Effective Heart Disease Prediction Using IBM Auto AI Service

INTRODUCTION

i. Overview

Heart disease is one of the major causes of life complicacies and subsequently leading to death. The heart disease diagnosis and treatment are very complex, especially in the developing countries, due to the rare availability of efficient diagnostic tools and shortage of medical professionals and other resources which affect proper prediction and treatment of patients. Inadequate preventive measures, lack of experienced or unskilled medical professionals in the field are the leading contributing factors. Although, large proportion of heart diseases is preventable but they continue to rise mainly because preventive measures are inadequate. In today's digital world, several clinical decision support systems on heart disease prediction have been developed by different scholars to simplify and ensure efficient diagnosis.

ii. Purpose

The heart is one of the most essential organs in humans. It is a kind of muscular organ which pumps blood into the body and is the central part of the body's cardiovascular system. The cardiovascular system is composed of all blood vessels such as arteries, veins, and capillaries that form a complex network of blood circulation all over the body . Any obstruction or abnormality in normal blood circulation or flow from the heart would result in several and severe complications of heart diseases. These are commonly called cardiovascular diseases (CVDs) and are among the deadliest diseases in the world. CVDs include diseases of the heart, vascular diseases of the brain, and diseases of blood vessels [48]. The World Health Organization (WHO) Report Global Atlas on Cardiovascular Disease Prevention and Control states that CVDs are the leading causes of deaths and disability in the world . Although CVDs can be prevented through life style changes and other related measures but from all indications they are still on rise on daily basis, as stated in various reports by the WHO. However, various reports by the WHO have indicated the rise of CVDs globally, which is very alarming. More people die from CVDs worldwide than from any other cause-an estimated 17.5million people in 2012 . According to another WHO report, 17.9 million people die each year from CVDs, an estimated 31% of all deaths worldwide. Of these, 85% are due to heart attack and stroke . The various reports by the WHO have indicated that deaths due to heart diseases have been on the increase, which are mainly attributed to inadequate preventive measures despite of the increasing risk factors. Medical proofs have shown that there are certain risk factors that increase a person's chances of having a cardiovascular or more specifically a heart disease. Some of these factors as enumerated by [9] include family history of cardiovascular diseases, high level of LDL (bad) cholesterol, low level of HDL (good) cholesterol, hypertension, high fat diet, lack of regular exercise, and obesity. Other risk factors include smoking, diabetes, age and gender. With these factors and more, physicians generally make diagnoses by evaluating a patient's current health status and previous diagnoses made on other patients with the same status. Cardiovascular diseases are of many types, some of which were listed by [36]: 1. Coronary Heart Diseases: Damage or disease in the major blood vessels. 2. Cardiomyopathy: An acquired or hereditary disease of the heart muscles. 3. Ischemic Heart Disease: Heart problems caused by narrowed heart arteries, which causes less blood and oxygen to reach the heart muscles. 4. Heart Failure: A chronic condition in which the heart does not pump blood as well as required. 5. Hypertensive Heart Disease: Heart problems caused by high blood pressure. 6. Inflammatory Heart Disease: Heart problems or conditions caused by viral or bacterial infections. 7. Valvular Heart Disease: Damage or defect in one of the heart valves. The increasing rate of heart diseases has become a global concern. Therefore, the healthcare industry needs to shape and intensify the way these diseases are handled in order to minimize the impact in the society. Huge data is available in the healthcare industry , more importantly the heart disease data, which needs to be efficiently analyzed for effective decision making. Based on data, statistics, clinical records and hospital management, it is claimed that in every 3 years, medical data doubles up and making health industry a multi-billion dollar domain . Machine learning and data mining techniques play a very vital role in the medical data analysis and knowledge extraction. The increasing morbidity and mortality due heart diseases worldwide has attracted the attention of researchers to conduct many studies in their effort to minimize the rates. Data mining and machine learning techniques have been widely used in the implementation of clinical decision support systems for heart disease prediction. The data mining applications are used for better health policy-making and prevention of hospital errors, early detection, prevention of diseases and preventable hospital deaths .

