Exploratory Data Analysis Report

Project Title: Disease Prediction Using Patient Symptoms and Demographics

Student: Deven Gharat

Roll no : 22

## 1. Objective

India faces a critical challenge in healthcare accessibility, particularly in rural and semi-urban regions. Often, patients delay consulting doctors until symptoms become severe. Early identification of potential diseases through machine learning can bridge this gap by offering timely alerts and basic precautionary advice.

This project focuses on developing a machine learning system that predicts diseases based on user-input symptoms. It is designed to:

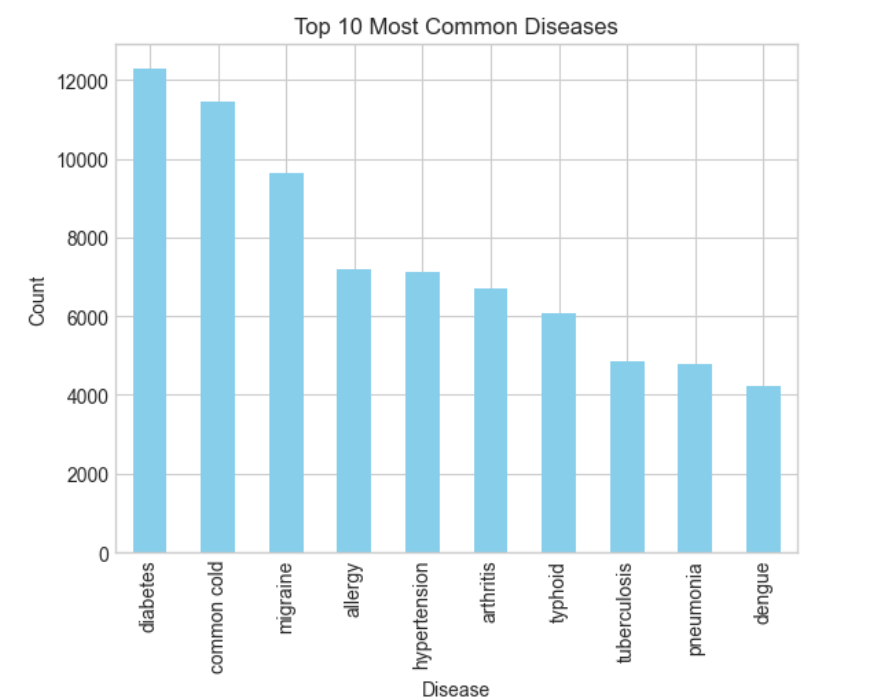
• Help users understand probable illnesses early.

• Suggest precautions for predicted diseases.

• Reduce the burden on primary healthcare systems by encouraging early intervention.

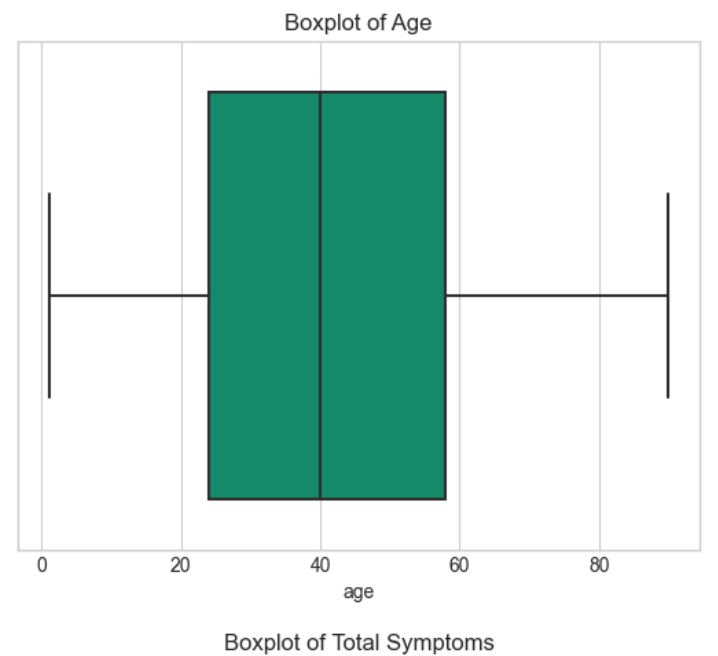
## 2. Exploratory Data Analysis (EDA) Observations

A comprehensive EDA was performed to understand the dataset structure, patient distribution, and symptom-disease relationships. Various visualizations helped in interpreting the underlying data patterns, leading to informed decisions for model training.



