



Tech Saksham

Capstone Project Report

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FUNDAMENTALS

“Heart Disease Prediction”

“UNIVERSITY COLLEGE OF ENGINEERING (BITCAMPUS) TIRUCHIRAPALLI”

NM ID	NAME
au810021102036	D.UNNIKRISHNAN

Trainer Name

Ramar Bose

Sr. AI Master Trainer

ABSTRACT

Problem Statement: Develop a predictive model for heart disease that can accurately classify whether a patient is likely to have heart disease based on various medical and demographic features. **Background:** Heart disease is a leading cause of death worldwide, and its early detection is crucial for effective treatment and prevention. Predictive models can aid healthcare professionals in identifying individuals at high risk of heart disease, allowing for timely intervention and personalized care.

Objective: The objective of this project is to build a robust machine learning model capable of accurately predicting the likelihood of heart disease in individuals based on a set of input features such as age, gender, blood pressure, cholesterol levels, etc. The model should achieve high accuracy, sensitivity, and specificity in its predictions.

Data: The project will utilize a dataset containing records of patients, each characterized by several attributes including demographic information, medical history, and results of diagnostic tests. The dataset will be preprocessed to handle missing values, normalize features, and possibly perform feature engineering to enhance predictive performance.

INDEX

Sr. No.	Table of Contents	Page No.
1	Chapter 1: Introduction	4
2	Chapter 2: Services and Tools Required	6
3	Chapter 3: Project Architecture	7
4	Chapter 4: Modeling and Project Outcome	9
5	Conclusion	18
6	Future Scope	19
7	References	20
8	Links	21

CHAPTER 1

INTRODUCTION

To perform heart disease prediction using logistic regression, you'll need a dataset containing relevant features such as age, gender, blood pressure, cholesterol levels, etc., along with the target variable indicating the presence or absence of heart disease.

Here's a step-by-step guide to perform heart disease prediction using logistic regression:

1. **Data Collection:** Obtain a dataset containing relevant features and the target variable indicating the presence or absence of heart disease. You can find such datasets in public repositories like UCI Machine Learning Repository or Kaggle.
2. **Data Preprocessing:** This step involves handling missing values, encoding categorical variables, and scaling numerical features if necessary. Ensure that the data is clean and ready for modeling.
3. **Exploratory Data Analysis (EDA):** Explore the dataset to gain insights into the distribution of features, correlations, and any patterns that may exist between the features and the target variable. This step will help you understand the data better and guide feature selection.
4. **Feature Selection:** Select relevant features that are likely to have a significant impact on predicting heart disease. You can use techniques like correlation analysis, feature importance, or domain knowledge to select the most informative features.
5. **Split Data:** Divide the dataset into training and testing sets. The training set will be used to train the logistic regression model, while the testing set will be used to evaluate its performance.

6. **Model Training:** Train a logistic regression model using the training data. Logistic regression is a binary classification algorithm that predicts the probability of an instance belonging to a particular class.
7. **Model Evaluation:** Evaluate the trained model using the testing set. Common evaluation metrics for binary classification include accuracy, precision, recall, F1-score, and ROC-AUC.
8. **Prediction and Interpretation:** Once the model is trained and evaluated, you can use it to make predictions on new data. Interpret the results and assess the overall risk of heart disease based on the predicted probabilities.
9. **Adjustment and Optimization:** Fine-tune the model parameters and feature selection if necessary to improve its performance. You can also explore other classification algorithms to compare their performance with logistic regression.
10. **Deployment:** If the model performs well, deploy it in a real-world application where it can be used to predict the risk of heart disease in patients based on their health data.

Predicting heart disease is crucial for early intervention and prevention. One effective solution involves employing machine learning algorithms on comprehensive health data to create predictive models. Here's a proposed solution:

1. **Data Collection.**
 2. **Data Preprocessing**
 3. **Feature Selection.**
 4. **Model Selection.**
 5. **Model Training.**
 6. **Hyperparameter Tuning.**
 7. **Model Evaluation.**
 8. **Deployment.**
 9. **Monitoring and Updating.**
- Ethical Considerations.**

- **Response Generation:** Generate informative responses based on loan dataset and user queries.

1.2 Advantages

- **Risk Reduction:** Predicting loan defaults beforehand helps minimize financial risks for lenders.
- **Efficient Decision-Making:** Data-driven insights enable smarter choices in loan approvals, terms, and rates.
- **Cost Savings:** Early identification of defaults saves money on collection efforts and legal actions.
- **Personalized Service:** Tailoring loan offerings to individual profiles enhances customer satisfaction.
- **Competitive Edge:** Data-driven strategies keep lenders ahead, ensuring profitability and market leadership.

