# Parkinson's Disease Detection

Gopal Kumar Shaw

5th Semester

Bangabasi College

Period of Internship:25th August 2025 -19th September 2025

Report submitted to :IDEAS-Institute of Data Engineering ,Analytics and Science Foundation ,ISI Kolkata

## Abstract:

**Purpose**

This project focuses on analyzing heart disease data to build predictive models that could help healthcare professionals make better diagnostic decisions and understand risk factors.

**Key Findings and Approach**

1. **Getting Quality Data**: We used the well-known Heart Disease dataset from UCI's Machine Learning Repository, which contains real patient information that researchers worldwide have been using to study cardiovascular health patterns.
2. **Smart Data Organization**: The dataset comes neatly organized with patient characteristics (like age, cholesterol levels, chest pain types) separated from the actual diagnosis outcomes, making it perfect for building prediction models.
3. **Easy Data Access**: Our approach allows researchers to quickly grab this dataset either by its ID number or name, making the research process more straightforward and accessible to others.
4. **Understanding What We're Working With**: We made sure to explore the dataset's background information, including how many patients are included and what the original researchers discovered, giving us important context for our analysis.
5. **Clear Variable Breakdown**: The code helps us see all the different patient factors in an organized table format, making it easier to understand what information we have to work with for predictions.
6. **Ready for Machine Learning**: Our setup works seamlessly with popular tools like scikit-learn, meaning we can quickly move from data exploration to building actual prediction models without technical hurdles.
7. **Consistent and Reliable**: By using UCI's standardized system, our research can be easily repeated by other scientists, which is crucial for validating findings in healthcare research.
8. **Room to Grow**: This foundation makes it simple to expand our analysis to include other heart-related datasets from UCI, potentially leading to more comprehensive insights.
9. **Proper Research Ethics**: We maintain proper attribution and follow established research standards by using UCI's official repository system, ensuring our work contributes responsibly to the scientific community.

10. **Building Block for Better Healthcare**: This initial setup creates a solid starting point for developing heart disease prediction tools that could eventually help doctors identify at-risk patients earlier.

## Introduction To Project:

**Heart Disease Analysis Project**

### 1. Introduction

Healthcare data analysis has become incredibly important in today's world, where early detection can literally save lives. For this project, we're diving into the world of medical data by exploring heart disease patterns using real patient information from the UCI Machine Learning Repository.

The UCI Repository is like a treasure trove for anyone learning data science - it contains tons of well-documented datasets that researchers and students worldwide use to practice and develop their skills. We specifically chose the Heart Disease dataset because it represents a real-world problem where machine learning can make a genuine difference in people's lives.

Our goal is simple: understand what factors might indicate someone is at risk for heart disease, and learn how to build predictive models that could eventually help doctors make better decisions for their patients.

### 2. What We're Working With

**Programming Tools:**

- **Python** - Our main programming language (it's perfect for data analysis!)
- **Essential Libraries:**
  - ucimlrepo - Helps us easily grab datasets from UCI
  - pandas & numpy - Our go-to tools for organizing and cleaning data
  - scikit-learn - Makes building machine learning models surprisingly straightforward
  - matplotlib & seaborn - Creates beautiful charts and visualizations

**Our Starting Model:** Linear Regression (think of it as drawing the best line through data points to make predictions)

### 3. About Our Data

**The UCI Repository** is basically the gold standard for machine learning datasets. It's free, well-maintained, and used by researchers everywhere, which means our work can be compared with studies from around the world.

**The Heart Disease Dataset** contains real patient information (anonymized, of course) including things like:

- Age and gender
- Cholesterol levels and blood pressure readings
- Heart rate measurements
- Various test results

The target we're trying to predict is simple: does this person have heart disease or not?

**Why Linear Regression?** It's like the "hello world" of machine learning - easy to understand and interpret, making it perfect for getting started and understanding our data.

