**Dataset Overview**

The dataset contains 19 columns and 2,357 rows, providing information about users, their medication, and related side effects. Below is the detailed structure of the dataset:

**Data Types**

The dataset includes a mix of data types:

• Integer (int64):

• 1 column (Kullanici\_id)

• Floating-point numbers (float64):

• 2 columns (Kilo, Boy)

• Datetime (datetime64):

• 4 columns (Dogum\_Tarihi, Ilac\_Baslangic\_Tarihi, Ilac\_Bitis\_Tarihi, Yan\_Etki\_Bildirim\_Tarihi)

• Categorical/String (object):

• 12 columns, including Cinsiyet, Uyruk, Il, Ilac\_Adi, Yan\_Etki, Kan Grubu, and family health history columns.

**Non-null Values**

Some columns have missing values, with non-null counts as follows:

• Cinsiyet: 1,579 non-null entries

• Il: 2,130 non-null entries

• Kilo: 2,064 non-null entries

• Boy: 2,243 non-null entries

• Other columns, such as Kullanici\_id, Dogum\_Tarihi, and Ilac\_Adi, have no missing data.

**Descriptive Statistics**

The following key metrics describe the dataset’s numerical features:

• Kullanici\_id: IDs range from 1 to 196, with an average ID of 97.

• Dogum\_Tarihi: Dates range from October 12, 1939, to April 25, 2011.

• Ilac\_Baslangic\_Tarihi and Ilac\_Bitis\_Tarihi: The medication dates span from January 1, 2022, to March 19, 2022.

• Kilo (Weight): The average weight is 80.86 kg, with a range from 50 kg to 110 kg.

• Boy (Height): The average height is 174.64 cm, with a range from 145 cm to 203 cm.

**Standard Deviations**

• Weight (Kilo): The standard deviation is 18.64 kg, indicating variability in the users’ weights.

• Height (Boy): The standard deviation is 16.52 cm, indicating moderate variability in height.

**Key Findings**

1.**Missing Values Heatmap**:

• **Analysis**: The heatmap clearly indicates that several columns in the dataset contain missing values, particularly in columns like Cinsiyet, Il, Alerjilerim, Kronik Hastaliklarim, Kan Grubu, Kilo, and Boy. These gaps highlight the need for data cleaning and imputation strategies to ensure data completeness before proceeding to further analysis.

• **Key Finding**: Missing data is pervasive in many important fields, suggesting that further imputation or removal of rows with incomplete data is crucial for accurate model predictions.

2.**Histograms of Weight (Kilo) and Height (Boy)**:

• **Analysis**: The distribution of weight (Kilo) and height (Boy) shows moderate variance, with most weights falling between 60 to 100 kg and heights ranging between 160 to 190 cm. This indicates a relatively balanced distribution for these features, with a few outliers.

• **Key Finding**: Both height and weight follow a somewhat uniform distribution with slight variations, suggesting that these variables can contribute well to any predictive modeling.

3.**Distribution Plots for Date Fields (e.g., Medication Dates)**:

• **Analysis**: The visualizations for date fields like Ilac\_Baslangic\_Tarihi and Yan\_Etki\_Bildirim\_Tarihi show distinct patterns where the majority of data points cluster within a specific range of dates. There seems to be consistency in the medication start and end times, indicating most patients have similar medication periods.

• **Key Finding**: Medication start and side effect reporting dates exhibit clustered behavior, showing that the data captures a consistent period of observation for most users.

4.**Correlation Matrix between Height and Weight**:

• **Analysis**: The correlation matrix between Kilo (Weight) and Boy (Height) reveals a weak negative correlation (-0.15). This indicates that height and weight are not strongly correlated, suggesting that each feature should be treated independently in modeling processes.

• **Key Finding**: Height and weight do not have a strong linear relationship, allowing them to provide distinct information when used as features for machine learning models.

5.**Scatter Plot of Height vs Weight**:

• **Analysis**: The scatter plot visualizes the distribution of height and weight across the population. The spread is wide, showing that individuals with the same height can have varying weights and vice versa. This variability should be taken into account when modeling for any predictive purpose related to body metrics.

• **Key Finding**: There is significant variability in the relationship between height and weight, indicating that these features may contribute diverse information when used together in prediction models.

6.**Drug Usage Distribution**:

• **Analysis**: The histogram of drug usage duration shows a near-normal distribution, with most individuals using medication for 50-75 days. This consistent drug usage pattern suggests that the dataset captures an extended follow-up period for most patients.

• **Key Finding**: Drug usage duration follows a near-normal distribution, making it an important feature for understanding patient behavior and possible side effects.