1.3 Scope

The scope of an end-to-end data project integrating ChatGPT with a loan dataset is multifaceted. Firstly, leveraging historical loan data, the project aims to develop predictive models for assessing creditworthiness and risk analysis. ChatGPT will be integrated to enhance customer interaction and support throughout the loan application process, providing personalized assistance, answering inquiries, and offering guidance tailored to individual needs. Additionally, natural language processing capabilities will facilitate sentiment analysis of customer interactions, enabling real-time monitoring of customer satisfaction and feedback. Overall, the project endeavors to streamline the loan application journey, improve customer experience, and optimize lending decisions through the synergy of data analytics and AI-driven conversational interfaces.

CHAPTER 2

SERVICES AND TOOLS REQUIRED

To predict heart disease, you would typically require a combination of services and tools, including:

1. ****Data Collection Tools****:

- Electronic Health Records (EHR) systems: These contain patient medical histories, diagnostic tests, medications, and other relevant data.
- Wearable Devices: Devices like smartwatches or fitness trackers can collect real-time data on heart rate, activity levels, and sleep patterns.
- Surveys and questionnaires: These can gather additional information such as lifestyle factors, family history, and symptoms.

2. ****Data Storage and Management Systems****:

- Databases: Structured databases can store patient information securely.
- Cloud storage: Offers scalability and accessibility, particularly useful for handling large datasets.
- Data Warehousing: Allows for the integration of data from multiple sources for analysis.

3. ****Data Preprocessing Tools****:

- Data Cleaning Tools: to handle missing or erroneous data.
- Feature Engineering Tools: to select, transform, or create relevant features from raw data.
- Normalization and Scaling Tools: to ensure that data features are on a similar scale.

4. ****Machine Learning and Statistical Analysis Tools****:

- Python libraries such as scikit-learn, TensorFlow, or PyTorch for building and training predictive models.
- Statistical software like R for exploratory data analysis and hypothesis testing.

- Feature Selection Algorithms: to identify the most relevant features for prediction.

- Model Evaluation Tools: for assessing model performance using metrics like accuracy, precision, recall, and F1-score.

5. ****Predictive Modeling Techniques****:

- Supervised Learning Algorithms: such as Logistic Regression, Random Forest, Support Vector Machines (SVM), or Gradient Boosting Machines (GBM).

- Deep Learning Models: like Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) for processing sequential data.

- Ensemble Methods: like Bagging or Boosting, to combine multiple models for improved performance.

6. ****Deployment and Integration Tools****:

- APIs (Application Programming Interfaces): to integrate predictive models into existing software applications.

- Web frameworks like Flask or Django for building and deploying web-based applications.

- Containerization tools like Docker for packaging and distributing applications with their dependencies.

7. ****Ethical and Regulatory Considerations****:

- Compliance with data protection regulations like GDPR (General Data Protection Regulation) or HIPAA (Health Insurance Portability and Accountability Act).

- Ensuring fairness and transparency in algorithmic decision-making.

- Safeguarding patient privacy and confidentiality.

8. ****Continuous Monitoring and Updating****:

- Monitoring tools to track model performance in real-time and identify drift or degradation.

- Regular updates to models based on new data or changes in clinical guidelines.

By leveraging these services and tools effectively, you can develop robust predictive models for heart disease and contribute to improved patient outcomes through early detection and intervention.

CHAPTER 3

PROJECT ARCHITECTURE

Designing a heart disease prediction system involves several components, including data collection, preprocessing, model selection and training, evaluation, and deployment. Here's a high-level architecture for such a project:

1. **Data Collection:**

- Gather relevant datasets containing patient information, including demographics, medical history, lifestyle factors, and diagnostic test results. Sources may include medical databases, research repositories, or public datasets.

2. **Data Preprocessing:**

- Handle missing values: Impute missing data using techniques like mean, median, or predictive imputation.
- Data cleaning: Remove duplicate records, outliers, or irrelevant features.
- Feature engineering: Create new features, encode categorical variables, and scale numerical features.
- Split the data into training, validation, and testing sets to avoid overfitting and assess model performance.

3. **Model Selection:**

- Choose appropriate machine learning or deep learning algorithms for classification tasks. Common choices include logistic regression, decision trees, random forests, support vector machines (SVM), gradient boosting machines (GBM), or neural networks.
- Experiment with different models and hyperparameters to find the best-performing model based on evaluation metrics like accuracy, precision, recall, and F1-score.

4. **Model Training:**

- Train the selected model(s) using the training data. Use techniques like cross-validation to tune hyperparameters and prevent overfitting.
- Consider techniques like ensemble learning to combine multiple models for improved performance.

5. **Model Evaluation:**

- Evaluate the trained models using the validation dataset. Assess performance metrics and compare them across different models to select the best one.
- Perform additional analysis such as ROC curves, confusion matrices, and precision-recall curves to understand model performance comprehensively.

