MIT-ADT
UNIVERSITY
PUNE, INDIA
A leap towards World Class Education

A PROPOSED IMPLEMENTATION OF END TO END MACHINE LEARNING PROJECT FOR DISEASE PREDICTION

SUBMITTED BY -

VINCE D’SOUZA (2203041)

TANUJ SINHA (2203046)

AADITYA NAIR (2203391)

ABHISHEK TEKAVADE (2203003)

Submitted in partial fulfillment of the requirements for qualifying

MINI PROJECT IV – SEM VI

ACKNOWLEDGEMENT

First and foremost I would like to thank the Lord Almighty for His presence and immense blessings throughout the project work. I wish to express my heartfelt gratitude to Dr Ganesh Pathak, Head of the Department, School of Computation (SOC) for much of his valuable support encouragement in carrying out this work. I would like to thank my internal guide Prof. Amol Dande Sir for continually guiding and actively participating in my project, giving valuable suggestions to complete the project work. I would like to thank all the technical and teaching staff of the School of Computation (SOC), who extended directly or indirectly all support.

Last, but not least, I am deeply indebted to my parents who have been the greatest support while I worked day and night for the project to make it a success.

By: -

Vince D’Souza (2203321)

Tanuj Sinha (2203046)

Aaditya Nair (2203391)

Abhishek Tekavade (2203003)

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 REALITY OF DISEASES

Now-a-days, people face various diseases due to the environmental condition and their living habits. Especially due to Covid, one never knows what problem he can encounter. So, the prediction of disease at earlier stage becomes important task. Disease prediction using patient treatment history and health data by applying data mining and machine learning techniques is ongoing struggle for the past decades. The correct prediction of disease is the most challenging task. The recent success of deep learning in disparate areas of machine learning has driven a shift towards machine learning models that can learn rich, hierarchical representations of raw data with little pre-processing and produce more accurate results.

With the help of disease data, Machine Learning finds hidden pattern information in the huge amount of data, which in turn helps the patient to identify the problem. The main focus is on to use machine learning in healthcare to supplement patient care for better results. Machine learning has made easier to identify different diseases and diagnosis correctly. Predictive analysis with the help of efficient multiple machine learning algorithms helps to predict the disease more correctly and help treat patients. Machine learning in healthcare aids the humans to process huge and complex medical datasets and then analyse them into clinical insights. This then can further be used by physicians in providing medical care. Hence machine learning when implemented in healthcare can leads to increased patient satisfaction.

1.2 Motivation behind the idea

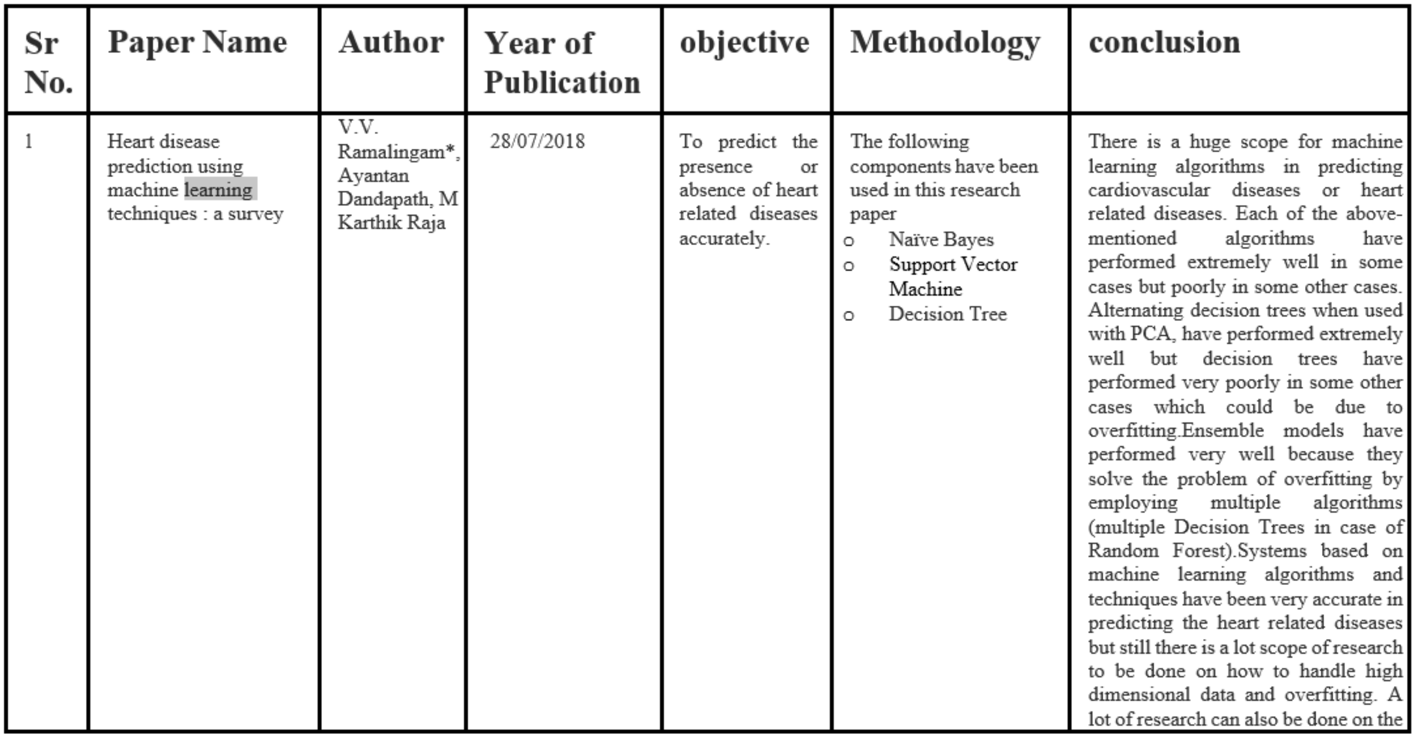
There is a demand to make such a system that will help end users to predict diseases on the basis of symptoms given in it without visiting hospitals.