2. LITERATURE SURVEY

2.1. Existing problem

Reference [3] presented a heart disease prediction framework using some supervised machine learning algorithms in R programming language. The algorithms used include Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Naïve Bayes (NB). The Cleveland datasets from the University of California, Irvine (UCI) machine learning repository consisting 303 instances and 76 features were used. The data was preprocessed due to missing values and the sample became 302 instances and only 14 heart disease features in size. The data was split into 70% and 30% for models training and testing respectively. It was a comparative analysis of the selected techniques in which the experimental results showed that the NB classifier performed the heart disease prediction better than the SVM and KNN, with an accuracy of 86.6%. Reference proposed a diagnostic system for predicting heart disease using Multi-Layer Perceptron Neural network (MLP) with back propagation as the training algorithm. The performance of the developed system was evaluated based on sensitivity, specificity, precision and accuracy. The Cleveland data of the UCI machine learning repository containing 303 instances and 76 features was employed for model training and testing. Data preprocessing was performed to remove 6 instances which contain missing values. Of the 76 features, only 14 were used as the most relevant to heart disease. Based on the experiments performed, the MLP-NN proposed model gave high accuracy of 93.39% for 5 neurons in hidden layer with running time of 3.86 seconds in the heart disease prediction. Reference proposed a logistic regression (LR) based approach of machine learning for heart disease prediction. Other algorithms such as NB, SVM, DT, and KNN were also explored using SK-Learn library for performance comparisons with the LR algorithm. According to them, the experimental results showed that the LR algorithm performed better at 86.89% accuracy. While other algorithms performed at 77.85% for KNN, 86% for NB, 78.69% for DT and 82% for SVM. Datasets used for model training and testing processes were not specified. Reference [5] performed a comparative study between ANN and SVM classification algorithms based on Positive Predictive value (PPV) of cardiovascular diseases. Their data was obtained from three selected hospitals affiliated to AJA University of Medical Sciences, Iran. The sample is composed of 1324 instances and 25 features. The sample is a medical records of patients with coronary artery diseases who were hospitalized in the three mentioned hospitals between March 2016 and March 2017. The data was collected based on the variables used in the guideline of the Cleveland heart disease data policy in UCI machine learning repository. The collected data were controlled using different methods, such data preparation, integration, cleaning, normalization and reduction. The data was fed SPSS (v23.0) and Microsoft Excel 2013, then R 3.3.2 was used for statistical computing. The sample was divided into 70% and 30% for algorithm training and testing respectively. Results of their experiments showed that SVM algorithm presented higher accuracy and better performance than the ANN model, and was characterized by higher power and sensitivity. Reference studied the case of predicting the risk of cardiovascular diseases (CVDs) by comparing Auto Machine Learning techniques against a graduate student using several important metrics, including the total amount of time required for building machine learning models and the final classification accuracies on unseen test datasets. They proposed Auto-SKlearn model, which was motivated by Scikit-Learn, a popular generic machine learning toolbox. Their model utilizes a large number of machine learning classifiers and preprocessing steps in the Scikit-learn toolbox. The classifiers used include LR, SVM, RF, Boosting, NN, and KNN. Given the training data, Auto-SKlearn first selects an appropriate set of data preprocessing steps, such as imputation of missing values. It then passes the processed data to feature processing block which further normalizes the data or reduces their dimensions using standard techniques, principal component analysis. Finally, the datasets are passed to the estimator block which selects and trains machine learning algorithms to predict desirable outputs from input data samples. Training, testing and evaluation were performed on two different cardiovascular disease datasets. The Cleveland heart disease datasets from the UCI machine learning repository, comprising 303 instances with 76 features from which only 13 were used. The other datasets were CVD data (source not specified) comprising 70,000 instances with 11 features. In the experiments, each dataset category was split into three: training set, validation set and testing set. The 303 instances of the UCI data was divided into 100 for testing, and 203 for training and validation. Kfold cross validation method was adopted. The 70,000 records of the CVD data was divided into 14,000 for testing and 56,000 for training and cross validation. According to the results of comparative analysis performed on the two different datasets, the Auto-ML model took only 30 minutes to build a competitive classifier for each dataset, compared to long periods of time (432 hours for UCI data and 360 hours for CVD datasets) taken by the graduate student to develop similar classifiers. The Auto-ML model is too slow and inefficient, taking 30 minutes to build a classifier. Reference [21] presented a machine learning-based technique for detection of heart disease using sampling techniques to handle unbalanced datasets. The sampling techniques used include Random Over-Sampling, Synthetic Minority Over-Sampling (SMOTE) and Adaptive Synthetic Sampling Approach (ADASYN). Framingham datasets from the Kaggle website, which contains 4239 instances with 15 features were used for the algorithm training and testing. Based on the features, the aim was to predict whether a patient had a 10-year risk of future coronary heart disease. The machine learning techniques used include LR, KNN, AdaBoost, DT, NB, and RF. The performances of these classification algorithms were measured and evaluation based on precision, recall, and accuracy. Each of these parameters varies according to the sampling technique used. From their experimental results, SVM classifier with Random OverSampling technique appeared the best in the heart disease prediction with an accuracy of 99%. However, RF performed better with SMOTE technique at 91.3% accuracy while DT classifier and RF again performed better with ADASYN technique at 90.3% accuracy. Therefore, the classification accuracy of this approach was solely based on the sampling techniques, which are not always necessary in all types of datasets.

