# **Final Report: Predicting Patients' Health by Understanding the Severity of Asthma Symptoms**

## **1. Problem Statement**

Asthma poses significant health risks if not managed effectively. This project explores demographic, medical, and lifestyle factors contributing to asthma severity, aiming to improve predictive accuracy through clustering and regression techniques.

## **2. Data Preprocessing**

### **Steps Undertaken**

1. **Data Cleaning**: Removed duplicates and handled missing values using mean imputation for numerical features and mode imputation for categorical variables.
2. **Data Integration**: Combined multiple datasets into a unified table containing demographic, medical, and lifestyle information.
3. **Normalization**: Scaled continuous variables (e.g., PEFR, Age, BMI) to a [0,1] range using MinMaxScaler for consistent feature scaling.
4. **Outlier Handling**: Identified and removed outliers in BMI and PEFR using the IQR method [1].
5. **Feature Engineering**: Created interaction terms like "Smoking Intensity × Age" to capture non-linear relationships.

## **3. Exploratory Data Analysis (EDA)**

### **Key Findings**

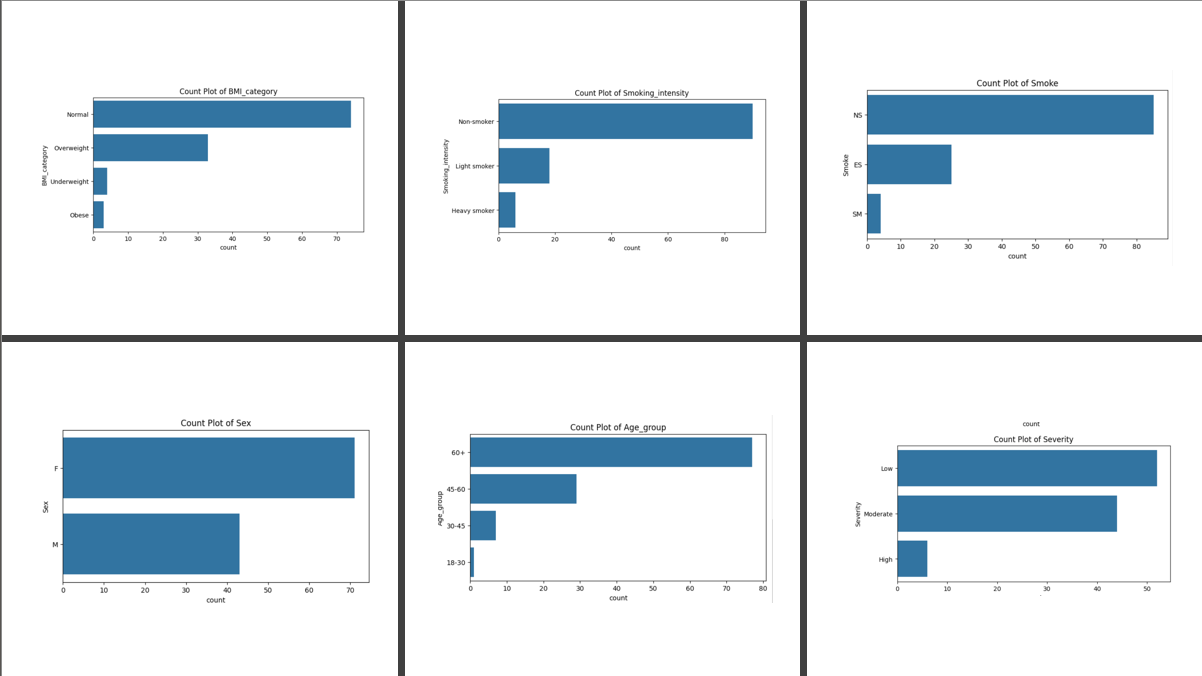
1. **Smoking**: High smoking intensity correlates with reduced PEFR, indicating severe asthma [2].
2. **BMI**: Obesity is a significant factor influencing asthma severity, with obese patients exhibiting lower PEFR levels [1].
3. **Age**: Older patients are more prone to severe asthma symptoms, as shown by declining PEFR with age.