By doing so, it will decrease the rush at OPD‟s of hospitals and bring down the workload on medical staff. Not only this, this system will reduce the costly treatment and panic moment at the end stages so that proper medication can be provided at the right time and we can lower down the death rate as well.

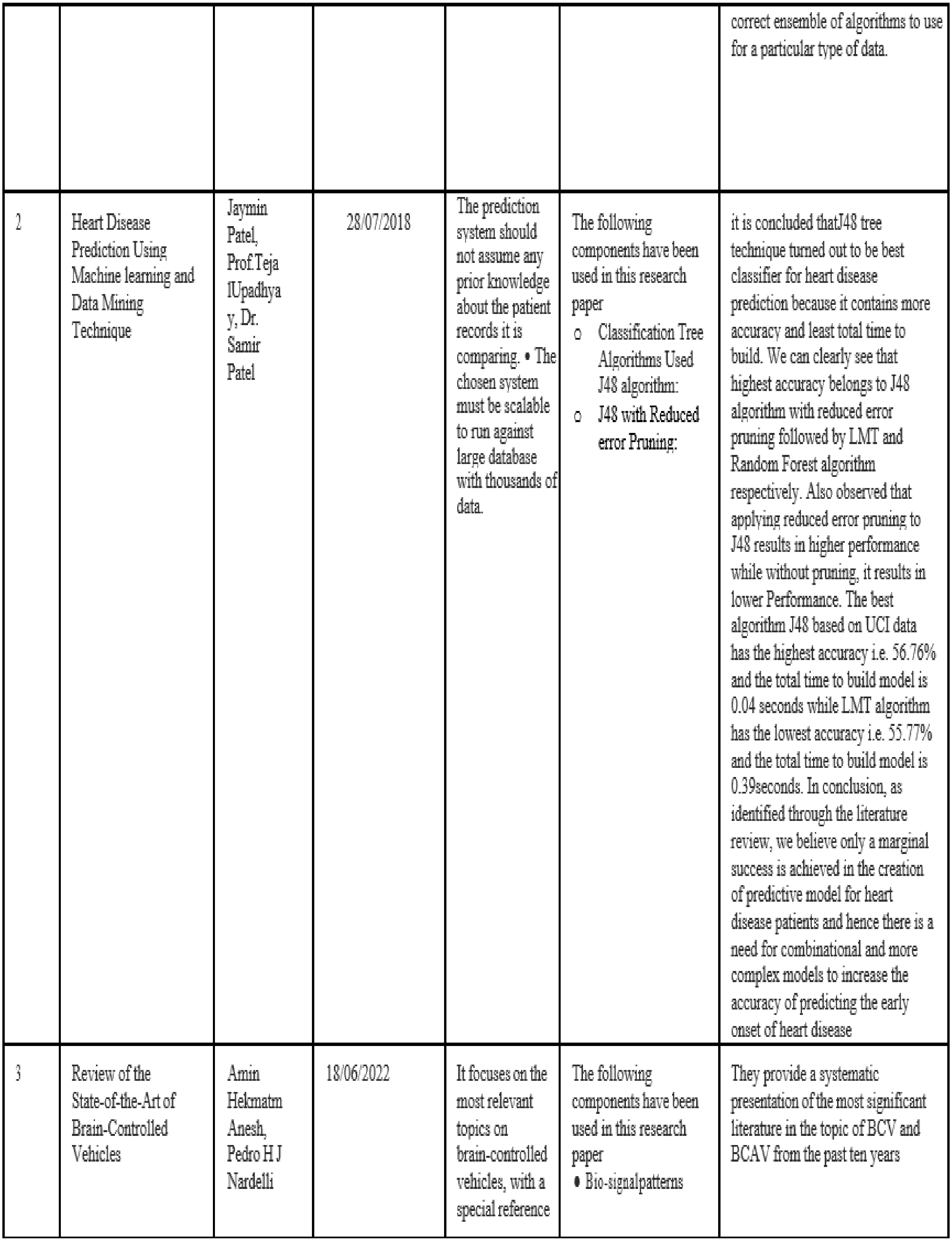
This system also consists of a feature of Database which stores the data entered by the end users and the name of the disease the patient is suffering from that can be used as a past record and will help in further treatment in future. The analysis accuracy is increased by using Machine Learning algorithms. Altogether this system will help in easier health management.

