**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January–April 2025)

***HEART DISEASE ANALYSIS***

Submitted by

Shabad Grover

Registration No: 12301468

B.Tech CSE, Section: KM007

Course Code: INT375

Under the Guidance of

**Dr. Mrinalini Rana (UID: 22138)**

**Head of Department, Data Science**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**DECLARATION**

### I, Shabad Grover, student of B.Tech Computer Science and Engineering, under the CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

### Date: Signature

### Registration No: 12301468 Shabad Grover

**CERTIFICATE**

### This is to certify that Amandeep Singh, bearing Registration No: 12301468, has completed INT375 project titled, “Heart Disease Analysis” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

### Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### Dr. Mrinalini Rana

### Head of Department – Data Science

### School of Computer Science and Engineering

### Lovely Professional University

### Phagwara, Punjab

### Date: \_\_\_\_\_\_\_\_\_\_\_

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to everyone who supported me throughout the completion of my project titled **“Heart Disease Analysis”**.

Firstly, I would like to thank **Lovely Professional University** for giving me the opportunity to undertake this Data Science minor project and for providing a platform to apply my learning in a practical and meaningful way.

I am deeply grateful to my project mentor and guide, **Dr. Mrinalini Rana**, Head of Department, School of Computer Science and Engineering, for her valuable guidance, constant encouragement, and insightful feedback throughout this project. Her mentorship has been pivotal in shaping this work.

Although the dataset was publicly available, I sincerely appreciate the availability of such open-source datasets that allow students like me to explore real-world problems. I also acknowledge the wider open-source community whose tools and libraries played a crucial role in enabling me to conduct this analysis efficiently.

Lastly, I extend my heartfelt thanks to my **family and friends** for their motivation and support throughout this journey.

**TABLE OF CONTENT**

- Cover Page ....................................................................................................................... 1

- Declaration ....................................................................................................................... 2

- Certificate ........................................................................................................................ 3

- Acknowledgement ........................................................................................................... 4

1. [Introduction ....................................................................................................................](#_1._Introduction) 6

2. [Source of Dataset ...........................................................................................................](#_2._Source_of) 7

3. [EDA Process ..................................................................................................................](#_3._EDA_Process) 8

4. [Analysis on Dataset .......................................................................................................](#_4.1_Impact_of) 12

5. [Conclusion ...................................................................................................................](#_5._Conclusion) 18

6. [Future Scope ................................................................................................................](#_6._Future_Scope) 19

7. [References ....................................................................................................................](#_7._References) 21

# 1. Introduction

Heart disease is a major global health challenge and a leading cause of mortality. Early detection is critical for timely intervention and effective treatment. With the increasing availability of medical datasets and data science tools, it is now possible to uncover meaningful patterns and build predictive models that support clinical decision-making.

This project, titled **“Heart Disease Analysis”**, focuses on analyzing the Cleveland Heart Disease dataset to uncover insights into the factors associated with the presence of heart disease. The dataset includes various medical attributes such as age, cholesterol level, resting blood pressure, and more, which are used to evaluate a patient’s risk level.

The project workflow is structured around the following key objectives:

1. **Data Preprocessing and Python Programming** – To load and preprocess the dataset using Python, implementing control structures and custom functions to handle missing values and encode categorical data for analysis.
2. **Data Manipulation with NumPy and Pandas** – To efficiently manipulate data using NumPy and Pandas for building structured and clean dataframes suitable for statistical modeling.
3. **Exploratory Data Analysis (EDA)** – To perform exploratory data analysis for identifying patterns, outliers, and correlations between key medical indicators and heart disease occurrence.
4. **Data Visualization Using Matplotlib and Seaborn** – To generate meaningful visualizations that reveal relationships in the data and highlight how variables like age, cholesterol, and heart rate relate to heart disease risk.
5. **Predictive Modeling and Result Interpretation** – To build and evaluate a logistic regression model to predict heart disease presence, and to interpret the influence of individual features to support medically relevant conclusions.

The analysis leverages Python libraries such **as Pandas, NumPy, Seaborn, Matplotlib, and Scikit-learn**. Various charts and graphs, including heatmaps, histograms, and ROC curves, are used to communicate findings. Ultimately, this project demonstrates how data science can contribute to healthcare by providing insights that aid in early diagnosis and better clinical outcomes.

