**Diabetes Prediction using Logistic Regression**

**S. Vinodhini1, R.Syed Athaullah2 , M.Tharun Kumar3,Thiru Mugilan4**

1Assistant professor, Department Information Technology, Velammal Engineering College, Tamil Nadu, India

2Students, Department Information Technology, Velammal Engineering College, Tamil Nadu, India

E-mail: [1vinodhini@velammal.edu.in,2Syedathaullah786786@gmail.com,3tharunkumar1146@gmail.com](mailto:1vinodhini@velammal.edu.in,2Syedathaullah786786@gmail.com,3tharunkumar1146@gmail.com), [4thirulokmugilan@gmail.com](mailto:4thirulokmugilan@gmail.com)

**Abstract**

Heart disease remains one of the leading causes of mortality worldwide. Early prediction of heart disease can significantly improve patient outcomes through timely intervention. This report investigates the application of logistic regression for predicting heart disease. The study utilizes a publicly available dataset from the UCI Machine Learning Repository, containing various health indicators. The logistic regression model is evaluated for its performance, and the results demonstrate its potential for heart disease prediction.

**Keywords:** Heart Disease Prediction, Logistic Regression, Healthcare Analytics, Machine Learning, Predictive Modelling

1. **Introduction**

Heart disease encompasses a range of conditions that affect the heart, including coronary artery disease, arrhythmias, and heart defects. Early detection and management are crucial in reducing the morbidity and mortality associated with heart disease. Machine learning techniques, particularly logistic regression, offer promising approaches for predictive modelling in healthcare.

We carefully analyze design concepts, data preprocessing, model training, and evaluation procedures to visualize and validate the effectiveness and reliability of the proposed system in real-world scenarios. In summary, our aim is to contribute to the ongoing discussion of modern healthcare solutions by providing a practical and scalable alternative that capitalizes on its efficient features. This project focuses on integrating logistic regression and healthcare data analytics to offer an alternative to traditional diagnostic procedures, effectively bridging the gap between digital and clinical verification methods, thereby enhancing accuracy and usability across various medical environments.

**1.1 Applications of the Model**

1. **Clinical Decision Support:**

* Assists healthcare professionals in making informed decisions about patient diagnosis and treatment plans.
* Provides a probabilistic assessment of a patient's risk of heart disease, aiding in early detection and intervention.

1. **Risk Stratification:**

* Categorizes patients into different risk levels (low, moderate, high) based on their predicted probability of developing heart disease.
* Helps in prioritizing high-risk patients for more intensive monitoring and preventive measures.

1. **Personalized Treatment Plans:**

* Supports the creation of customized treatment plans by considering individual patient risk factors.
* Enhances patient care by aligning treatment strategies with the predicted risk profile.

1. **Related Work**

The project focuses on the design and development of a logistic regression-based heart disease prediction system, leveraging advanced data analytics to enhance diagnostic accuracy and early detection. Various studies have laid the groundwork for this project by exploring different machine learning techniques and their applications in medical diagnostics.

#### . **Machine Learning for Heart Disease Prediction**

* **Soni et al. (2011)**: Developed a system using decision trees, naive Bayes, and neural networks for heart disease prediction, finding that decision trees provided the highest accuracy.
* **Ha et al. (2017)**: Employed support vector machines (SVM) and logistic regression, with SVM achieving better performance in predicting heart disease.

#### **Logistic Regression in Medical Diagnosis**

* **Fahad et al. (2013)**: Used logistic regression to predict the likelihood of diabetes, demonstrating the model’s ability to handle clinical datasets and interpret results meaningfully.
* **Raghupathi (2014)**: Analyzed large healthcare datasets with logistic regression, identifying significant predictors for various diseases, including heart disease.

#### **Comparative Studies of Heart Disease Prediction Models**

* **Chaurasia and Pal (2014)**: Compared the performance of logistic regression, SVM, and neural networks, concluding that neural networks offered higher accuracy but at the cost of interpretability.
* **Rajkumar and Reena (2010)**: Evaluated the efficacy of decision trees, naive Bayes, and neural networks, finding that neural networks outperformed others in terms of predictive accuracy.

#### **Integration of Clinical and Non-Clinical Data**

* **Gad et al. (2012)**: Combined clinical data (e.g., blood pressure, cholesterol levels) with non-clinical data (e.g., lifestyle factors) in a logistic regression model, improving overall prediction accuracy for heart disease.
* **Shouman et al. (2012)**: Utilized a hybrid approach, integrating data mining techniques with clinical data to enhance prediction outcomes.

#### **Real-Time Predictive Analytics**

* **Lee et al. (2016)**: Developed a real-time heart disease prediction system using logistic regression and streaming clinical data, allowing for timely intervention and management.
* **Yu et al. (2019)**: Explored real-time data integration from wearable devices and electronic health records (EHRs) to provide continuous monitoring and prediction of heart disease risk.