CHAPTER 2

LITERATURE SURVEY



Sr No.	Paper Name	Author	Year of Publication	objective	Methodology	conclusion
1	Heart disease prediction using machine learning techniques : a survey	V.V. Ramalingam*, Ayantan Dandapath, M Karthik Raja	28/07/2018	To predict the presence or absence of heart related diseases accurately.	The following components have been used in this research paper o Naïve Bayes o Support Vector Machine o Decision Tree	There is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart related diseases. Each of the above-mentioned algorithms have performed extremely well in some cases but poorly in some other cases. 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 overfitting.Ensemble models have performed very well because they solve the problem of overfitting by employing multiple algorithms (multiple Decision Trees in case of Random Forest).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 overfitting. A lot of research can also be done on the



						correct ensemble of algorithms to use for a particular type of data.
2	Heart Disease Prediction Using Machine learning and Data Mining Technique	Jaymin Patel, Prof.Teja lUpadhyay, Dr. Samir Patel	28/07/2018	The prediction system should not assume any prior knowledge about the patient records it is comparing. • The chosen system must be scalable to run against large database with thousands of data.	The following components have been used in this research paper ○ Classification Tree Algorithms Used J48 algorithm: ○ J48 with Reduced error Pruning:	it is concluded thatJ48 tree technique turned out to be best classifier for heart disease prediction because it contains more accuracy and least total time to build. We can clearly see that highest accuracy belongs to J48 algorithm with reduced error pruning followed by LMT and Random Forest algorithm respectively. Also observed that applying reduced error pruning to J48 results in higher performance while without pruning, it results in lower Performance. The best algorithm J48 based on UCI data has the highest accuracy i.e. 56.76% and the total time to build model is 0.04 seconds while LMT algorithm has the lowest accuracy i.e. 55.77% and the total time to build model is 0.39seconds. In conclusion, as identified through the literature review, we believe only a marginal success is achieved in the creation of predictive model for heart disease patients and hence there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of heart disease
3	Review of the State-of-the-Art of Brain-Controlled Vehicles	Amin Hekmatm Anesh, Pedro H J Nardelli	18/06/2022	It focuses on the most relevant topics on brain-controlled vehicles, with a special reference	The following components have been used in this research paper ● Bio-signalpatterns	They provide a systematic presentation of the most significant literature in the topic of BCV and BCAV from the past ten years

CHAPTER 3

PROBLEM STATEMENT & OBJECTIVES

PROBLEM STATEMENT –

In day to day life, it is difficult for one person to go to a doctor and get a check of diseases as it takes quite a lot amount of money as well as time. In our developing and technology dependent life we totally rely on gadgets. So, there should be a way, with whose help a person can at least check whether he has a particular disease or not. Using tech like machine learning in predicting the diseases using symptoms or concerned medical data is a need of the hour.