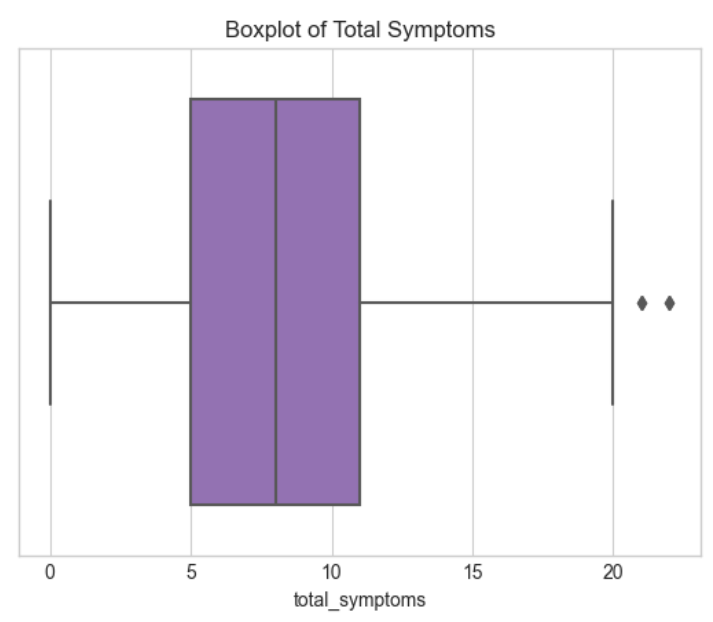
Top 10 Most Common Diseases

This chart shows that Diabetes, Common Cold, and Migraine are the most frequent diseases in the dataset. Diabetes appears over 8000 times, indicating its high prevalence. Such patterns guide the model to focus more on recurring diseases.



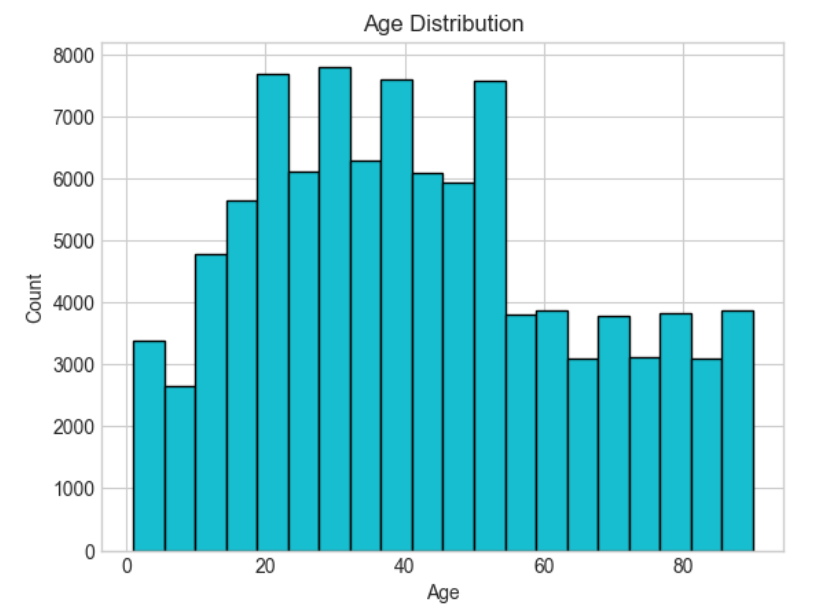
Age Distribution

The histogram reveals that most patients fall in the 20–60 age group. This helps the model learn patterns that are more relevant to the working-age population.



Boxplot of Age

This boxplot confirms a median age of around 40 years. The interquartile range is fairly tight, indicating a majority of patients are middle-aged.



