**An Industry Oriented Major Project Report on**

# A Healthcare System using Machine Learning Systems for Disease Prediction with Chatbot Assistance

*Submitted to the*

## JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD

*In partial fulfilment of the requirement for the award of the degree of*

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**BY**

|  |  |
| --- | --- |
| ANAGONDI SHIVA PRASAD | (19WJ1A0519) |
| BADEMPET PAVAN | (19WJ1A0531) |
| CHEERALA PRASHANTH | (19WJ1A0561) |

Under the Esteemed Guidance of

# Mr. K. CHANDRA SHEKAR

**Associate Professor, Department of CSE**



**Department of Computer Science & Engineering**

**GURU NANAK INSTITUTIONS TECHNICAL CAMPUS (AUTONOMOUS)**

**School of Engineering & Technology**

**Ibrahimpatnam, R.R District- 501506**

**2023**



Date: 11-01-2023

## Department of Computer Science & Engineering

**CERTIFICATE**

This is to certify that the project report titled **“MOVIE RECOMMENDATION SYSTEM BASED ON USER PREFERENCES”** by **ANAGONDI SHIVA PRASAD (19WJ1A0519), BADEMPET PAVAN (19WJ1A0531) & CHEERALA PRASHANTH (19WJ1A0561)** is submitted in partial fulfilment of the requirements for the degree of **Bachelor of Technology** in **Computer Science & Engineering** of the **Jawaharlal Nehru Technological University Hyderabad** during the academic year 2022 - 23, is a Bonafede record of work carried out under our guidance and supervision.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

### INTERNAL GUIDE PROJECT CO-ORDINATOR HOD CSE

**Mr. K. Chandra Shekar Mr. B. Samirana Acharya Dr. J. Rajeshwar**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**EXTERNAL EXAMINER**

**ACKNOWLEDGEMENT**

We wish to express our sincere thanks to **Mr. Sardar Tavinder Singh Kohli, Chairman, GNITC** for providing us with the conducive environment for carrying through our academic schedules and Project with ease.

We wish to express our sincere thanks to **Mr. Sardar Gagandeep Singh Kohli, Vice Chairman, GNITC** for providing us with the conducive environment for carrying through our academic schedules and Project with ease.

We wish to express our sincere thanks to **Dr. Harvinder Singh Saini, Managing Director, GNITC** for providing us with the conducive environment for carrying through our academic schedules and Project with ease.

We wish to express our sincere thanks to **Dr. Koduganti Venkat Rao, Director, GNITC** for providing us with the conducive environment for carrying through our academic schedules and Project.

We wish to express our sincere thanks to **Dr. Rishi Sayal, Associate Director, GNITC** for providing us with the conducive environment for carrying through our academic schedules and Project with ease.

We would like to say our sincere thanks to **Dr. J. Rajeshwar, Professor & Head, Department of CSE, GNITC** for providing seamless support and right suggestions are given in the development of the project.

We would like to say our sincere thanks to **Mr. B. Samirana Acharya, Assistant Professor, Department of CSE, Project Coordinator,** for providing seamless support and right suggestions are given in the development of the project.

We have been truly blessed to have a wonderful internal guide **Mr. K. Chandra Shekar, Associate Professor, Department of CSE,** for guiding us to explore the ramification of our work and we express our sincere gratitude towards him for leading me through the completion of Project.

Finally, we would like to thank our family members for their moral support and encouragement to achieve goals.

**ANAGONDI SHIVA PRASAD (19WJ1A0519)**

**BADEMPET PAVAN (19WJ1A0531)**

**CHEERALA PRASHANTH (19WJ1A0561)**

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE NO** | **NAME OF THE FIGURE** | **PAGE NO.** |
| 4.1 | SYSTEM ARCHITECTURE | 13 |
| 4.2 | LEVEL 0 DATA FLOW DIAGRAM | 14 |
| 4.3 | LEVEL 1 DATA FLOW DIAGRAM | 15 |
| 4.4 | E-R DIAGRAM | 16 |
| 4.5 | USE CASE DIAGRAM | 17 |
| 4.6 | CLASS DIAGRAM | 18 |
| 4.7 | OBJECT DIAGRAM | 19 |
| 4.8 | SEQUENCE DIAGRAM | 20 |
| 4.9 | COLLABORATION DIAGRAM | 21 |
| 4.10 | STATE DIAGRAM | 22 |
| 4.11 | ACTIVITY DIAGRAM | 23 |
| 4.12 | COMPONENT DIAGRAM | 24 |
| 4.13 | DEPLOYMENT DIAGRAM | 25 |