OBJECTIVES –

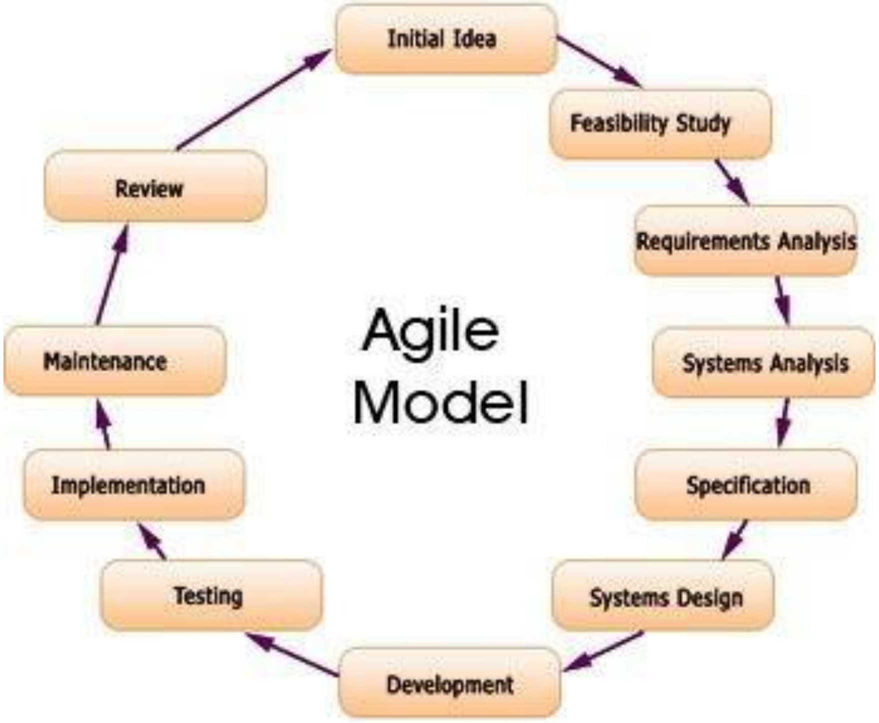
The aim of this study is to test the proposed hypothesis that supervised ML algorithms can improve health care by the accurate and early detection of diseases. In this study, we investigate studies that utilize more than one supervised ML model for each disease recognition problem. This approach renders more comprehensiveness and precision because the evaluation of the performance of a single algorithm over various study settings induces bias which generates imprecise results. The analysis of ML models will be conducted on few diseases located at heart, kidney, breast, and brain. For the detection of the disease, numerous methodologies will be evaluated such as RF, DT, SVM, and LR. At the end of this literature, the best performing ML models in respect of each disease will be concluded.

CHAPTER 4

SYSTEM ARCHITECTURE

4.1 AGILE MODEL

Agile model is selected for the project. We are planning to implement the system with basic facilities only. So many future enhancements are possible with this model. Agile model can satisfy this requirement efficiently. Since it follows the plan-do-check-act for improvement, backtracking can be done easily in Agile model.



Agile modelling is a practice-based methodology for effective modelling and documentation of software-based systems. This can be applied on a software development project in an effective and light-weight manner. An Agile Model Driven Development (AMDD) approach enables high-level modelling at the beginning of a project to understand the scope and potential architecture of the system, and then during development iterations, it requires modelling as part of iteration planning activities and then requires just in time (JIT) model storming approach.

1.2 HARDWARE AND SOFTWARE REQUIREMENTS –

- **Hardware requirements-**

  Dual 2GHz+ CPU

  2GB+ RAM

  80MB database space

  1GB disk space.

- **Software requirements-**

  Web Browser (Chrome, Mozilla Firefox, Internet Explorer, etc.) Internet connection.

CHAPTER 5

TECHNOLOGIES USED

PYTHON, ML, STREAMLIT, NUMPY

5.1 PYTHON

Python is a high-level, interpreted programming language that was first released in 1991. It is designed to be easy to read and write, with a simple and straightforward syntax that makes it accessible to beginners, yet powerful enough to be used for complex applications.

Python is a versatile language that is widely used in a variety of fields, including web development, data science, machine learning, and artificial intelligence. It has a large and active community of developers who contribute to its development and create useful libraries and frameworks that extend its capabilities.

One of the key strengths of Python is its ease of use and readability, which makes it an ideal language for teaching programming to beginners. It is also known for its high productivity, as it allows developers to write code quickly and efficiently.