i. Proposed solution

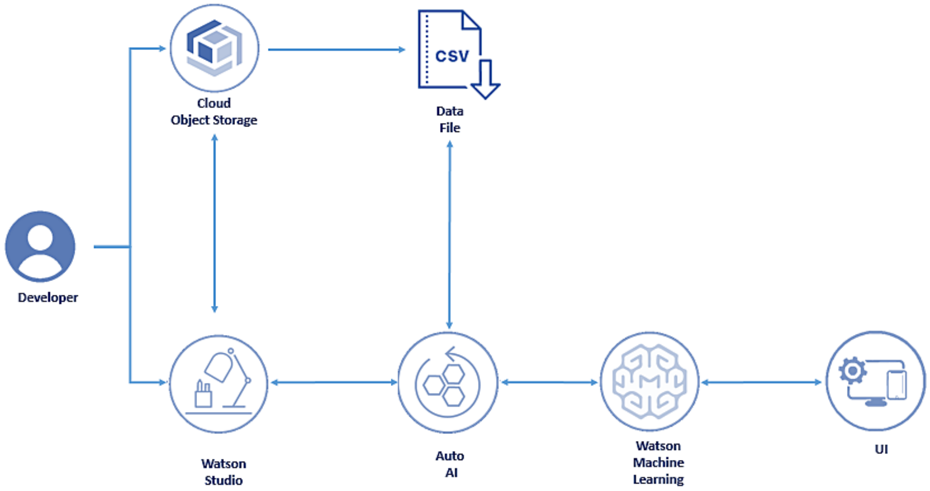
Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide.

Heart failure is a common event caused by CVDs and this dataset contains 9 features that can be used to predict mortality by heart failure.

In this project,we need to build a model using Auto AI and build a web application where we can showcase the prediction of heart failure.

3. THEORITICAL ANALYSIS

3.1. Block diagram



3.2. Hardware / Software designing

- ➢ 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

- IBM Watson Studio

- IBM Watson Machine Learning
- Node-RED
- IBM Cloud Object Storage

4. EXPERIMENTAL INVESTIGATIONS

1. Log in to IBM account
2. Create IBM Watson Studio and Node-RED Service
    - Watson studio
    - Node-RED
    - Cloud Object Storage service (COS)
    - Machine Learning service (ML)
3. Create a Watson studio project
4. ADD Auto AI Experiment
5. Run the Auto AI Experiment to build a Machine learning model on the desired dataset
6. Save the model
7. Deploy the model as a web server and generate scoring End Point
8. Create a WEB application Using Node-RED to take user input and showcase Prediction on UI

In this activity, you will learn how you can save a pipeline as a Watson Machine Learning model, deploy the model, and score it to view a prediction.

