3 provided best results among other pipelines. So we save that model and run it by giving values and predict the output whether the person is prone to heart failure or not.

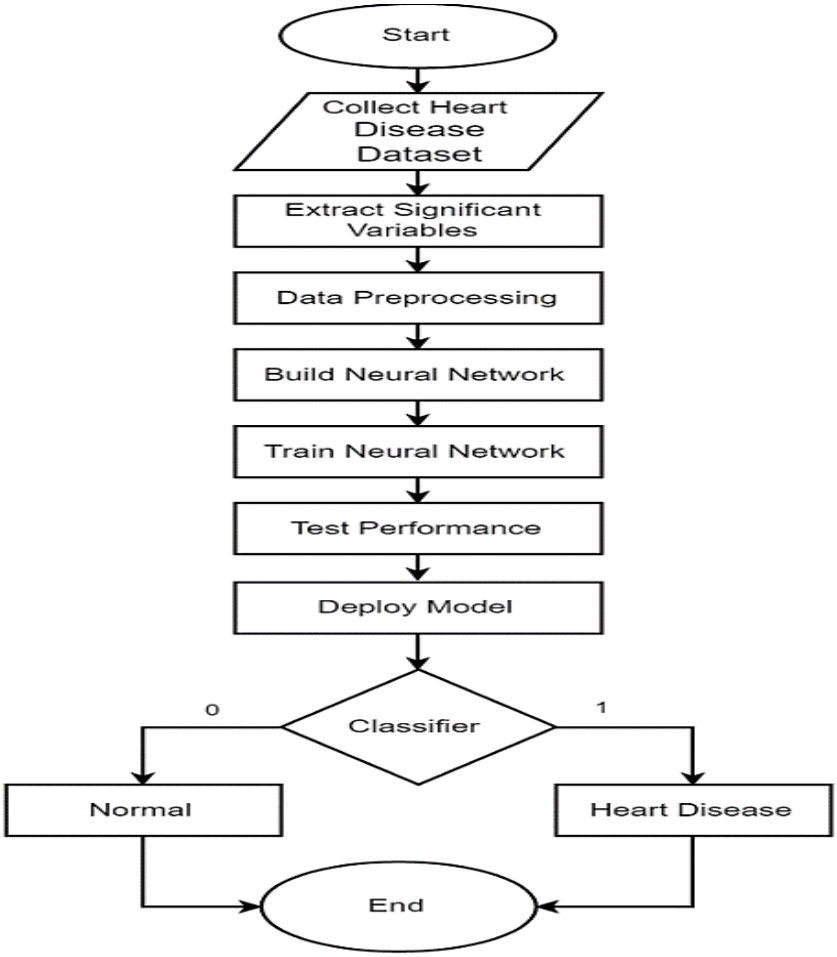
The accuracy of the model is 0.873 with build time 34sec.

Finally we integrate the Node-red with Auto AI Experiment and create an user interface.

5. FLOW CHART

Diagram showing the control flow of the solution





6. RESULT

Using IBM Watson service and Node-red service the heart disease is predicted. For this project I used 8 pipelines, among those pipelines one pipeline that is XGB Classifier pipeline gave the best result.

7. ADVANTAGES AND DISADVANTAGES

Advantages

1. Increased accuracy for effective heart disease diagnosis.
2. Handles roughest(enormous) amount of data using random forest algorithm and feature selection.
3. Reduce the time complexity of doctors.
4. Cost effective for patients.

Disadvantages

1. Prediction of cardiovascular disease results is not accurate.
2. Data mining techniques does not help to provide effective decision making.
3. Cannot handle enormous datasets for patient records.

8. APPLICATIONS

1.Using machine learning techniques, the heart disease can be predicted.

2. The medical data such as Blood pressure, hypertension, diabetes, cigarette smoked per day and so on is taken as input and then these features are modeled for prediction.

3.This model can then be used to predict future medical data.

9. CONCLUSION

The objective of our work is to provide a study of different data mining techniques that can be employed in automated heart disease prediction systems. Various techniques and data mining classifiers are defined in this work which has emerged in recent years for efficient and effective heart disease diagnosis. The analysis shows that different technologies are used in all the papers with taking different number of attributes.

Alternating decision trees when used with PCA, have performed extremely well but decision trees have performed very poorly in some other cases which could be due to overfilling. Random Forest and Ensemble models have performed very well because they solve the problem of over fitting by employing multiple algorithms (multiple Decision Trees in case of Random Forest). Models based on Naïve Bayes classifier were computationally very fast and have also performed well.SVM performed extremely well for most of the cases. Systems based on machine learning algorithms and techniques have been very accurate in predicting the heart related diseases but still there is a lot scope of research to be done on how to handle high dimensional data and over fitting. A lot of research can also be done on the correct ensemble of algorithms to use for a particular type of data.

10. FUTURE SCOPE

In Future Genetic algorithm will be used in order to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction. Prediction of the heart disease will be evaluated according to the result produced from it. Improvement is done to increase its consistency and efficiency. Benefit of using genetic algorithm is the

prediction of heart disease can be done in a short time with the help of reduced dataset.