### 4. How We're Approaching This

Here's our step-by-step game plan:

1. **Explore What's Available** - See what datasets UCI has to offer
2. **Get Our Data** - Download the Heart Disease dataset
3. **Understand What We Have** - Look at the actual patient data and what each column means
4. **Build Our First Model** - Train a simple linear regression to make predictions
5. **Learn About Our Dataset** - Dig into the background information and understand the medical context

**Our Workflow Looks Like This:**

### 5. Why This Project Matters

This isn't just about learning to code or passing a class - we're working on something that connects to real healthcare challenges:

- **Hands-on Learning**: Working with actual medical data gives us experience with the kind of messy, complex information data scientists deal with every day
- **Practical Skills**: We're learning tools and techniques that are used in hospitals, research labs, and healthcare companies right now

- **Making a Difference**: Even though this is a learning project, we're developing skills that could eventually contribute to better patient care
- **Building Confidence**: Starting with well-understood datasets helps us learn without getting overwhelmed

### 6. What We Learned in Our First Two Weeks

Our training covered everything we needed to tackle this project:

**Week 1 - Getting Our Feet Wet:**

1. Python basics for data work (no prior experience needed!)
2. Understanding different types of machine learning problems
3. Getting comfortable with data tables using NumPy and Pandas
4. Creating our first charts and graphs with Matplotlib and Seaborn
5. Discovering the UCI Repository and how to use it

**Week 2 - Diving Deeper:**

6. Learning to fetch datasets using the ucimlrepo library
7. Exploring data like a detective - finding patterns and interesting details
8. Building our first models with scikit-learn (surprisingly fun!)
9. Understanding what all those dataset details and medical terms mean
10. Getting comfortable with linear regression and how it works

The best part about this training was how everything built on itself - each new skill made the next one easier to understand, and by the end, we felt ready to tackle real healthcare data with confidence.

This project represents the perfect blend of learning technical skills while working on something meaningful that could genuinely help people in the future.

## Objectives of the Project:

The main objectives of this project are:

1. **To explore real-world datasets** from the UCI Machine Learning Repository using Python.
2. **To perform data preprocessing and exploratory data analysis (EDA)** for better understanding of dataset characteristics.
3. **To apply machine learning techniques**, such as Linear Regression, for predictive modeling.
4. **To analyze healthcare-related data** (Heart Disease dataset) and identify key factors influencing disease prediction.
5. **To gain practical experience** in using Python libraries like ucimlrepo, pandas, numpy, and scikit-learn.
6. **To understand the importance of predictive analytics** in real-life applications, particularly in healthcare decision-making.
7. **To build a foundation for advanced machine learning projects** by learning end-to-end workflow: dataset fetching → preprocessing → modeling → evaluation.

## Project Methodology and Work Process:

The project follows a structured methodology to ensure systematic execution and meaningful results. The methodology involves **six major stages**, as outlined below:

### 1. Problem Identification

- Understanding the importance of predictive analytics in healthcare.
- Selecting the **Heart Disease dataset** from the UCI Machine Learning Repository to demonstrate prediction of disease likelihood.

### 2. Dataset Collection

- Using the Python library **ucimlrepo** to fetch the dataset from UCI.
- Commands like list_available_datasets() and fetch_ucirepo(id=45) were used to identify and retrieve the Heart Disease dataset.

### 3. Data Preprocessing and Exploration

- Separating **features (X)** and **target (y)** from the dataset.
- Performing **Exploratory Data Analysis (EDA)** to understand patterns, correlations, and variable distributions.
- Checking dataset **metadata** such as number of instances, attributes, and additional information.
- Handling missing values and preparing the dataset for modeling.

### 4. Model Development

- Applying **Linear Regression** as an initial supervised learning model.
- Training the model using sklearn.linear_model.LinearRegression().
- Testing the model's ability to learn relationships between patient health indicators and the target variable (presence/absence of heart disease).

**5. Model Evaluation**

- Evaluating model accuracy and performance.
- Interpreting coefficients to identify which features have more influence on heart disease prediction.