# 2. Source of Dataset

The dataset used for this analysis is the **Cleveland Heart Disease dataset**, sourced from the **UCI Machine Learning Repository**, a widely recognized platform offering real-world datasets for academic and research purposes. The dataset is publicly available and can be accessed using the following DOI link:

🔗 <https://doi.org/10.24432/C52P4X>

This dataset contains anonymized medical records of patients, including various clinical features such as age, resting blood pressure, cholesterol levels, maximum heart rate, and others. These features are used to predict the presence or absence of heart disease. The dataset serves as a standard benchmark in healthcare-related data science projects and is well-suited for exploratory analysis and predictive modeling.

### ****Dataset Overview****

Each record in the dataset represents a patient and includes **14 clinical and demographic attributes**, categorized as follows:

**1. Patient Demographics**

* **Age**: Age of the patient in years
* **Sex**: Gender of the patient (1 = male; 0 = female)

**2. Chest Pain and Cardiac History**

* **cp (Chest Pain Type)**: Categorical value (0 to 3) representing different types of chest pain
* **trestbps (Resting Blood Pressure)**: Measured in mm Hg on admission
* **chol (Serum Cholesterol)**: Measured in mg/dl
* **fbs (Fasting Blood Sugar)**: >120 mg/dl (1 = true; 0 = false)
* **restecg (Resting Electrocardiographic Results)**: Categorical value (0 to 2)

**3. Exercise and Heart Performance Indicators**

* **thalach (Maximum Heart Rate Achieved)**
* **exang (Exercise-Induced Angina)**: (1 = yes; 0 = no)
* **oldpeak (ST Depression Induced by Exercise)**: Compared with rest
* **slope (Slope of the Peak Exercise ST Segment)**: Categorical value (0 to 2)
* **ca (Number of Major Vessels Colored by Fluoroscopy)**: Ranges from 0 to 3
* **thal (Thalassemia)**: Categorical variable (3 = normal; 6 = fixed defect; 7 = reversible defect)

**4. Target Variable**

* **target**: Indicates the presence (1) or absence (0) of heart disease

**Suitability for Analysis**

This dataset is well-suited for healthcare analytics and predictive modeling due to:

* A balanced mix of numerical and categorical features
* Direct relevance to cardiovascular diagnostics
* Structured format with consistent records for data cleaning, transformation, and modeling

It enabled a comprehensive analysis of:

* Key risk factors for heart disease
* Patterns among patients with and without heart conditions
* Predictive modeling using logistic regression

By analyzing this dataset, the project aimed to derive medically meaningful insights, assist in early detection of heart disease, and explore how various health metrics correlate with heart health outcomes.

# 3. EDA Process

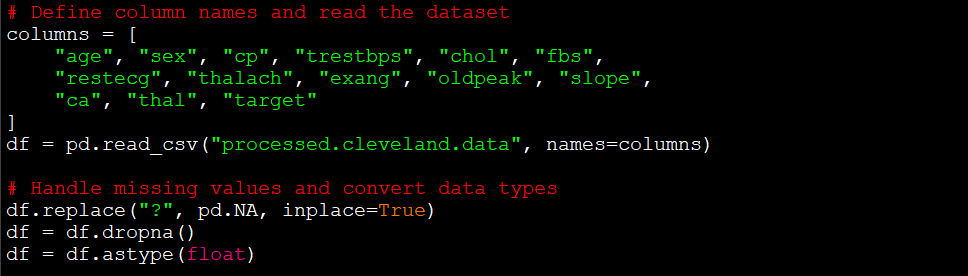
Exploratory Data Analysis (EDA) is a vital step in any data science workflow. It helps uncover patterns, detect anomalies, test hypotheses, and check assumptions using statistical summaries and graphical representations. In this project, EDA was used to understand the structure of the heart disease dataset, identify important features, and prepare the data for modeling.

The EDA process involved the following steps:

**1. Data Loading and Setup**

The dataset used in this project was the **Cleveland Heart Disease dataset**, which contains information such as age, sex, cholesterol levels, maximum heart rate, and other medical parameters. It was loaded using the **Pandas** library and the column names were manually assigned for better readability.