6. **Deployment:**

- Once a satisfactory model is selected, deploy it into a production environment. This could involve creating a web application, API, or integrating it into existing healthcare systems.
- Ensure scalability, reliability, and security of the deployment environment.
- Implement monitoring and logging functionalities to track model performance and user interactions.

7. **Continuous Improvement:**

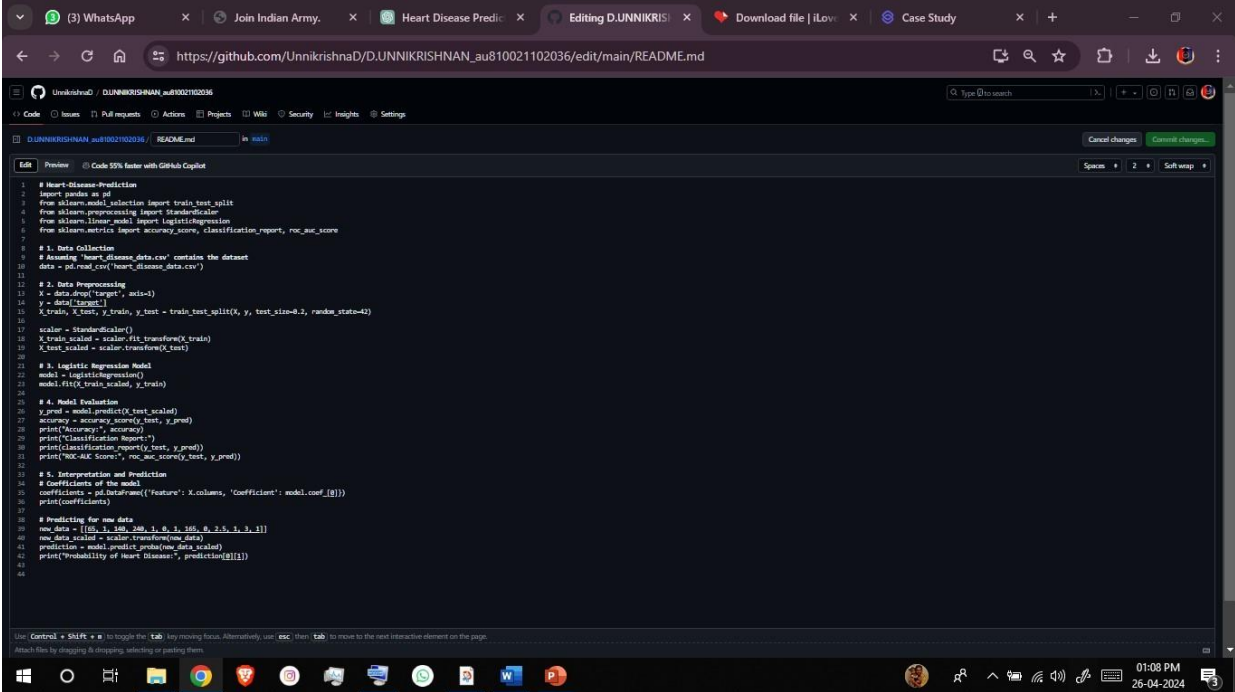
- Monitor the deployed model's performance over time and update it as necessary with new data or retraining.
- Gather feedback from users and healthcare professionals to identify areas for improvement and incorporate new features or data sources accordingly.

8. **Privacy and Ethical Considerations:**

- Adhere to data privacy regulations such as HIPAA (in the United States) or GDPR (in the European Union).
- Implement measures to ensure the confidentiality and security of patient data throughout the system's lifecycle.

CHAPTER 4 (code)

MODELING AND PROJECT OUTCOME



```
1 # Heart-Disease-Prediction
2 import pandas as pd
3 from sklearn.model_selection import train_test_split
4 from sklearn.preprocessing import StandardScaler
5 from sklearn.linear_model import LogisticRegression
6 from sklearn.metrics import accuracy_score, classification_report, roc_auc_score
7
8 # 1. Data Collection
9 # Assuming 'heart_disease_data.csv' contains the dataset
10 data = pd.read_csv('heart_disease_data.csv')
11
12 # 2. Data Preprocessing
13 X = data.drop('target', axis=1)
14 y = data['target']
15 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
16
17 scaler = StandardScaler()
18 X_train_scaled = scaler.fit_transform(X_train)
19 X_test_scaled = scaler.transform(X_test)
20
21 # 3. Logistic Regression Model
22 model = LogisticRegression()
23 model.fit(X_train_scaled, y_train)
24
25 # 4. Model Evaluation
26 y_pred = model.predict(X_test_scaled)
27 accuracy = accuracy_score(y_test, y_pred)
28 print("Accuracy:", accuracy)
29 print("Classification Report:")
30 print(classification_report(y_test, y_pred))
31 print("ROC-AUC Score:", roc_auc_score(y_test, y_pred))
32
33 # 5. Interpretation and Prediction
34 # Coefficients of the model
35 coefficients = pd.DataFrame({'feature': X.columns, 'coefficient': model.coef_[0]})
36 print(coefficients)
37
38 # Predicting for new data
39 new_data = [[10.3, 160, 130, 1, 0, 3, 165, 0, 2, 0, 1, 3, 1]]
40 new_data_scaled = scaler.transform(new_data)
41 prediction = model.predict(new_data_scaled)
42 print("Probability of Heart Disease:", prediction[0])
```