This Activity contains the Following Tasks

- **Collect the data set**

In this Activity, We gonna build a machine learning model that predicts heart failure based on the following parameters

- AVGHEARTBEATSPERMIN
- PALPITATIONSPERDAY
- CHOLESTEROL
- BMI
- AGE
- SEX
- FAMILY HISTORY
- SMOKERLAST5YRS

- EXERCISEMINPERWEEK
- Create Watson Studio project
- **Add Auto AI experiment**

  The AutoAI graphical tool in Watson Studio automatically analyzes your data and generates candidate model pipelines customized for your predictive modeling problem. These model pipelines are created iteratively as AutoAI analyzes your dataset and discovers data transformations, algorithms, and parameter settings that work best for your problem setting. Results are displayed on a leaderboard, showing the automatically generated model pipelines ranked according to your problem optimization objective.

To add the project

- Click on Add Project from Watson Studio project
- Select Auto AI
- Specify a name and description for your experiment.
- Select a machine learning service instance and click Create.
- Run AI Experiment
- Save the model

Once the pipeline creation is complete, you can view and compare the ranked pipelines in a leaderboard.

Choose Save model from the action menu for the pipeline with the highest accuracy or low error rate This saves the pipeline as a Machine Learning asset in your project. A notification gives you the link to view the saved model in your project.

- **Deploy the model**

  Before you can use your trained model to make predictions on new data, you must deploy the model.You can deploy the model from the model details page. You can access the model details page in one of these ways:
- Click on the model name in the notification displayed when you save the model.
- Open the Assets page for the project containing the model and click the model name in the Machine Learning Model section.

From the model details page:

- Click the Promote to deployment space.
- Choose an existing deployment space or create a new one.
- Click Add Deployment.
- In the page that opens, fill in the fields:
- Specify a name for the deployment.
- Select “Web service” as the Deployment type.

- Click Save.

Build Node-RED Application

Let's build a User interface which takes inputs from the user. The Model Analyses the Inputs and returns the Prediction that is showcased on the User interface.

5. RESULT

https://drive.google.com/drive/u/0/folders/1yLWwCSwNGrwfs5uZIrYvkB3Iq13f5V2c

6. ADVANTAGES & DISADVANTAGES

It will be more useful in remote area maily for Trible people.

Also the network facility needed is the disadvantage.

7. APPLICATIONS

From a clinical point of view this is critical – there could be more than 1,000 patient factors used in a model, but no healthcare professional will want to adopt something that will require such an extensive number of variables to be input. These findings suggest possible guidelines for the minimum amount and type of data needed to train effective predictive disease models.

8. CONCLUSION

This project led us to a deeper understanding about the tradeoffs between certain data types and their usefulness in helping detect an individual's likelihood of heart failure. For example, we found that the model's performance improved when more diverse data types are used, but, the combination of diagnosis, medication order, and hospitalization data was most important, respectively. We leveraged knowledge-driven ontologies of medications and diagnoses to summarize variables into higher level concepts and developed data-driven methods to identify and select the most salient variables to create a smaller and more robust subset of variables. This led us to develop predictive models that were high in both performance and practicality.

9. FUTURE SCOPE

By using Auto AI model,we can build humaoid robot for mility purpose.

BIBILOGRAPHY

https://www.researchgate.net/publication/344998779_A_Comprehensive_Review_on_Heart_Disease_Prediction_Using_Data_Mining_and_Machine_Learning_Techniques.