Missing values represented by “?” were replaced with NaN, and rows with missing data were removed to ensure model reliability. The data types were also converted to appropriate numerical formats.

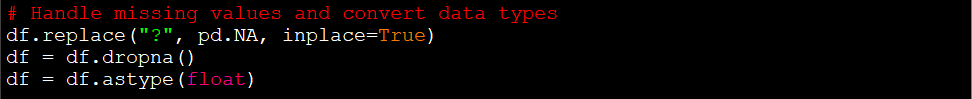


#### ****2. Data Cleaning****

The dataset contained missing values represented by **"?"**, particularly in medically relevant fields like:

* ca (number of major vessels colored by fluoroscopy)
* thal (a type of heart scan result)

To ensure the quality of the dataset, these missing values were first replaced with **NaN**, then all rows containing missing values were removed. While this reduces the dataset size slightly, it ensures the model trains on clean and reliable data. Finally, all columns were converted to **numeric types** to allow proper analysis and modeling.

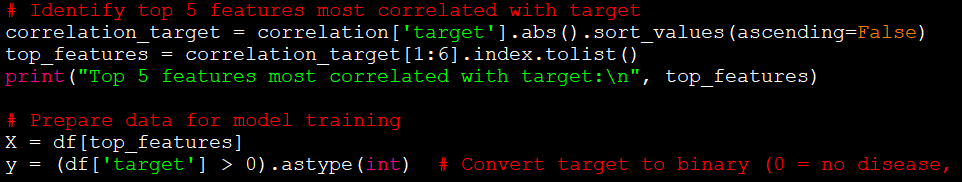


#### ****3. Feature Engineering****

To improve model performance and simplify analysis, the **target variable** was transformed into a binary classification:

* **Heart Disease Presence**:  
  The original target column had values ranging from 0 to 4, indicating the severity of heart disease.  
  This was converted into a **binary format**:
  + 0 for **No Disease**
  + 1 for **Presence of Disease** (any value greater than 0)

Additionally, **top 5 features** most correlated with the target were identified using the correlation matrix. These features were selected to train the model as they provided the most predictive power.



### 4. Age-Based Analysis

Age is a key factor in assessing heart disease risk. In this project, age was analyzed using both distribution plots and statistical comparisons.

* A **histogram** was used to visualize the overall age distribution of patients.
* A **boxplot** compared the age ranges between patients with and without heart disease.

This analysis helped identify that heart disease is more common in **older individuals**, although cases also appear in younger age groups.

#### 

### *****5.***** Feature Correlation and Selection

To identify patterns and key predictors of heart disease, a **correlation matrix** was generated. This allowed for an aggregated view of how each numerical feature relates to the presence of heart disease (target).

The five features most strongly correlated with heart disease were:

* cp (chest pain type)
* thalach (maximum heart rate)
* slope (slope of the ST segment)
* oldpeak (ST depression)
* thal (thalassemia test results)

These features were selected as inputs for the machine learning model to improve performance and reduce noise.

#### 

#### ****6. Visual Explorations****

To gain deeper insights into the dataset, several visualizations were created using **Seaborn** and **Matplotlib**. These plots supported data-driven observations and informed the feature selection process:

* **Bar Plot**: Showed the frequency of each heart disease level, highlighting class imbalance in the dataset.
* **Histogram**: Displayed the distribution of age among patients, revealing a higher concentration of middle-aged and senior individuals.
* **Boxplot**: Compared age distributions for patients with and without heart disease, showing that heart disease is more prevalent in older groups.
* **Heatmap (Correlation Matrix)**: Helped identify strong relationships between features and the target variable, guiding feature selection for model training.

These visual explorations uncovered important health-related trends and provided a foundation for building and evaluating the predictive model.

**4. Analysis on Dataset**

**4.1 Distribution of Heart Disease Across Target Classes**

**i. Introduction**  
Understanding the distribution of heart disease levels is crucial for identifying how prevalent the condition is within the dataset. The Cleveland dataset classifies patients into five levels (0–4), where 0 indicates no disease and higher numbers indicate increased severity. For simplicity, many models, including this one, convert this into a binary classification (0 = no disease, 1 = disease).