Understanding limitations of Linear Regression for classification-type datasets.

**6. Documentation and Reporting**

- Documenting each step of the process, including dataset details, preprocessing methods, and model results.
- Presenting findings in a structured format for academic and professional reporting.

## Finding and Results:

**1. Dataset Insights**

- The Heart Disease dataset consisted of multiple attributes such as **age, sex, chest pain type, resting blood pressure, cholesterol level, maximum heart rate achieved, fasting blood sugar, and ECG results**, among others.
- The dataset had **sufficient instances and diverse variables**, making it useful for predictive modeling.
- Metadata exploration confirmed the dataset's real-world healthcare relevance and provided details about the number of instances and variable descriptions.

**2. Data Analysis Outcomes**

- Exploratory Data Analysis (EDA) revealed certain patterns:
  - **Age and cholesterol levels** showed a visible influence on the likelihood of heart disease.
  - Features such as **maximum heart rate** and **exercise-induced angina** were significant indicators.
- Correlation analysis highlighted strong relationships between specific medical factors and heart disease occurrence.

### 3. Model Performance

- A **Linear Regression model** was trained on the dataset:
    - The model successfully established a relationship between patient features (X) and the target variable (y).
    - Coefficients of the model indicated which features had greater weight in prediction.
- However, since the target variable in the Heart Disease dataset is **categorical (presence or absence of disease)**, Linear Regression provided limited accuracy compared to classification algorithms (e.g., Logistic Regression, Random Forest).

### 4. Key Findings

- The project **validated the process of fetching, preprocessing, and modeling datasets** directly from the UCI Repository using Python.
- While Linear Regression is mathematically useful for understanding feature influence, **classification models would be more suitable** for predicting disease presence/absence.
- The project achieved its purpose of **familiarizing with machine learning workflows**, from dataset collection to model evaluation.

### 5. Results Summary

- **Successfully implemented dataset import and preprocessing** using ucimlrepo.
- **Developed and trained a predictive model** using scikit-learn.
- **Gained insights into the dataset's features and their influence** on heart disease.
- **Identified limitations** of using regression for categorical prediction tasks, paving the way for future use of more advanced models.

# Conclusion:

This project successfully demonstrated the complete workflow of a machine learning task, starting from **dataset collection using the UCI Machine Learning Repository** to **model development and evaluation** in Python. By applying Linear Regression on the **Heart Disease dataset**, we explored the relationships between patient health indicators and the likelihood of heart disease.

The project highlighted the importance of **data preprocessing, exploratory analysis, and metadata understanding** before applying predictive models. While Linear Regression provided insights into feature influence, it also revealed its limitations for categorical target prediction, suggesting that **classification models (e.g., Logistic Regression, Decision Trees, Random Forests)** would be more effective for such healthcare datasets.

Overall, this project provided valuable **hands-on experience with real-world data and machine learning libraries**, laying a strong foundation for more advanced predictive analytics tasks. The key learning outcome was not only the ability to train models but also the **critical understanding of when and how to apply appropriate algorithms** based on dataset characteristics.

# Appendices:

## Appendix A: References

1.Andreas C. Müller & Sarah Guido – *Introduction to Machine Learning with Python* – O'Reilly.

2.Trevor Hastie, Robert Tibshirani & Jerome Friedman – *The Elements of Statistical Learning* – Springer.

3. Jake VanderPlas – *Python Data Science Handbook* – O'Reilly.

## Appendix B: Github Link for Codes Developed:

The Python codes for data preprocessing, visualization, and machine learning model implementation are uploaded on GitHub:
🔗 https://github.com/gopal-i/IRIS-FLOWER-CLASSIFICATION-USING-LOGISTIC-REGRESSION-AND-RANDOM-FOREST-PARKINSON

## Appendix C: Other Document Links:

- Project Report (this document): https://github.com/gopal-i/IRIS-FLOWER-CLASSIFICATION-USING-LOGISTIC-REGRESSION-AND-RANDOM-FOREST-PARKINSON