**LIST OF SNAPSHOTS**

|  |  |  |
| --- | --- | --- |
| **SNAPSHOT NO.** | **NAME OF THE SNAPSHOT** | **PAGE NO.** |
| 7.1 | INTERFACE BEFORE RECOMMENDING | 45 |
| 7.2 | MOVIE SELECTION DROPDOWN | 46 |
| 7.3 | UPON RECOMMEND ACTION, RUNNING OF  MODEL | 46 |
| 7.4 | FETCHING OF BACKDROPS | 47 |
| 7.5 | GENERATION OF SIMILAR MOVIES | 47 |

**LIST OF ABBREVATION**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **ABBREVATION** | **EXPANSION** |
| 1**.** | ML | Machine Learning |
| 2. | GUI | Graphical User Interface |
| 3. | NLP | Natural Language Processing |
| 4. | NLTK | Natural Language Toolkit |
| 5. | AI | Artificial Intelligence |
| 6. | DL | Deep Learning |

**A HEALTHCARE SYSTEM USING MACHINE LEARNING TECHNIQUES FOR DISEASE PREDICTION WITH CHATBOT ASSISTANCE**

**ABSTRACT:** This project aims to develop a healthcare system that can predict multiple diseases, including liver disease, diabetes and heart disease using machine learning techniques. These are the most common types of Diseases that plaque the Indians today. This system will use XGBoost (eXtreme Gradient Boosting) for these diseases while maintaining accuracy, speed, resource utilization, stability and reducing complexity. The system will also include a healthcare chatbot that can provide personalized health recommendations to users.

The system will be developed using Python and its many Machine Learning Libraries & Streamlit Framework. Machine learning models will be trained on publicly available datasets to predict diseases accurately. The developed models will be evaluated using performance metric like accuracy of training and testing data. The healthcare chatbot will be developed with NLP and/or OpenAI API Integration Techniques to provide personalized health recommendations to users. The system will be deployed on a cloud-based platform to ensure scalability and availability.

The Healthcare System has the potential to improve the accuracy of disease diagnosis, care, and early intervention, thereby reducing healthcare costs and improving patient outcomes. The healthcare chatbot can provide accessible and personalized health recommendations, leading to a healthier and happier life.

**CHAPTER 1**

**INTRODUCTION**

In today's era, healthcare systems face numerous challenges in accurately diagnosing and predicting diseases, leading to increased healthcare costs and compromised patient outcomes. To address these challenges, this project aims to develop a comprehensive healthcare system that utilizes machine learning techniques to predict multiple diseases, including liver disease, diabetes, and heart disease. These diseases are prevalent in the Indian population, making them a primary focus of this endeavour.

To ensure the system's effectiveness, the project will employ XGBoost (eXtreme Gradient Boosting), a powerful machine learning algorithm known for its accuracy, speed, and resource utilization. By leveraging the capabilities of XGBoost, the system will deliver robust predictions while reducing complexity and maintaining stability.

Additionally, the developed healthcare system will incorporate a healthcare chatbot that offers personalized health recommendations to users. This chatbot will utilize Natural Language Processing (NLP) techniques and/or potentially integrate with OpenAI APIs to provide users with tailored advice and support.

The implementation of the healthcare system will be carried out using Python programming language, which offers a rich ecosystem of machine learning libraries. Furthermore, the Streamlit framework will be utilized to create a user-friendly interface for seamless interaction with the system.

To ensure the accuracy of disease predictions, the machine learning models will be trained on publicly available datasets that encompass a wide range of patient information. Performance evaluation of the models will be conducted using metrics such as accuracy on both training and testing data, ensuring their reliability and effectiveness.

For scalability and availability, the healthcare system will be deployed on a cloud-based platform. This approach ensures that the system can accommodate increasing user demands while maintaining consistent accessibility.

The ultimate goal of this healthcare system is to significantly improve disease diagnosis accuracy, promote early intervention, and enhance patient care. By doing so, it has the potential to reduce healthcare costs and, more importantly, improve patient outcomes. Moreover, the integration of a healthcare chatbot into the system will enable users to access personalized health recommendations conveniently, fostering a healthier and happier life.

With this project, we envision a future where advanced machine learning techniques and intelligent chatbot systems collaborate to revolutionize healthcare, making it more efficient, accurate, and accessible to individuals from all walks of life.

**1.1 AIM OF PROJECT**

To build a Movie Recommender System with Cosine Similarity to determine the most similar movies. In this case, the system converts the information about the movies into vectors based on common features such as genre, cast, and plot. These vectors are then compared to each other using Cosine Similarity, which measures the similarity between two vectors based on the cosine of the angle between them. This could be considered as a microservice in a larger web application, as it is a specific, self-contained component that can be integrated into a larger system to provide a specific functionality.