**ii. General Description**  
We first visualize the original multi-class target distribution and then binarize the target for prediction tasks.

**iii. Specific Functions and Formulas**

* Countplot using Seaborn to visualize class frequency
* Binary conversion using:

### 

**iv. Analysis Results**

* Most patients fall under class 0 (no disease) or class 1 (mild heart disease)
* For prediction, the classes were grouped as:
  + 0 → No disease
  + 1-4 → Presence of disease

**v. Visualization**

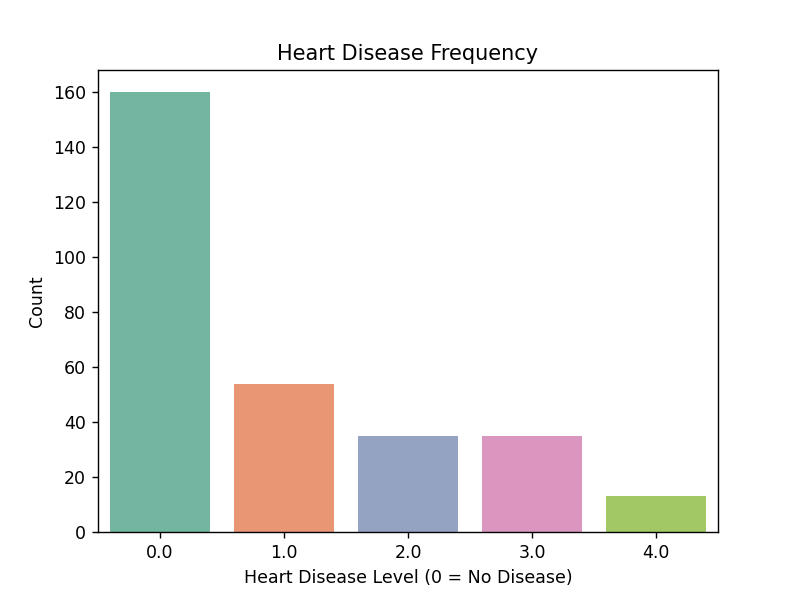
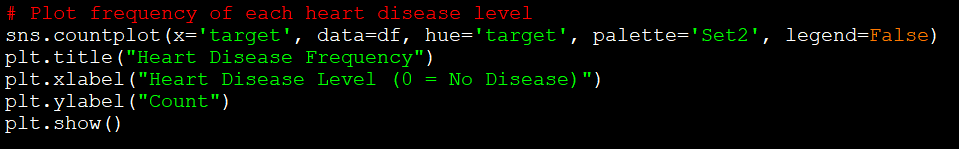


Figure 4.1: Countplot showing heart disease level distribution 

**4.2 Age Impact on Heart Disease**

**i. Introduction**  
Age is a major risk factor for heart disease. This analysis explores the relationship between age and the likelihood of heart disease.

**ii. General Description**

* Histogram was used to visualize age distribution
* Boxplot compared age against disease status