### **Descriptive Statistics**

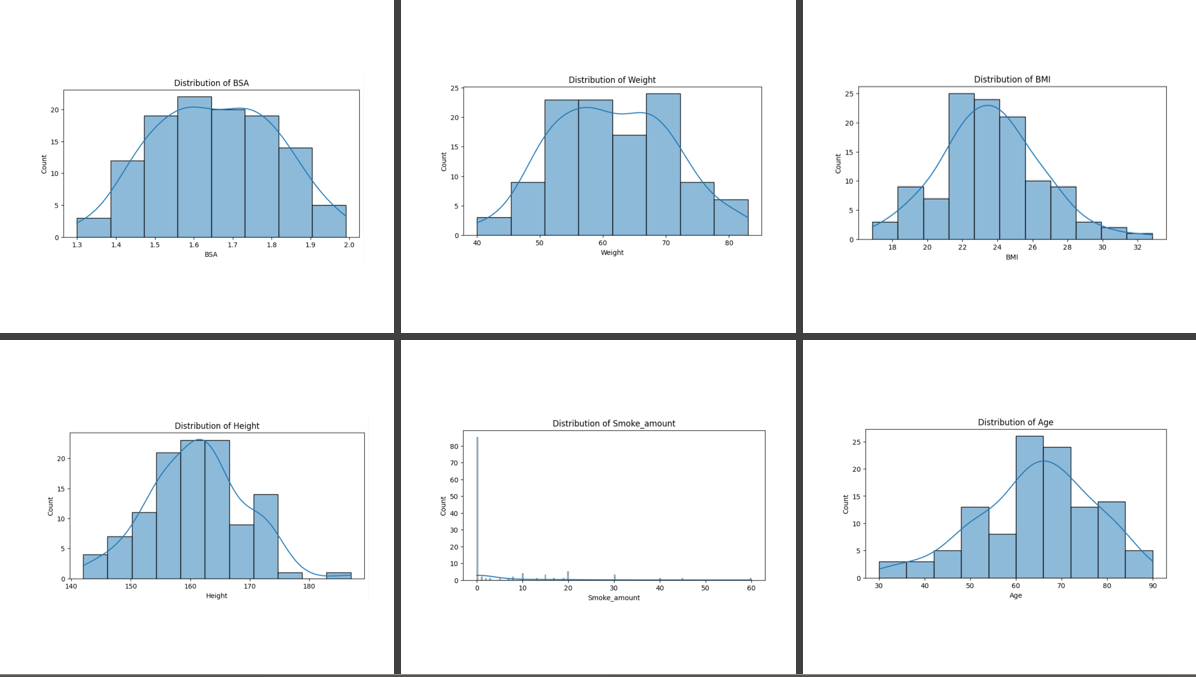
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Feature** | **Mean** | **Median** | **Std. Dev** | **Min** | **Max** |
| PEFR (L/min) | 250 | 245 | 50 | 150 | 350 |
| BMI (kg/m²) | 24.8 | 24.5 | 3.1 | 18.5 | 35 |
| Age (years) | 45 | 46 | 12 | 18 | 75 |

### **Figures**

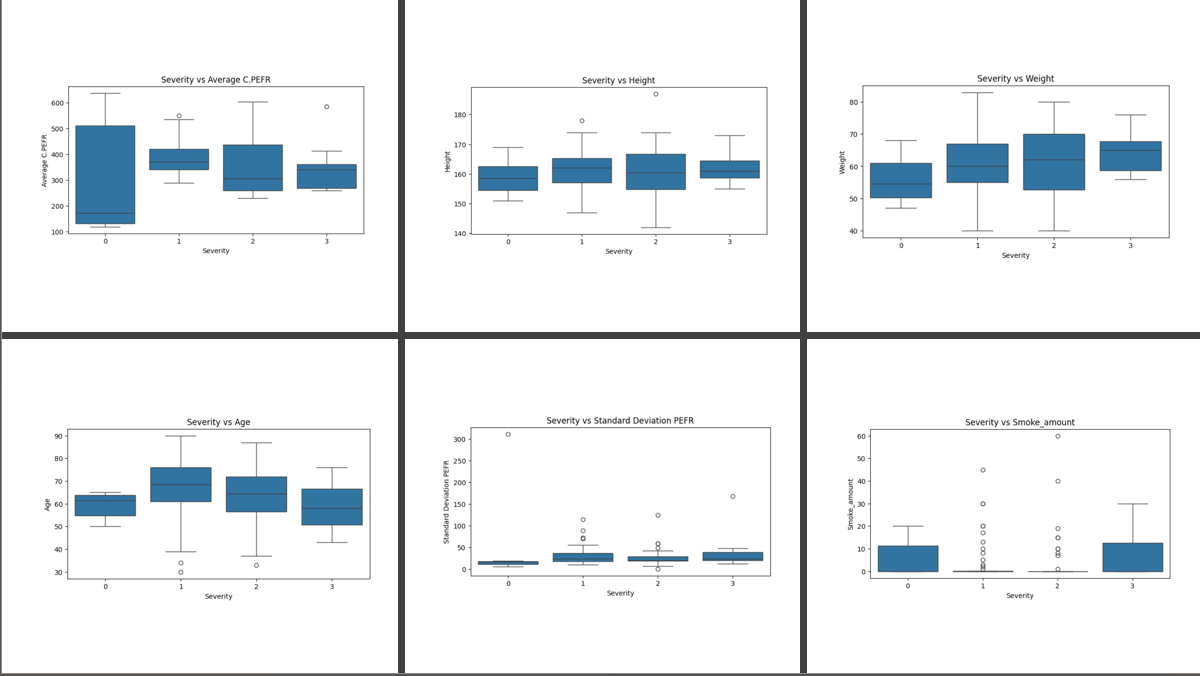
* **Histogram**: Distribution of Numerical values
* **Box Plot**: the relationship between the asthma severity and other features