Python is an open-source language, which means that the source code is freely available and can be modified and distributed by anyone. This has contributed to its popularity and widespread adoption, as developers can use Python for a wide range of projects without having to pay for licenses or worry about proprietary restrictions.

5.2 ML

ML stands for machine learning, which is a subset of artificial intelligence (AI) that involves the use of algorithms and statistical models to enable computer systems to learn from data and improve their performance on specific tasks over time, without being explicitly programmed.

Machine learning algorithms can be broadly categorized into three types: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning involves training a model on a labeled dataset, where the desired output is already known, and then using this model to predict new outputs for new inputs. Common applications of supervised learning include image classification, sentiment analysis, and speech recognition.

Unsupervised learning, on the other hand, involves training a model on an unlabeled dataset and having it identify patterns or structures in the data. This type of learning is often used for clustering or anomaly detection.

Reinforcement learning involves training a model through trial-and-error interactions with an environment, where the model learns to take actions that maximize a reward signal. This type of learning is often used in game AI or robotics.

Machine learning has numerous applications in a variety of fields, including finance, healthcare, marketing, and entertainment. Some popular machine learning libraries and frameworks include scikit-learn, Tensor Flow, and PyTorch.

5.3 STREAMLIT

Streamlit is an open-source Python library that enables developers to quickly build web applications for data science and machine learning projects. It allows developers to create interactive web applications without having to learn complex web frameworks, such as Django or Flask.

With Streamlit, developers can write code using familiar Python libraries, such as Pandas and Matplotlib, and then easily deploy and share their applications on the web. Streamlit provides a simple and intuitive interface for creating interactive data visualizations, such as charts, maps, and graphs, as well as for creating user input forms and widgets.

Streamlit is built on top of Flask and leverages web sockets to enable real-time updates and interactive components in the web application. It is designed to work seamlessly with popular data science and machine learning libraries, such as Tensor Flow, PyTorch, and Scikit-learn.

Streamlit provides a number of built-in components and widgets, such as sliders, checkboxes, and dropdowns, that can be used to create interactive user interfaces for your application. It also provides an easy way to deploy your application to the web, either through Streamlit's cloud platform or by deploying to a cloud service provider, such as Amazon Web Services or Google Cloud Platform.

Streamlit has gained popularity among data scientists and machine learning engineers because of its ease of use and rapid development capabilities, which can help to accelerate the prototyping and experimentation phases of a data science project.

5.4 NUMPY

NumPy is a Python library that is used for numerical computing. It provides a powerful array object that can be used to perform mathematical operations on large datasets quickly and efficiently. NumPy is often used in data science, scientific computing, and machine learning applications.

NumPy's main feature is its ndarray (n-dimensional array) object, which is a multidimensional array that can be used to store and manipulate large datasets. The ndarray object provides a wide range of mathematical functions and operations, such as element-wise arithmetic, linear algebra operations, and statistical analysis.

NumPy also provides a number of other useful features, such as the ability to create random arrays and matrices, the ability to slice and reshape arrays, and the ability to perform broadcasting, which allows for efficient computation on arrays of different shapes and sizes.

In addition to the ndarray object, NumPy provides a number of other tools and functions for numerical computing, such as Fourier transforms, signal processing, and optimization. NumPy also integrates well with other Python libraries, such as Pandas and Matplotlib, making it a valuable tool for data analysis and visualization.

NumPy is widely used in scientific and engineering applications, such as physics, chemistry, and biology, as well as in finance, economics, and other fields where large amounts of numerical data are processed. Its speed and efficiency make it a popular choice for high-performance computing applications.