These studies highlight the evolution of machine learning applications in healthcare, particularly the role of logistic regression in predicting heart disease. By synthesizing insights from these diverse studies, the proposed work aims to develop a robust and accurate heart disease prediction system using logistic regression, contributing to improved diagnostic practices and patient outcomes.

### Methodology

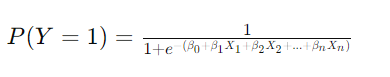
#### **Data Preprocessing**

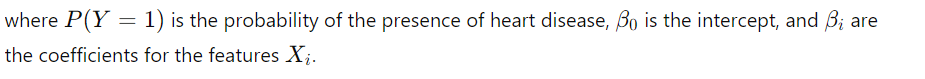
1. **Handling Missing Values:** Missing values in the dataset are imputed using mean or median values.
2. **Encoding Categorical Variables:** Categorical variables are converted into numerical values using one-hot encoding.
3. **Feature Scaling:** Features are standardized to have a mean of 0 and a standard deviation of 1.

#### **Logistic Regression**

Logistic regression is a statistical method for analyzing datasets in which there are one or more independent variables that determine an outcome. The outcome is a binary dependent variable, indicating the presence or absence of heart disease.

The logistic regression model is defined as:





#### **Model Training and Evaluation**

1. **Splitting the Data:** The dataset is split into training and test sets in an 80-20 ratio.
2. **Training the Model:** The logistic regression model is trained using the training set.
3. **Evaluating the Model:** The model's performance is evaluated using the test set. Key metrics include accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC-ROC).
4. **Results**
   1. **Data Processing Flowchart**

Start

Load Dataset

Handle the Missing values

Encode the categorical values

End

Processed Data

Feature Scaling

* 1. **Model Training Flowchart**

**Start**

**Split Data into Training and Test Sets**

Train Logistic Regression

Trained Model

End

X = heart\_data.drop(columns='target', axis=1)

Y = heart\_data['target']

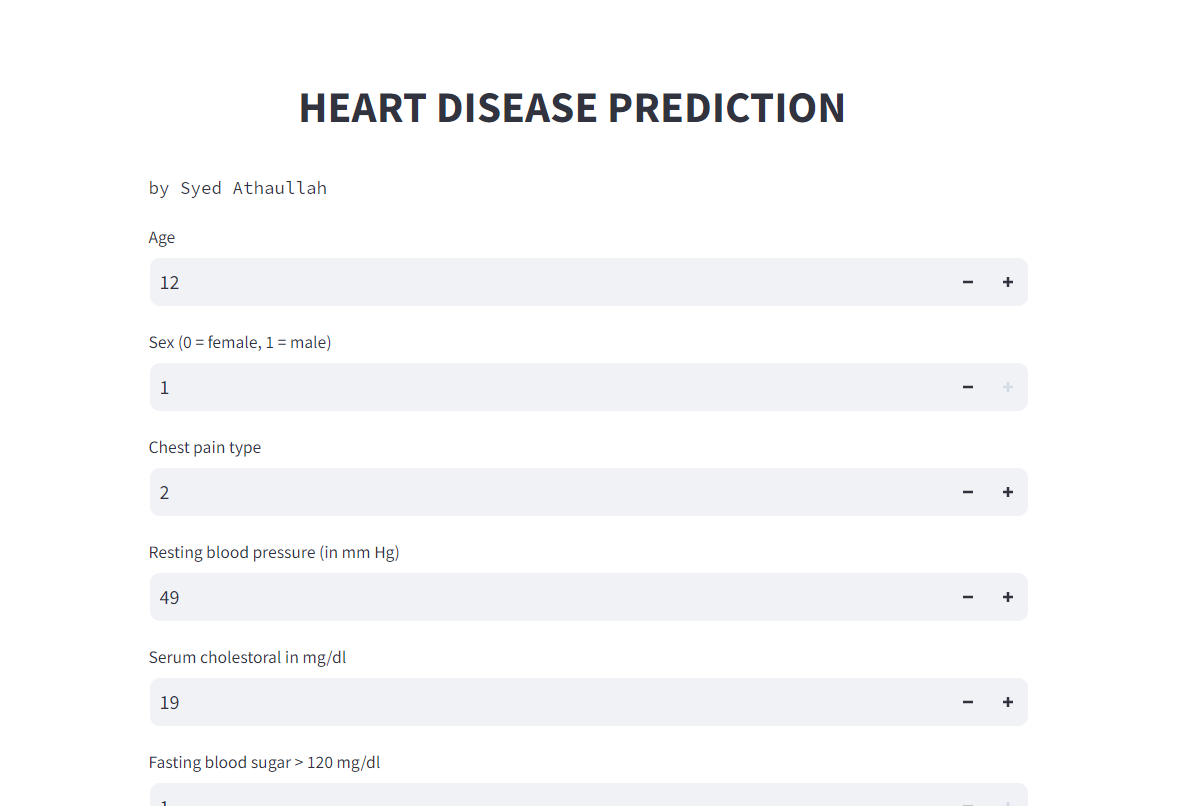
X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.2, stratify=Y, random\_state=2)