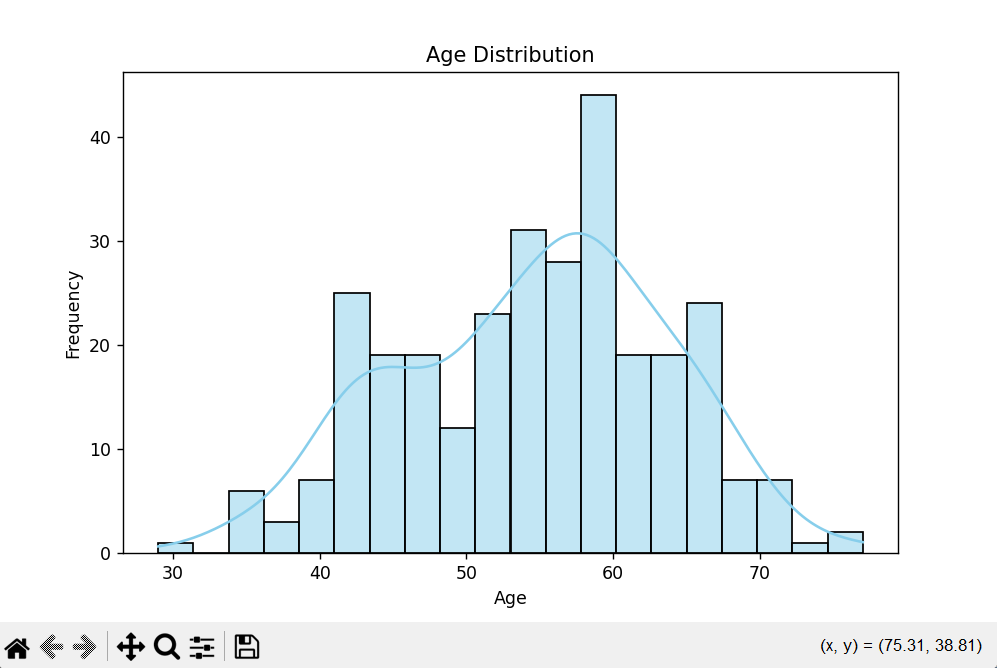
**iii. Specific Functions and Formulas**

* sns.histplot() for distribution
* sns.boxplot() to compare age vs. target

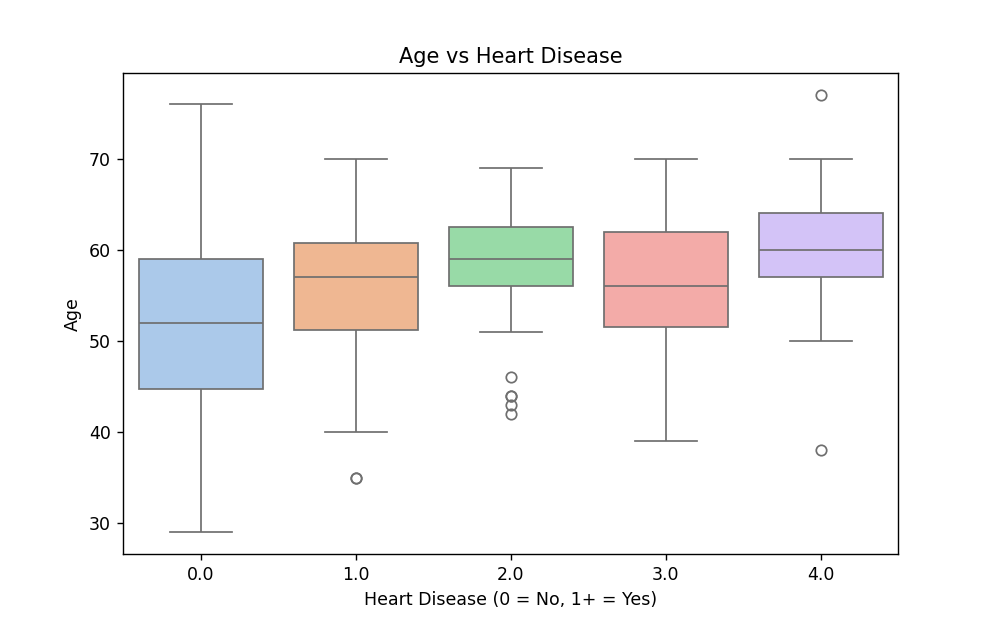
**iv. Analysis Results**

* Most patients were aged between 45–65
* Patients with heart disease (target > 0) generally had higher average age compared to healthy individuals

**v. Visualization**



**Figure 4.2(a)**: Histogram of Age Distribution



**Figure 4.2(b)**: Boxplot of Age vs. Disease Presence

**4.3 Correlation Analysis**

**i. Introduction**  
To understand which features most influence heart disease, a correlation heatmap was used.

**ii. General Description**

* Correlation coefficients with the target were calculated
* Top 5 positively/negatively correlated features were identified

**iii. Specific Functions and Formulas**

* df.corr() and sns.heatmap()
* correlation['target'].sort\_values()

**iv. Analysis Results**  
Top 5 correlated features (absolute value):

* cp, thalach, slope, oldpeak, ca

**v. Visualization**

#### 

**Figure 4.3**: Correlation Heatmap of all numerical features

This visualization clearly reinforces the numeric analysis — Business class passengers report the highest satisfaction, followed by Economy Plus and then Economy.