**1.2 OBJECTIVE**

The objectives of this project are to develop a healthcare system that leverages machine learning techniques to accurately predict prevalent diseases like liver disease, diabetes, and heart disease, with a specific focus on the Indian population. The system aims to utilize the XGBoost algorithm, renowned for its accuracy, speed, and efficiency, to ensure precise disease predictions while maintaining stability and reducing complexity. Additionally, the project aims to incorporate a healthcare chatbot into the system, utilizing techniques such as Natural Language Processing (NLP) and OpenAI API integration to provide personalized health recommendations to users. The implementation will be carried out using the Python programming language and various machine learning libraries, with the Streamlit framework facilitating a user-friendly interface. The machine learning models will be trained on publicly available datasets, enabling robust disease predictions, and their performance will be evaluated using metrics such as accuracy on both training and testing data. Deployment on a cloud-based platform will ensure scalability and availability of the healthcare system. Ultimately, the project seeks to enhance disease diagnosis accuracy, enable early intervention, reduce healthcare costs, and improve patient outcomes. The integration of a healthcare chatbot aims to offer accessible and tailored health recommendations, empowering users to lead healthier and happier lives.

**1.3 SCOPE OF THE PROJECT**

The scope of this project includes the development of a healthcare system that utilizes machine learning techniques to predict common diseases, such as liver disease, diabetes, and heart disease, with a focus on the Indian population. The project also involves integrating a healthcare chatbot to provide personalized health recommendations. The system will be developed using Python, machine learning libraries, and the Streamlit framework. The machine learning models will be trained on publicly available datasets, and the system will be deployed on a cloud-based platform for scalability and availability.

**1.4 EXISTING SYSTEM**

In the case of the existing system, there are very few existing systems in the market that have the potential to evolve into a comprehensive All-in-One Healthcare System for Disease Prediction. However, even those existing systems that do exist face significant limitations. These systems often require a substantial amount of computing resources, yet they struggle to achieve high levels of accuracy in disease prediction.

In terms of disease prediction algorithms, the existing system relies on Decision Trees for liver and heart disease prediction, and K-Nearest Neighbours (KNN) for diabetes prediction. However, each of these algorithms has its own drawbacks and limitations, leading to suboptimal performance in disease prediction.

Moreover, the existing systems lack a well-designed user interface (UI) and struggle with issues related to scalability and ease of use. They also fall short in providing personalized recommendations to users, which limits their effectiveness in addressing individual healthcare needs.

Given these limitations, there is a clear opportunity to develop a more advanced and comprehensive healthcare system that surpasses the existing solutions in terms of accuracy, usability, scalability, and personalized recommendations.

### **1.4.1 EXISTING SYSTEM DISADVANTAGES**

* KNN can be computationally expensive and slow when dealing with large datasets, as it requires calculating the distance between the new input point and every point in the training dataset.
* Decision Trees are prone to overfitting, especially when dealing with noisy data or complex datasets with many features. This can lead to poor generalization and low accuracy on new data.
* Both KNN and Decision Trees can suffer from the curse of dimensionality, where the performance of the model decreases as the number of features increases. This can lead to increased computation time and poor accuracy.

**1.5 LITERATURE SURVEY**

## 1.6 PROPOSED SYSTEM

The proposed system aims to overcome the limitations of existing systems by introducing advanced features and robust algorithms for disease prediction. In the proposed system, XGBoost (eXtreme Gradient Boosting), an Ensemble Boosting Technique, is utilized for accurate prediction of liver disease, heart disease, and diabetes.

By incorporating XGBoost, the proposed system ensures superior performance in disease prediction compared to traditional algorithms. XGBoost is known for its ability to handle complex and noisy medical datasets, resulting in more accurate and interpretable results.

Furthermore, the proposed system includes a healthcare chatbot that provides personalized health recommendations to users. This chatbot leverages the power of the XGBoost models to offer tailored advice and support based on individual healthcare needs.

To enhance usability and convenience, the proposed system incorporates a user-friendly interface (UI) that provides easy access to all the tools and functionalities in one place. This centralized UI allows users to interact with the disease prediction models and healthcare chatbot seamlessly.

The proposed system aims to be powerful, efficient, and comprehensive, delivering accurate disease predictions and personalized health recommendations. By leveraging advanced algorithms and providing a user-friendly interface, it seeks to provide a holistic solution for disease prediction and management, ultimately improving healthcare outcomes for individuals.

### **1.6.1 PROPOSED SYSTEM ADVANTAGES: -**

XGBoost uses optimized algorithms and parallel processing techniques to handle large datasets more efficiently.