7.**Boxplots for Weight and Height**:

• **Analysis**: The boxplots for weight and height show the median, quartiles, and the range of values for each feature. Both plots indicate that the distributions are fairly centered, with no significant outliers, which suggests a clean dataset for these particular variables.

• **Key Finding**: The boxplots for height and weight show consistent central tendencies with minimal outliers, making these features reliable for modeling purposes.

8.**Side Effects by Age Groups**:

• **Analysis**: The bar chart compares the rates of different side effects (e.g., fatigue, sleep issues, vision problems) across age groups. Younger age groups (0-18) tend to report the highest rate of side effects, while older age groups (66+) have increased reports of high blood pressure.

• **Key Finding**: Age significantly influences the prevalence of different side effects, with younger patients reporting more side effects overall, and older patients facing more serious conditions like high blood pressure.

**9. Most Used Drugs by City:**

• **Analysis**: The bar plot shows the most frequently used drugs in different cities. The usage patterns differ significantly between regions, with some cities having a preference for specific medications.

• **Key Finding**: Different cities exhibit varying patterns of drug usage, suggesting that geographical location might influence the choice of medication, potentially due to regional medical practices or demographics.

**10**.**Time to Onset of Side Effects**:

• **Analysis**: The distribution of how many days after the medication start date side effects occur shows a peak at around 30-35 days. This suggests that most patients begin to experience side effects about one month after starting medication.

•**Key Finding**: Side effects tend to manifest within the first month of drug usage, providing a crucial time window for monitoring patient reactions.

**11. Most Common Side Effects and Allergies:**

•**Analysis**: This bar chart compares common side effects such as fatigue, taste change, high blood pressure, and blurred vision with allergies like coffee, seafood, tomatoes, and spinach. The overlap between specific side effects and allergic reactions is apparent, highlighting potential areas for further research on allergenic responses to medications.

•**Key Finding**: Fatigue and taste alteration are the most frequently reported side effects, while coffee and seafood allergies are also common, suggesting a relationship between diet and medication side effects.

**12. Side Effects by Gender:**

•**Analysis**: This visualization compares the prevalence of various side effects among men and women. For example, women experience more high blood pressure and abdominal pain, while men report more symptoms like fatigue and taste alteration.

•**Key Finding**: There are gender differences in how side effects manifest, with women experiencing more cardiovascular-related symptoms, while men tend to report sensory issues like taste alteration and fatigue.

**13. Most Common Allergies by Gender:**

•**Analysis**: This set of bar charts compares the most common allergies reported by men and women. Men show a higher frequency of allergies to tomatoes and dust, while women are more allergic to seafood and spinach.

•**Key Finding**: Gender plays a role in allergy susceptibility, with significant differences in food allergies such as tomatoes for men and seafood for women.

**14. Chronic Diseases Most Suffered by Gender:**

•**Analysis**: This comparison highlights the chronic diseases most prevalent in men and women. Men are more likely to suffer from osteoporosis and diabetes, while women report higher instances of Alzheimer’s and hypertension.

•**Key Finding**: Chronic disease patterns vary significantly between genders, suggesting that healthcare strategies should be tailored to address gender-specific health issues.

**15. Chronic Disease Distribution by Cities:**

•**Analysis**: This series of bar charts presents the distribution of chronic diseases across various cities. Cities like Kayseri and Istanbul show higher cases of chronic diseases such as diabetes and heart disease.

•**Key Finding**: Chronic diseases like hypertension and diabetes have higher prevalence in urban areas like Istanbul and Kayseri, likely due to lifestyle factors associated with larger cities.

**16. Side Effects by Blood Group:**

•**Analysis**: This chart demonstrates the different side effects experienced by individuals based on their blood group. For example, those with B+ blood type report more fatigue and taste alteration, while those with O+ report more constipation and blurred vision.

•**Key Finding**: Blood type may influence how individuals experience medication side effects, suggesting a potential link between genetics and drug response.

**17. Allergies by Blood Group:**

•**Analysis**: This chart explores how different blood groups experience food allergies. Blood types like O+ report more allergies to spinach and coffee, while A+ types show more sensitivity to tomatoes.

•**Key Finding**: Blood type appears to correlate with specific food allergies, indicating that genetic factors could play a role in allergic reactions.

**Visualization Techniques**

**1.Heatmaps:**

•Purpose: To identify missing values and patterns of null data across the dataset.

•Application: A heatmap was used to visually display missing values in columns like Cinsiyet, Il, and Kronik Hastaliklarim. This visualization highlighted where data cleaning or imputation would be necessary before model building.