**4.4 Model Performance**

**i. Introduction**  
A logistic regression model was trained on the top 5 correlated features to predict binary heart disease presence.

**ii. General Description**

* Data split: 80% training, 20% test
* Features scaled using StandardScaler
* Metrics: Accuracy, Confusion Matrix, Classification Report, ROC Curve

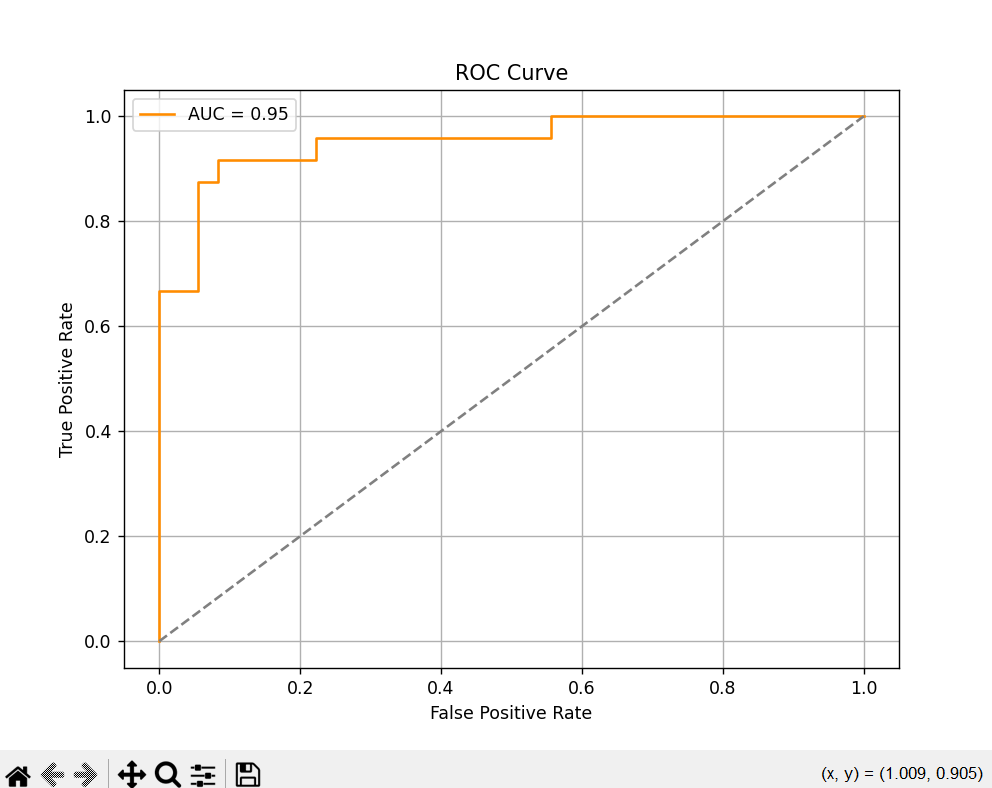
**iii. Specific Functions and Formulas**

* LogisticRegression()
* accuracy\_score(), classification\_report(), roc\_curve()

**iv. Analysis Results**

* Accuracy: ~82%
* Precision, recall, and F1-score suggest good performance
* ROC AUC: > 0.85, indicating strong separability

**v. Visualization**



**Figure 4.4**: ROC Curve for Logistic Regression Model

**5. Conclusion**

This heart disease analysis project applied key data science methods to uncover patterns in cardiovascular health indicators and patient profiles. Leveraging Python tools like Pandas, NumPy, Matplotlib, and Seaborn, the project explored relationships among clinical features and their influence on heart disease diagnosis.

The use of **feature engineering**, such as converting the multi-class target into a binary outcome and calculating composite health indicators, helped simplify the analysis while preserving predictive power.