• XGBoost addresses overfitting issue by using regularization techniques, such as L1 and L2 regularization, to prevent overfitting and improve the generalization of the model.

• XGBoost uses feature selection and feature engineering techniques to reduce the number of features and improve the performance of the model.

**CHAPTER 2**

**PROJECT DESCRIPTION**

This project involves the development of a healthcare system that utilizes XGBoost (eXtreme Gradient Boosting), an Ensemble Boosting Technique, for accurate disease prediction. The system focuses on predicting liver disease, heart disease, and diabetes. Additionally, a healthcare chatbot is integrated into the system to provide personalized health recommendations. The project aims to overcome the limitations of existing systems by employing robust algorithms, offering a user-friendly interface, and delivering accurate and interpretable results.

**2.1 METHODOLOGIES**

**2.1.1 MODULES**

* Data Collection
* Data Analysis
* Data Preprocessing
* Model Training
* Creating Web App Service
* Containerization & Deploying on Cloud

**1. Data Collection**

The Data Collection module involves the process of gathering relevant data for disease prediction. In this project, the data is obtained from various sources such as publicly available datasets or other reliable sources. Kaggle is a platform that hosts a variety of datasets that can be used for different projects. To ensure that the data is up to date, it is important to constantly revise the dataset and make minimal to no changes to the code. This ensures that the analysis is based on the most recent data available.

Two datasets were collected in this project, one consisting of information about the movies and the other containing credits information. The Movies Information dataset includes details such as the title, release date, budget, and revenue of a movie. The Credits dataset includes information about the cast and crew of a movie, including the actors, directors, and writers. These two datasets are used together to provide a complete picture of the movie industry.

**2. Data Analysis**

The Data Collection module involves the process of gathering relevant data for disease prediction. In this project, the data is obtained from various sources such as publicly available datasets or other reliable sources.

**3. Data Preprocessing**

The Data Preprocessing module plays a crucial role in preparing the collected data for training the machine learning models. It encompasses several essential steps, such as handling missing values, addressing outliers, encoding categorical variables, and scaling or normalizing numerical features. These preprocessing steps are crucial for improving the accuracy and reliability of the models' predictions, as they help to mitigate the impact of noisy or inconsistent data. By carefully preprocessing the data, the system can handle various data-related challenges and ensure that the input provided by users is appropriately transformed for accurate disease prediction.

**4. Model Training**

The Model Training module is a crucial component of the proposed system, responsible for training the machine learning models used for disease prediction. In this project, eXtreme Gradient Boosting (XGBoost) models are employed for predicting liver disease, heart disease, and diabetes. These models are trained using the preprocessed data, which includes relevant features and labels. The training process involves optimizing the models' parameters to achieve high accuracy and generalizability. Performance metrics like accuracy are used to evaluate the models on both training and testing datasets, ensuring their effectiveness in disease prediction. The trained models are then serialized and saved using the pickle library, allowing for efficient storage and future use. Through rigorous training and evaluation, the models are equipped to make accurate predictions and play a vital role in the overall functionality of the healthcare system.

**5. Creating web app service**

The Creating Web App Service module is a crucial component of the project that focuses on developing a web application using the Streamlit framework. Streamlit is utilized to create an interactive and user-friendly interface for the disease prediction system. The web app service provides different pages for each disease prediction (diabetes, heart disease, and liver disease) and integrates a healthcare chatbot. It allows users to input their health parameters, such as glucose level, blood pressure, age, etc., for disease prediction. The web app service utilizes the trained XGBoost models to generate accurate predictions based on the user inputs. The predictions are displayed on the web interface, providing users with immediate results. Additionally, the web app service ensures scalability and accessibility by deploying the system on a cloud-based platform. This enables users to access the disease prediction system from various devices and locations, enhancing the overall user experience and facilitating widespread adoption of the healthcare application.

**6. Containerization & Deploying on Cloud**

The Containerization & Deploying on Cloud module involves packaging the developed system into containers and deploying them on a cloud-based platform. Containerization enables encapsulating the application and its dependencies into portable units, ensuring consistency and ease of deployment across different environments. By using containerization technologies like Docker, the system can be bundled with all necessary components, making it self-contained and easily deployable. Furthermore, deploying the system on a cloud-based platform offers scalability, high availability, and efficient resource utilization. Cloud platforms such as Heroku and Azure provide infrastructure for hosting containers and managing their orchestration. This ensures that the healthcare system can handle increased user demand, allows for easy scaling, and provides reliable access to users from anywhere, anytime. Additionally, cloud-based deployment offers flexibility in terms of integrating with other services like databases, monitoring tools, or load balancers, further enhancing the system's capabilities and overall performance.