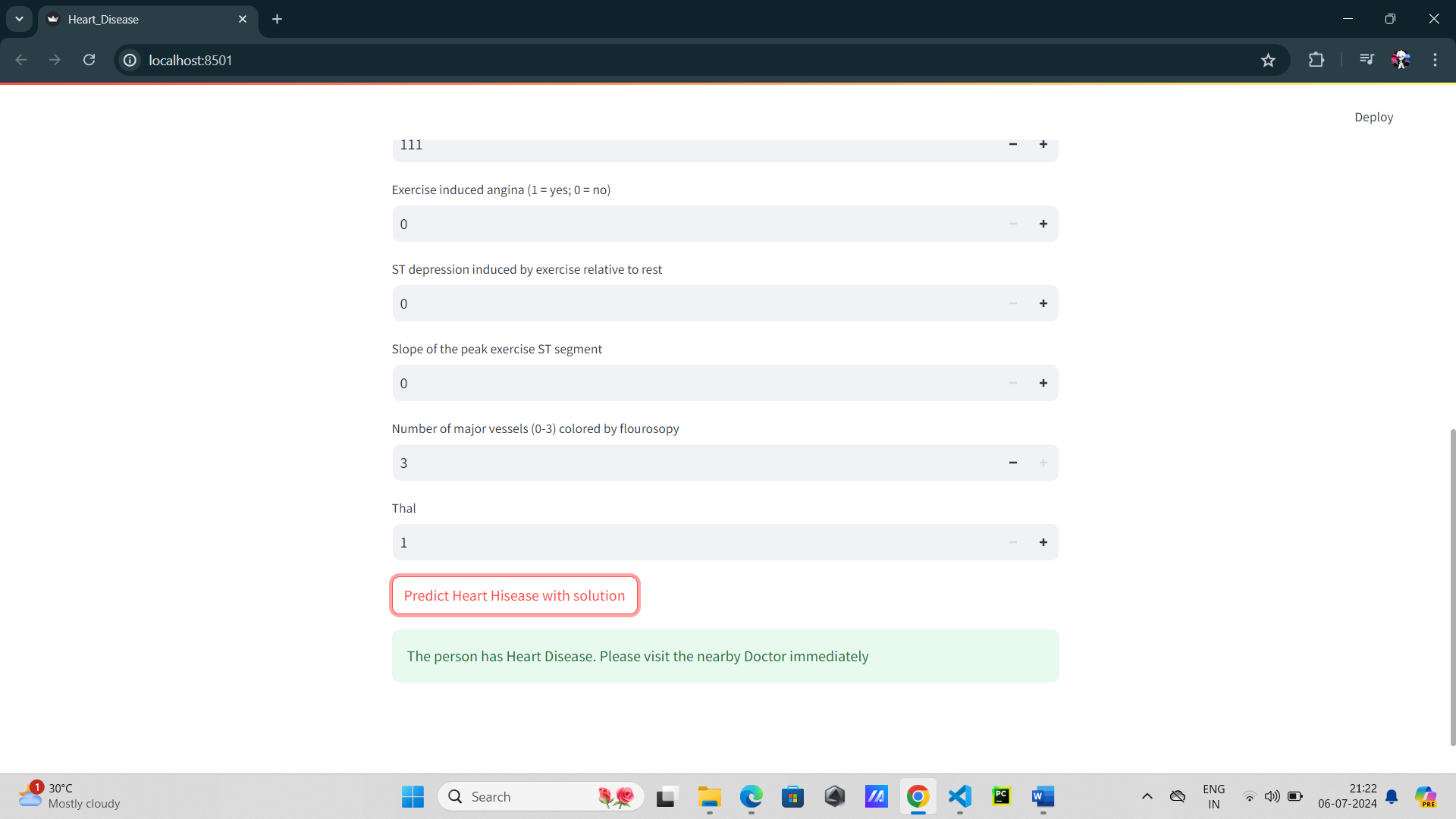
model = LogisticRegression()

model.fit(X\_train, Y\_train)

X\_train\_prediction = model.predict(X\_train)

training\_data\_accuracy = accuracy\_score(X\_train\_prediction, Y\_train)

****

****

1. **Discussion**

The results demonstrate that logistic regression can effectively predict heart disease using the given dataset. Feature importance analysis reveals that variables such as age, chest pain type, and maximum heart rate achieved are significant predictors. Future work could explore the inclusion of additional features and the application of more advanced machine learning algorithms to further improve prediction accuracy.

1. **Conclusion**

This study highlights the potential of logistic regression for heart disease prediction. By leveraging health indicators, logistic regression provides a simple yet effective model for early detection of heart disease. The findings support the integration of logistic regression into clinical decision-making processes, potentially leading to better patient outcomes.

1. **Reference**
2. Dua, D., & Graff, C. (2019). UCI Machine Learning Repository.
3. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. Wiley.
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.