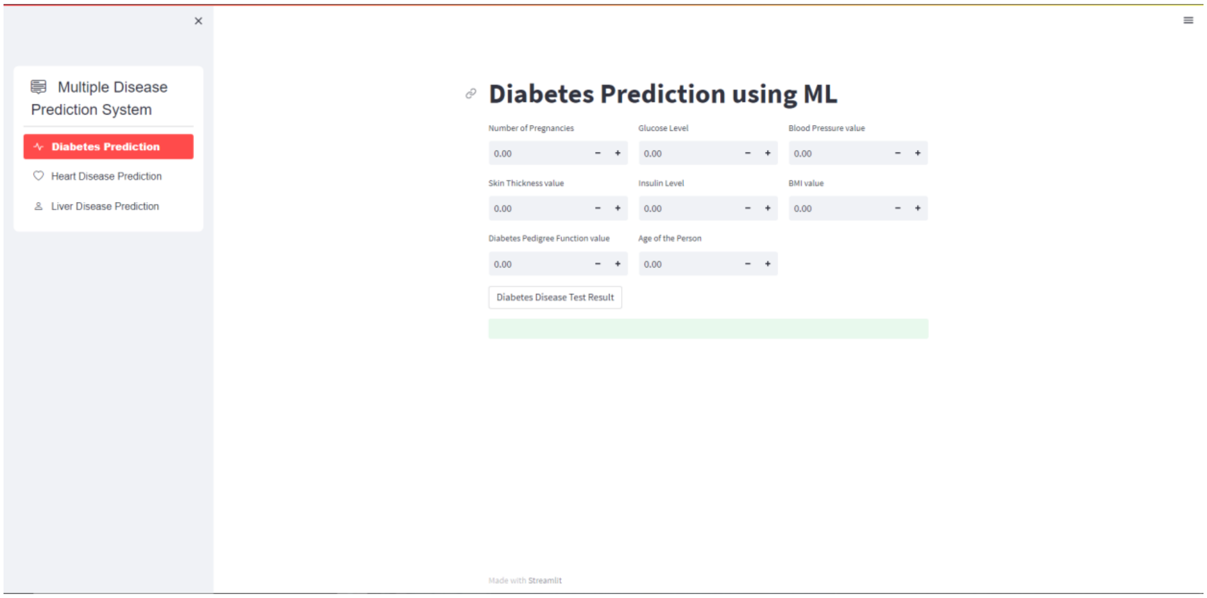
CHAPTER 6

PROJECT DETAILS

- For front-end we used Streamlit.
- For back-end we Googlecollab and various machine learning algorithms.

PROJECT IMAGES:

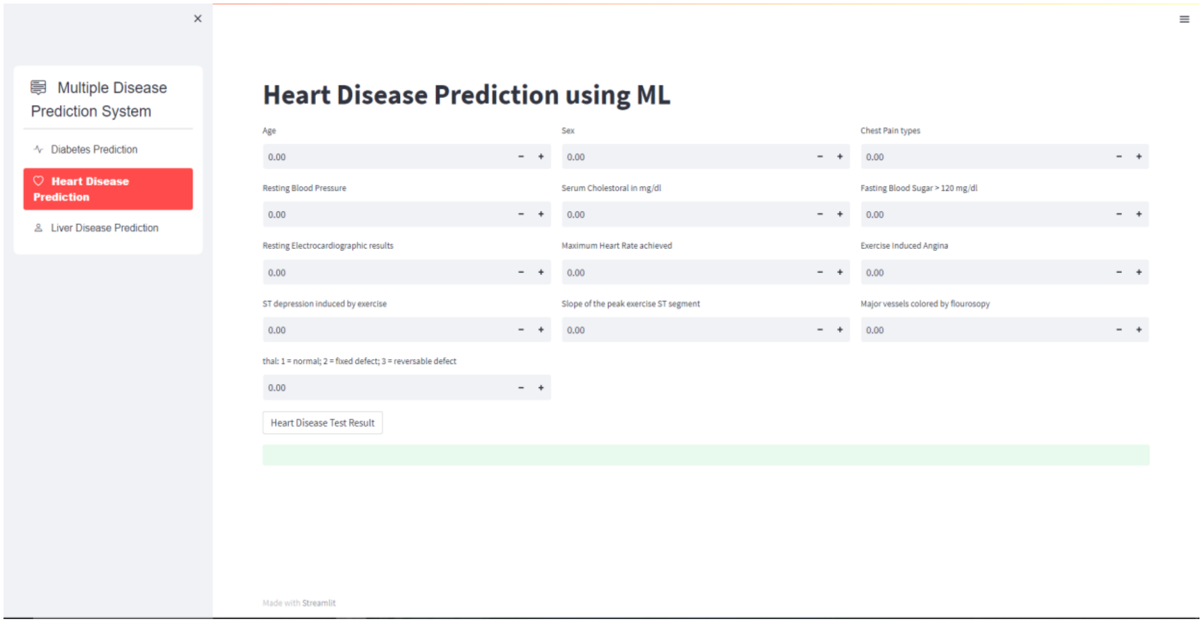
1. HOME PAGE



Diabetes Prediction using ML

Number of Pregnancies: 0.00
Glucose Level: 0.00
Blood Pressure value: 0.00
Skin Thickness value: 0.00
Insulin Level: 0.00
BMI value: 0.00
Diabetes Pedigree Function value: 0.00
Age of the Person: 0.00