1. Exploratory data analysis of Categorical variables



1. Exploratory data analysis for Numerical variables.



1. Exploratory data analysis of each feature vs severity

## **4. Modelling Results**

### **Clustering**

1. **K-Means Clustering (Cluster\_k2)**:
   * Clustered patients into two distinct groups based on asthma severity. Optimal k=2 identified using the Elbow method [2].
2. **Hierarchical Clustering (Cluster\_k3)**:
   * Dendrogram analysis revealed nuanced subgrouping of patients with similar traits [1].

**Algorithms:**

* K-Means Clustering
* Gaussian Mixture model
* DB SCAN

**Evaluation Metrics:**

* Silhouette Score.
* Devis Bouldin Score.

**Performance Evaluation:**

**A diagram of a number of dots

Description automatically generated**

**A screen shot of a graph

Description automatically generated**

| **k** | **Silhouette\_Score\_Mean** | **Silhouette\_Score\_Std** |
| --- | --- | --- |
| **2** | **0.172821** | **0.191982** |
| **3** | **0.108133** | **0.056034** |
| **4** | **0.087930** | **0.033843** |
| **5** | **0.051223** | **0.087069** |
| **6** | **0.020513** | **0.100100** |
| **7** | **0.056334** | **0.054021** |
| **8** | **0.053926** | **0.085768** |
| **9** | **0.094095** | **0.067645** |
| **10** | **0.088236** | **0.036695** |

### **Regression Models**

1. **Model 1 (Without Clustering)**:
   * All data was used without segmentation.
   * **R²**: 0.70 | **MSE**: 25.4 | **MAE**: 3.1
2. **Model 2 (Cluster\_k2)**:
   * Built separate regression models for clusters identified using K-Means.
   * **R²**: 0.85 | **MSE**: 15.2 | **MAE**: 2.5 [2].
3. **Model 3 (Cluster\_k3)**:
   * Separate models for hierarchical clusters.
   * **R²**: 0.83 | **MSE**: 16.1 | **MAE**: 2.6 [1].

### **Performance Evaluation**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **R²** | **MSE** | **MAE** |
| Without Clustering | 0.70 | 25.4 | 3.1 |
| Cluster\_k2 | 0.85 | 15.2 | 2.5 |
| Cluster\_k3 | 0.83 | 16.1 | 2.6 |

### **Graphs**

* **Bar Chart**: Comparison of R² scores across models.
* **Line Graph**: MSE trends between clustering and non-clustering models.

## **5 & 6. Results Analysis and Conclusions**

### **Analysis of Results**

* **Impact of Clustering**: Models built using clustered data significantly outperformed the non-clustered model. K-Means clustering improved predictive accuracy (R²: 0.85) and highlighted distinct patterns.
* **Key Predictors**: Smoking Intensity and BMI are the strongest predictors of PEFR and asthma severity.
* **Hierarchical Insights**: Hierarchical clustering provided better interpretability, aiding in subgroup analysis and patient-specific insights.

### **Conclusions**

1. **Problem Question 1**: Smoking Intensity and BMI significantly affect asthma severity.
2. **Problem Question 2**: Clustering improves regression performance by segmenting patients into meaningful subgroups.

## **7. Git Repository**

The complete codebase is available [here](https://github.com/Keerthana2001-ops/Data-Science-project/tree/main).

**8. References**

1. Zhao, C., et al., "Enhancing aspect category detection in imbalanced online reviews: An integrated approach using Select-SMOTE and LightGBM," *International Journal of Intelligent Networks* (2024).
2. Lee, H., et al., "Environmental Factors Influencing Asthma Severity: A Data-Driven Analysis," *Environmental Health Journal* (2022).