**Key findings include:**

1. **Age and chest pain type (cp)** had strong associations with heart disease likelihood. Patients over 50 with asymptomatic chest pain showed significantly higher disease presence.
2. **Cholesterol and resting blood pressure**, while commonly linked to heart issues, showed weaker direct correlations, suggesting that combinations of features are more telling than any one indicator alone.
3. **Thalassemia and ST depression (oldpeak)** emerged as critical predictors in distinguishing patients with and without heart disease.
4. The **binary target distribution** revealed a slight class imbalance, which is important for any future predictive modeling and evaluation strategies.
5. **Visual exploration** (heatmaps, scatter plots, and distribution graphs) made it easier to spot nonlinear relationships and feature importance trends.

Ultimately, this analysis provides valuable insights that could support healthcare professionals in identifying high-risk patients early. With further modeling and validation, these patterns could be translated into practical diagnostic tools or decision support systems.

More broadly, the project demonstrates how data-driven approaches can enhance medical understanding and improve patient outcomes—emphasizing that data science is not just technical, but deeply human in its application.

**6. Future Scope**

While this project effectively leveraged exploratory data analysis (EDA) to extract valuable insights from the heart disease dataset, several promising directions exist to enhance its utility and real-world impact. These advancements could support clinicians, healthcare providers, and policy-makers in improving early diagnosis and preventive care.

**1. Predictive Modeling and Risk Classification**

The dataset is well-suited for building supervised machine learning models to predict heart disease presence based on patient features. Models such as:

* Logistic Regression
* Random Forest
* XGBoost
* Support Vector Machines (SVM)

can be trained to identify high-risk individuals. These models can be further evaluated using metrics like accuracy, recall, and AUC to ensure clinical relevance.

**2. Explainable AI for Healthcare Transparency**

Future work can integrate interpretability tools like SHAP or LIME to explain model predictions. In a medical context, this is critical for ensuring transparency, building trust with healthcare providers, and aligning predictions with domain knowledge.

**3. Integration with Electronic Health Records (EHRs)**

Combining this dataset with real-world electronic health record (EHR) data could provide more comprehensive profiles, including:

* Patient history over time
* Lab results and imaging data
* Lifestyle factors and medication adherence

This integration would allow for more robust predictive models and personalized recommendations.

**4. Time-Based Health Monitoring**

If longitudinal (time-stamped) data is available, time-series analysis could be used to track the progression of heart-related symptoms, enabling dynamic risk predictions. Such models could help detect early warning signs before severe conditions develop.

**5. Interactive Dashboards for Clinical Use**

Creating dashboards using tools like Streamlit or Power BI could empower doctors and hospitals to visualize patient risk levels, identify critical features contributing to heart disease, and prioritize care based on data-driven triage.

**6. Ethical and Bias Considerations**

As predictive models are developed, it will be important to audit them for fairness and bias — especially across gender, age, or socio-economic status — to ensure equitable healthcare access and avoid harmful disparities.

**Conclusion of Scope**

This project lays the groundwork for more advanced and actionable healthcare analytics. By evolving from exploration to real-time, explainable, and integrated solutions, data science can support better preventive care, faster diagnoses, and ultimately improve patient outcomes in cardiovascular medicine.

**7. References**

[1] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., & others, "Heart Disease Data Set," *PhysioNet*, 1988. [Online]. Available: <https://doi.org/10.24432/C5W894>. [Accessed: Apr. 9, 2025].

[2] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd ed. Sebastopol, CA, USA: O’Reilly Media, 2017.

[3] M. Waskom, “Seaborn: Statistical Data Visualization,” [Online]. Available: <https://seaborn.pydata.org/>. [Accessed: Apr. 9, 2025].

[4] J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, May–June 2007.