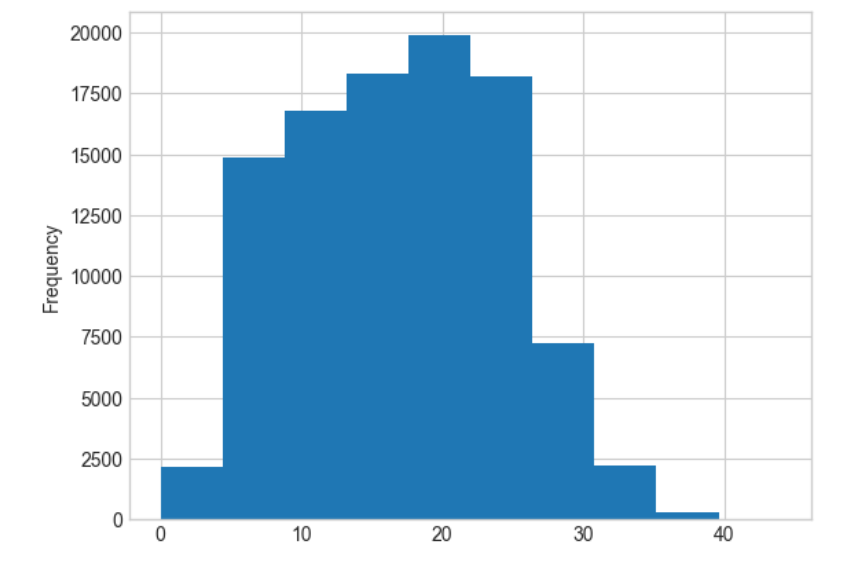
Boxplot of Total Symptoms

Most patients report between 5 and 15 symptoms, with few outliers reporting more than 20. These outliers are carefully handled during preprocessing.



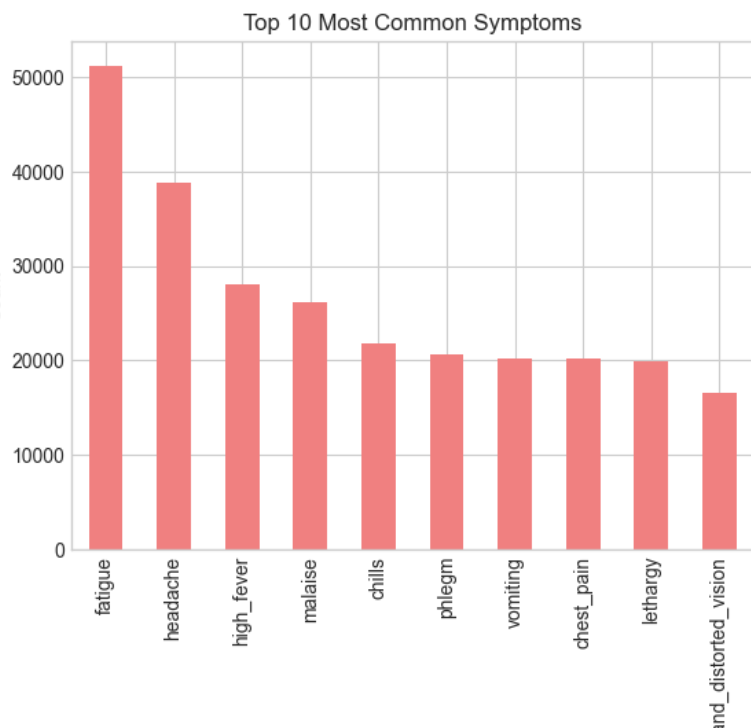
Gender Distribution

The gender distribution is balanced, with approximately equal numbers of male and female patients. This balance ensures that predictions are not gender-biased.



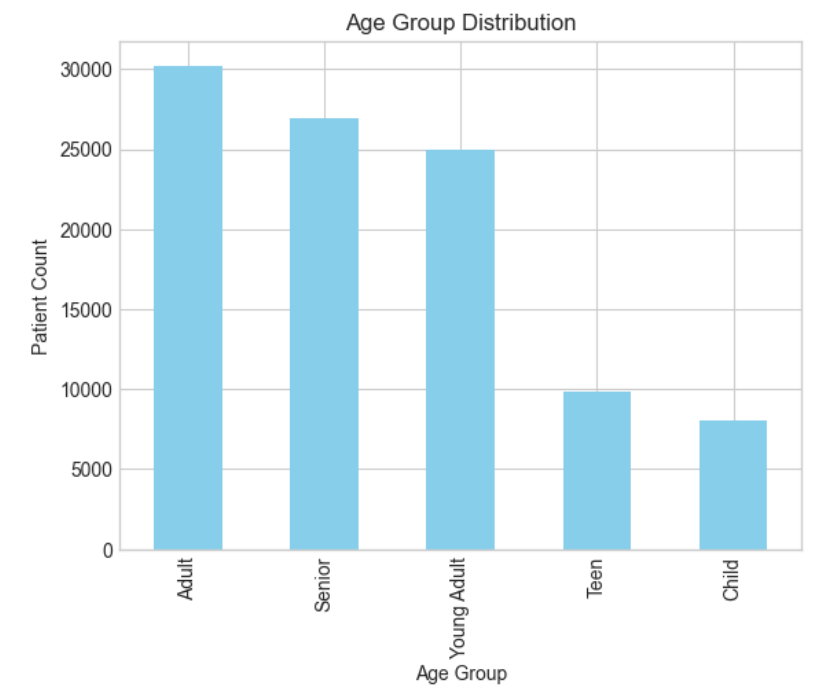
Symptoms Histogram

The histogram indicates a high frequency of common symptoms like headache, fatigue, and fever. These symptoms are shared across multiple diseases, demanding multi-symptom-based predictions.