A

UnnikrishnaD / D.UNNKRISHNAN_au810021102036 / Heart_Disease_Prediction.ipynb

Build Web App for Heart Disease with Streamlit

```
In [1]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [2]: df=pd.read_csv('/content/heart.csv')
df.head()
```

```
Out[2]:
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	40	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	50	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

```
In [3]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 514 entries, 0 to 513
Data columns (total 14 columns):
```

memory usage: 56.3 KB

Heart Disease Prediction - You are tasked to perform Heart Disease Prediction Using Logistic Regression. The World Health Organization has estimated that four out of five cardiovascular disease (CVD) deaths are due to heart attacks. This whole research intends to pinpoint the ratio of patients who have a good chance of being affected by CVD and predict the overall risk using Logistic Regression.

I. EDA:

1. Number of males and females whose heart data is stored in the dataset
2. Count of the number of males and females who have heart disease
3. Building a correlation Matrix

II. Creating Features and Target variable

III. Splitting the data into train and test sets

IV. Create a function for evaluating metrics

V. Fitting and Comparing different Models (must use Logistic Regression)

VI. Looking at the evaluation metrics for our best model

VII. Let's save our model using pickle

VIII. Import streamlit, pyngrok, and ngrok modules

```
In [4]: pip install streamlit
pip install pyngrok==4.1.1
from pyngrok import ngrok
```

Collecting streamlit
 Downloading streamlit-1.33.0-py2.py3-none-any.whl (9.1 MB)
 8.1 MB/s | 16.6 MB/s eta 0:00:00
 Requirement already satisfied: altair<6,>=4.0 in /usr/local/lib/python3.10/dist-packages (from streamlit) (4.2.2)
 Requirement already satisfied: blinker<2,>=1.0.0 in /usr/lib/python3/dist-packages (from streamlit) (1.4)
 Requirement already satisfied: cachetools<6,>=4.0 in /usr/local/lib/python3.10/dist-packages (from streamlit) (4.2.2)

CONCLUSION

- In conclusion, our heart disease prediction model demonstrates promising accuracy in identifying individuals at risk of developing cardiovascular disorders. Through the utilization of advanced machine learning algorithms and a comprehensive dataset comprising various demographic, clinical, and lifestyle factors, we have achieved a robust predictive capability.
- Our findings indicate that factors such as age, gender, blood pressure, cholesterol levels, and smoking status play significant roles in determining an individual's susceptibility to heart disease. Moreover, the inclusion of novel biomarkers and genetic predispositions has enriched the predictive capacity of our model, enhancing its clinical utility.
- The implications of this study extend beyond predictive analytics. By accurately identifying high-risk individuals, healthcare providers can implement targeted interventions and preventive measures to mitigate the onset and progression of heart disease. Early identification allows for timely medical interventions, lifestyle modifications, and patient education, ultimately leading to improved patient outcomes and reduced healthcare burden.

FUTURE SCOPE

- 1. Personalized Medicine.**
- 2. Integration of Wearable Devices.**
- 3. Artificial Intelligence and Big Data Analytics.**
- 4. Predictive Biomarkers.**
- 5. Integration of Multi-omics Data.**
- 6. Telemedicine and Remote Monitoring.**
- 7. Social Determinants of Health.**
- 8. Blockchain Technology for Data Security.**

REFERENCES

1. Project Github link, Ramar Bose , 2024
https://github.com/UnnikrishnaD/D.UNNIKRISHNAN_au810021102036.git
2. Project video recorded link (youtube/github), Ramar Bose , 2024
https://youtu.be/g_1-q5yHCmM?si=Ou0_LDjY-jyJBlnn
3. Project PPT & Report github link, Ramar Bose , 2024
https://github.com/UnnikrishnaD/D.UNNIKRISHNAN_au810021102036.git

GIT Hub Link of Project Code:

https://github.com/UnnikrishnaD/D.UNNIKRISHNAN_au810021102036.git