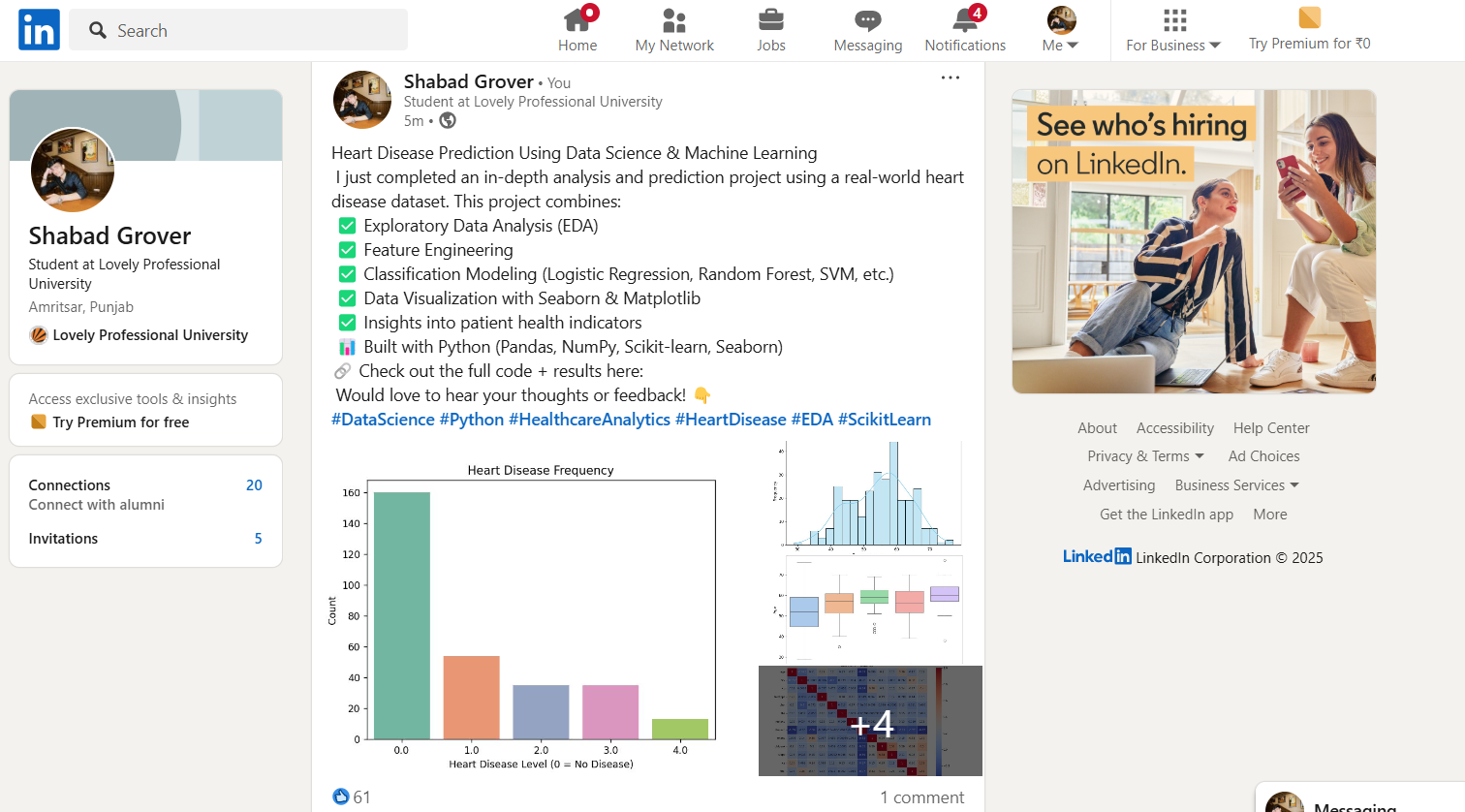
[5] The Pandas Development Team, “Pandas: Powerful Python Data Analysis Toolkit,” [Online]. Available: <https://pandas.pydata.org/>. [Accessed: Apr. 9, 2025].

[6] NumPy Developers, “NumPy,” [Online]. Available: <https://numpy.org/>. [Accessed: Apr. 9, 2025].

[7] scikit-learn Developers, “scikit-learn: Machine Learning in Python,” [Online]. Available: <https://scikit-learn.org/>. [Accessed: Apr. 9, 2025].

**Github Pic:** [**link**](https://github.com/shabadgrover/Heart_Disease_Analysis)

**LinkedIn Pic:** [**link**](https://www.linkedin.com/posts/shabad-grover-8982672a3_datascience-python-healthcareanalytics-activity-7316869159831642112-hC52?utm_source=share&utm_medium=member_desktop&rcm=ACoAAEkv4jQBDE36hG7qy0kHCii0j6C64jnmNS4)

****