1. Alotaibi, F. S. (2019). Implementation of machine learning model to predict heart failure disease. *International Journal of Advanced ComputerScience and Applications*, *10* (6), 261-268.
2. Amin, M. S., Chiam, Y. K., & Varathan,K. D. (2018). Identification of significant featuresand data mining techniques in predicting heart disease. *Telematics and Informatics*. doi: 10.1016/J.TELE.2018.11.007.
3. Anitha, S., & Sridevi,N. (2019). Heart disease prediction using data mining techniques. *Journal of Analysis and Computation*, *8* (2), 48-55.
4. Annepu, D., & Gowtham, G. (2019). Cardiovascular disease prediction using machine learning techniques. *International ResearchJournal of Engineering and Technology*, *6* (4), 3963-3971.
5. Ayatollahi, H., Gholamhosseini, L., & Salehi,M. (2019). Predicting coronary artery disease: a comparison between two data mining algorithms. *BMC Public Health*. doi: 10.1186/S12889-019-6721-5.
6. Banu, G. R., & Jamala, J. H. (2015). Heart attack prediction using data mining technique. *International Journal of Modern Trends in Engineering and Research*, *2* (5), 428-432.
7. Benjamin, H., David, F., & Belcy, S. A. (2018). Heart disease prediction using data mining techniques. *ICTACT Journal of Soft Computing*, *9* (1), 1824-1830.
8. Chaithra, N., & Madhu, B. (2018). Classification models on cardiovascular diseaseprediction using data mining techniques. *Journal of Cardiovascular Diseasesand Diagnosis*. doi: 10.4172/2329-9517.1000348.
9. D'Souza, A. (2015). Heart disease prediction using data mining techniques. *International Journalof Research in Engineering and Science*, *3* (3), 74-77.
10. Devi, S. K. (2016).Prediction of heart disease using data mining techniques. *Indian Journal of Science and Technology*. doi: 10.17485/ijst/2016/v9i39/102078.
11. Dulhare, U. N. (2018). Prediction system for heart disease using naïve bayes and particle swarm optimization. *Biomedical Research*, *29* (12), 2646-2649.
12. Gawali, M., & Shirwalkar, N. (2018). Heart disease prediction system using data mining techniques. *International Journal of Pure and Applied mathematics*, *120* (6), 499-506.
13. Haq, A. U., Li, J.-P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Hindawi Mobile Information System*. doi: 10.1155/2018/3860146.
14. Hariharan, K., Vigneshwar, W. S., Sivaramakrishnan, N., & Subramaniyaswamy, V. (2018). A comparative study on heart disease analysis using classification techniques. *InternationalJournal of Pure and Applied Mathematics*, *119* (12), 13357-13366.
15. Hussein, M. U. (2017, October 29). *Physics and the Cardiovascular System*.Retrieved from ResearchGate:

https://www.researchgate.net.

16. Jagtap, A., Malewadkar, P., Baswat, O., & Rambade,H. (2019). Heart disease prediction using machine learning.*International Journalof Research in Engineering*, *Scienceand Management*, *2* (2), 352-355.

17. Kashyap, A. (2018). Artificial intelligence and medicaldiagnosis. *Scholars Journal of Applied MedicalSciences*, 4982-4985. doi: 10.21276/sjams.2018.6.12.61.

18. Khan, S. N., Nawi, N. M., Shahzad, A., Ullah, A., &Mushtaq, M. F. (2019).Comparative analysis for heart diseaseprediction. *International Journal on Informatics Visualization*, *1* (4-2), 227-231.

19. Khourdifi, Y., & Bahaj, M. (2018). Heart disease prediction and classification using machine learningalgorithms optimized by particle swarm optimization and ant colony optimization. *International Journal of Intelligent Engineering and Systems,12* (1).

20. Kim, J. K., & Kang, S. (2017). Neuralnetwork-based coronary heart disease risk prediction using feature correlation analysis. *Hindawi Journal of Healthcare Engineering*. doi: 10.1155/2017/2780501.

APPENDIX

A. SourceCode

https://drive.google.com/drive/u/0/folders /1yLWwCSwNGrwfs5uZIrYvkB3Iq13f5V2c.

Note: Limit the report to 15 pages.