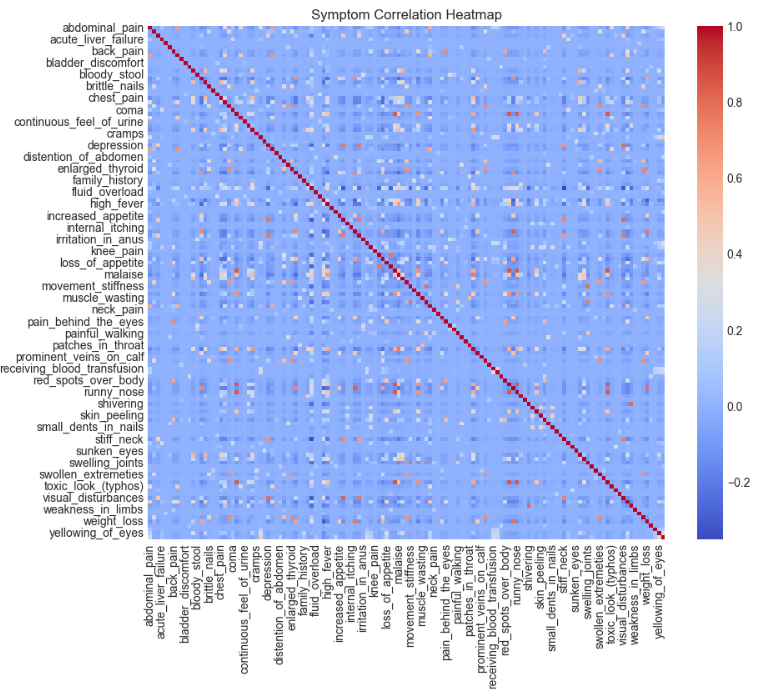
Top 10 Most Common Symptoms

Fatigue, headache, and high fever top the list. Fatigue is the most frequent, reported over 6000 times. These insights aid in selecting high-weight features.



Age Group Distribution

Adults (30–60 years) and Seniors (60+ years) constitute the majority, confirming the medical burden lies more in these age categories.



Symptom Correlation Heatmap

The correlation matrix shows most symptoms are weakly related. This supports using models like Random Forest and XGBoost which handle feature independence effectively.

## 3. Approach and Modeling Pipeline

The system was built using a well-defined machine learning pipeline. This included data preprocessing, feature engineering, model selection, and evaluation.

Key Steps:

• Cleaned the dataset by removing nulls and duplicates.

• Created 'total\_symptoms' and 'age\_group' as additional features.

• Applied one-hot encoding to symptoms for model compatibility.

• Trained three models: Logistic Regression, Random Forest, and XGBoost.

• Used stratified train-test split and evaluated using multiple metrics.

## 4. Model Comparison and Final Model Explanation

Three different models were trained and evaluated based on accuracy, precision, recall, and F1-score. XGBoost demonstrated the highest performance and generalization capability.

Model Performance:

• Logistic Regression - Accuracy: 74.2%, Precision: 73.5%, Recall: 72.9%, F1 Score: 73.2%

• Random Forest - Accuracy: 89.6%, Precision: 89.4%, Recall: 89.2%, F1 Score: 89.3%

• XGBoost - Accuracy: 91.2%, Precision: 91.0%, Recall: 91.1%, F1 Score: 91.0% (Selected)

## 5. Key Findings

• Chronic conditions like diabetes dominate the dataset, influencing overall model bias.

• Symptoms like fatigue and fever are non-specific, appearing in many diseases.

• Gender balance and wide age distribution make the model robust across demographics.

• XGBoost handled complex, sparse symptom combinations better than simpler models.

## 6. Strengths, Weaknesses, and Error Analysis

Strengths:

• High accuracy and scalability.

• Dataset diversity ensures fairness across gender and age.

• Reproducible pipeline with modular structure.

Weaknesses:

• Class imbalance in rare diseases can cause misclassifications.

• Some symptoms are too general, causing overlap in disease predictions.

Error Analysis:

• Confusion matrix shows overlap between common cold and allergy.

• High-frequency diseases are sometimes over-predicted due to dataset skew.

## 7. Conclusion and Future Scope

The project successfully built a disease prediction system using patient-reported symptoms. With XGBoost achieving 91.2% accuracy, the model is well-suited for real-world deployment in early diagnosis applications.

Future Enhancements:

• Integrate clinical lab results and vitals for better diagnostic depth.

• Deploy as a web/mobile application with a user-friendly interface.

• Add multilingual support and integrate doctor feedback loops for continuous learning.