2. Heart disease prediction



Heart Disease Prediction using ML

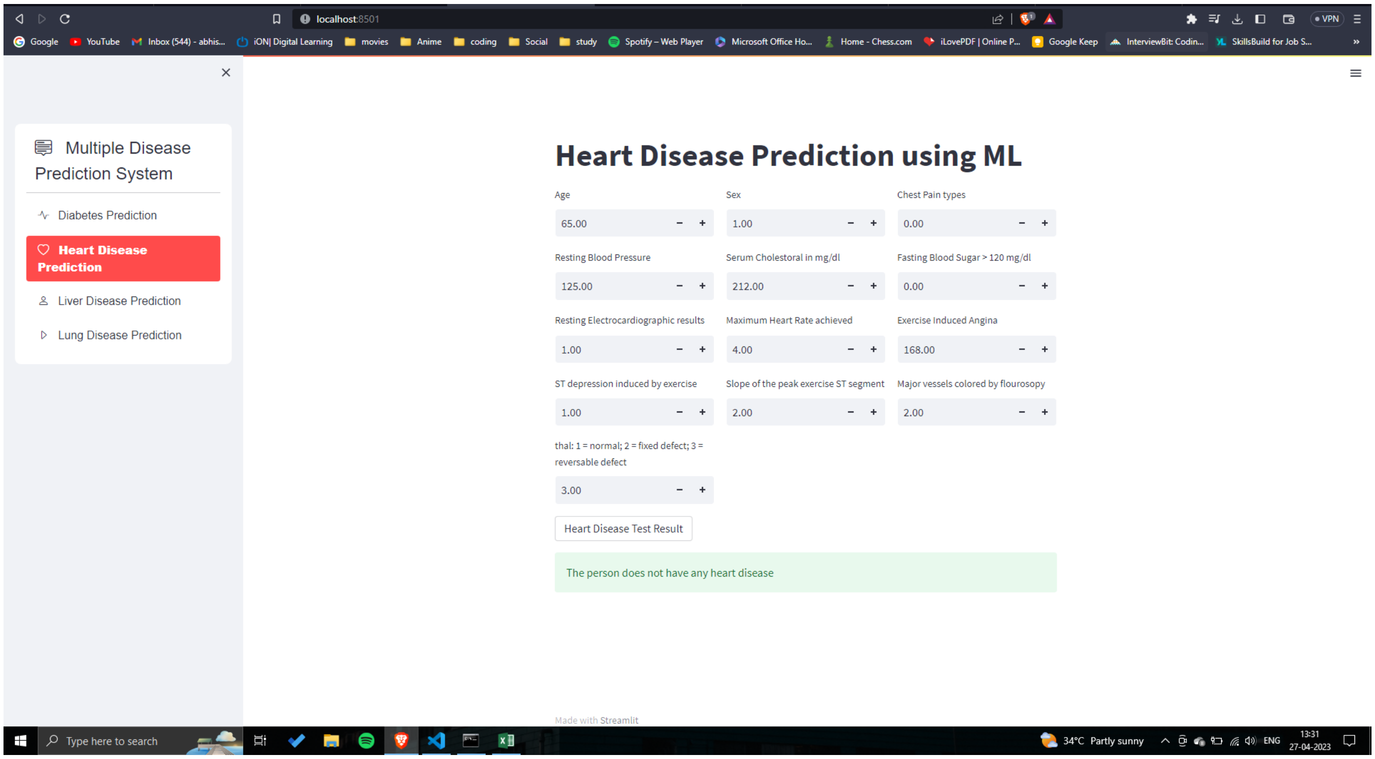
Age: 0.00
Sex: 0.00
Chest Pain types: 0.00
Resting Blood Pressure: 0.00
Serum Cholestoral in mg/dl: 0.00
Fasting Blood Sugar > 120 mg/dl: 0.00
Resting Electrocardiographic results: 0.00
Maximum Heart Rate achieved: 0.00
Exercise Induced Angina: 0.00
ST depression induced by exercise: 0.00
Slope of the peak exercise ST segment: 0.00
Major vessels colored by flourosopy: 0.00
thal: 1 = normal; 2 = fixed defect; 3 = reversable defect: 0.00

CHAPTER 7

OUTCOMES AND RESULTS

7.1 RESULTS

The expected outcome from this project was to successfully create a website which could predict disease.