•Impact: It provided a quick, comprehensive view of missing data, helping to decide on appropriate strategies for handling incomplete records.

**2.Bar Charts:**

•Purpose: To compare categorical data or frequencies.

•Application: Bar charts were used in several contexts, such as comparing the side effects experienced by different gender groups and identifying the most common allergies for both men and women. Another example is visualizing the most used drugs across cities, which gives insight into regional differences in drug prescriptions.

•Impact: These comparisons helped identify key trends, such as gender differences in medication side effects and geographical patterns in drug usage and chronic diseases.

**3.Histograms:**

•Purpose: To display the distribution of continuous variables.

•Application: Histograms were employed to analyze the distribution of variables like weight (Kilo), height (Boy), and the number of days between medication use and side effect onset. These plots allowed for the detection of outliers and patterns in data distribution.

•Impact: They provided a detailed understanding of the spread and central tendencies in continuous variables, aiding in feature engineering for predictive modeling.

**4.Boxplots:**

•Purpose: To visualize the central tendency, dispersion, and potential outliers in continuous data.

•Application: Boxplots were used to visualize the range, median, and quartiles of weight and height distributions. This technique helped highlight potential outliers in the dataset.

•Impact: It gave a concise summary of the data’s distribution, making it easier to identify whether the data is skewed or contains outliers.

**5.Scatter Plots:**

•Purpose: To display the relationship between two continuous variables.

•Application: Scatter plots were used to explore the relationship between weight and height, helping to determine if there was any significant correlation between the two variables.

•Impact: This visualization confirmed the weak correlation between height and weight, providing insight into how these variables should be treated in further analyses.

**6.Correlation Matrices:**

•Purpose: To show the strength of relationships between numerical variables.

•Application: A correlation matrix was used to explore relationships between variables like Kilo and Boy. This method quickly highlighted weak or strong correlations, such as the weak correlation between height and weight.

•Impact: It was crucial in understanding the underlying relationships between features, guiding which variables to include in the modeling process.

**2. Data Preprocessing**

**Handling Missing Data:**

• **Missing Values**: Missing data, particularly in the Cinsiyet (Gender), Kilo (Weight), and Boy (Height) columns, were handled using various techniques. For missing gender values, a RandomForestClassifier model was trained to predict the missing values based on available data (Kan Grubu, Boy, Kilo), and the predicted values were filled in.

• **Imputation for Missing Numerical Values**: Mean imputation was performed on missing weight and height values. This was done by grouping the data by gender and calculating the mean for each gender, then filling the missing values with these calculated averages.

• **Categorical Missing Values**: Missing categorical values were filled with ‘Bilinmiyor’ (Unknown) to retain completeness in the dataset without losing too much information.

**Encoding Categorical Variables:**

• **Label Encoding**: Categorical variables like Cinsiyet (Gender) and Kan Grubu (Blood Type) were label-encoded. This was necessary to convert these string-based categories into numerical format so they could be used in the RandomForest model for predicting missing values and further analysis.

• **OneHot Encoding**: For larger categorical variables, such as Uyruk (Nationality), Il (City), Ilac\_Adi (Drug Name), and various family health columns, OneHotEncoder was used. This ensured that all categorical variables were transformed into numerical representations without any ordinal assumptions, preventing bias in the machine learning models.

**Splitting Data:**

• **Train-Test Split**: The data was split into training and test sets using an 80/20 split. This allowed the model to be trained on the majority of the data while keeping a portion aside to evaluate its performance. The model trained on features such as Kan Grubu, Boy, and Kilo, and predicted the target variable Cinsiyet.

**Model Selection:**

• **RandomForest Classifier**: A RandomForestClassifier was selected to predict missing gender values and handle the classification task. This model was chosen due to its robustness in handling both categorical and numerical variables, as well as its ability to handle missing data through its decision tree structure. The model was trained using the data with complete Cinsiyet values and then applied to predict missing values in the dataset.

**3. Conclusion and Further Steps:**

• **Summary**: The preprocessing steps involved handling missing data, encoding categorical variables, and scaling numerical features. These steps ensured that the data was ready for further analysis or model development. The combination of label encoding and OneHotEncoding allowed for appropriate handling of both small and large categorical variables, while scaling ensured that the numerical features were standardized for better model performance.

• **Future Improvements**: Potential improvements could include experimenting with different machine learning models or tuning the hyperparameters of the RandomForestClassifier. Additionally, further analysis could be conducted to explore interactions between encoded categorical features.