Heart Disease Prediction using ML

Age: 65.00

Sex: 1.00

Chest Pain types: 0.00

Resting Blood Pressure: 125.00

Serum Cholestoral in mg/dl: 212.00

Fasting Blood Sugar > 120 mg/dl: 0.00

Resting Electrocardiographic results: 1.00

Maximum Heart Rate achieved: 4.00

Exercise Induced Angina: 168.00

ST depression induced by exercise: 1.00

Slope of the peak exercise ST segment: 2.00

Major vessels colored by flourosopy: 2.00

thal: 1 = normal; 2 = fixed defect; 3 = reversable defect: 3.00

Heart Disease Test Result

The person does not have any heart disease

After several changes in code and after running several tests, we can say with utmost surety to the best of our knowledge that these objectives have successfully turned into our project outcomes.

7.2 FUTURE PROSPECTS

The main enhancement which we plan to do is to collect more and more data so that the training of our model can be done in the best possible way. Also, the use several other Machine Learning algorithms could also help in taking our accuracy a level up. We would work on the UI a bit more wherein the user need not input everything only the important features (features selected after feature selection). This will not only make our model light weight but also provide the user a better experience.

Also use of different techniques such as hyperparameters tuning on each model to get the best out of the models, techniques such as boosting and bagging etc., are in our to-do for the future version of the project.

CHAPTER 8

CONCLUSION

The project presented the technique of predicting the disease based on the symptoms, age, and gender of an individual patient. Different Machine learning algorithms were used to carry out the project such as the Random Forest, Decision Tree and Logistic Regression. Almost all the ML models gave good accuracy values. As some models were dependent on the parameters, they couldn’t predict the disease and the accuracy percentage was quite low. Once the disease is predicted, we could easily manage the medicine resources required for the treatment. This model would help in lowering the cost required in dealing with the disease and would also improve the recovery process.

We have also created a GUI for better interaction with the system by users which is very easy to operate .This paper shows that Machine Learning algorithm can be used to predict the disease easily with different parameters and models. To conclude, our system is helpful to those people who are always worrying about their health and they need to know what happens with their body. Our main motto to develop this system is to know them for their health. Especially, people who are suffering from mental illness like depression, anxiety. They can come out of these problems and can live their daily lives easily. Besides, our system provides better accuracy of disease prediction according to symptoms of the user, and also it will provide motivational thoughts and images. In the end, we can say that our system has no boundary of the user because everyone can use this system.

THANK YOU!

CHAPTER 9

REFERENCES

1. A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of heart disease using machine learning," in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275–1278.

2. Y. Hasija, N. Garg, and S. Sourav, "Automated detection of dermatological disorders through image-processing and machine learning," in 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 1047–1051

3. S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1– 16, 2019

4. Shrestha,Ranjit.,& Chatterjee,Jyotir. Moy. (2019).Heart Disease Prediction System Using Machine Learning . LBEF Research Journal of Science Technology and Management , 1(2), 115-132

5. Godse, Rudra A.,Gunjal,Smita S., JagtapKaran A .,Mahamuni ,Neha S., &Wankhade, Prof. Suchita. (2019). Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively.International Journal of Advance Research in Computer and Communication Engineering, 8(12), 50-52

6. Anitha ,Dr.S.,& Sridevi,Dr.N. (2019). Heart Disease Prediction Using Data Mining Techniques.Journal of analysis and Computation ,13(2) , 48-55.

7. Bindhika,Galla Siva Sai., Meghana,Munaga., ReddyManchuriSathvika. , &Rajalakshmi. (2020). Heart Disease Prediction Using Machine Learning Techniques. International Research Journal of Engineering and Technology, 7(4) , 5272-5276.

8. Pingale,Kedar., Surwase, Sushant., Kulkarni,Vaibhav.,Sarage ,Saurabh., &Karve, Prof. Abhijeet .(2019). Disease Prediction using Machine Learning.International Research Journal of Engineering and Technology, 6(12) , 2810-2813.

9. Chauhan Raj H., NaikDaksh N. ,Halpati,Rinal A., Patel,Sagarkumar J. , &PrajapatiMr. A.D. (2020). Disease Prediction using Machine Learning.International Research Journal of Engineering and Technology, 7(5) , 2000-2002.