Devika S Vinod

3 MSDS B

23122113

Github Link :

Statistical Profiling of a Clinical UTI Dataset for Machine Learning-Based Diagnosis

Introduction :

Step into the realm of urinary tract infections (UTIs) with this comprehensive analysis of a dataset sourced from a local clinic in Northern Mindanao, Philippines. Spanning from April 2020 to January 2023, this dataset was meticulously collected for the groundbreaking research study titled "Optimizing UTI Diagnosis with Machine Learning and Artificial Neural Network for Reducing Misdiagnoses," authored by Agdeppa et al. (2023). Embedded within are diverse features ranging from demographic indicators like age and gender to intricate urine characteristics (colour, transparency, glucose, protein, pH, specific gravity), blood cell counts (WBC, RBC), epithelial cells, mucous threads, amorphous urates, bacteria presence, and the decisive UTI diagnosis (NEGATIVE or POSITIVE). Our journey through this analysis aims not just to preprocess and explore the data, but to unearth profound insights critical for sculpting machine learning models that refine UTI diagnosis and alleviate misdiagnoses.

Dataset Description:

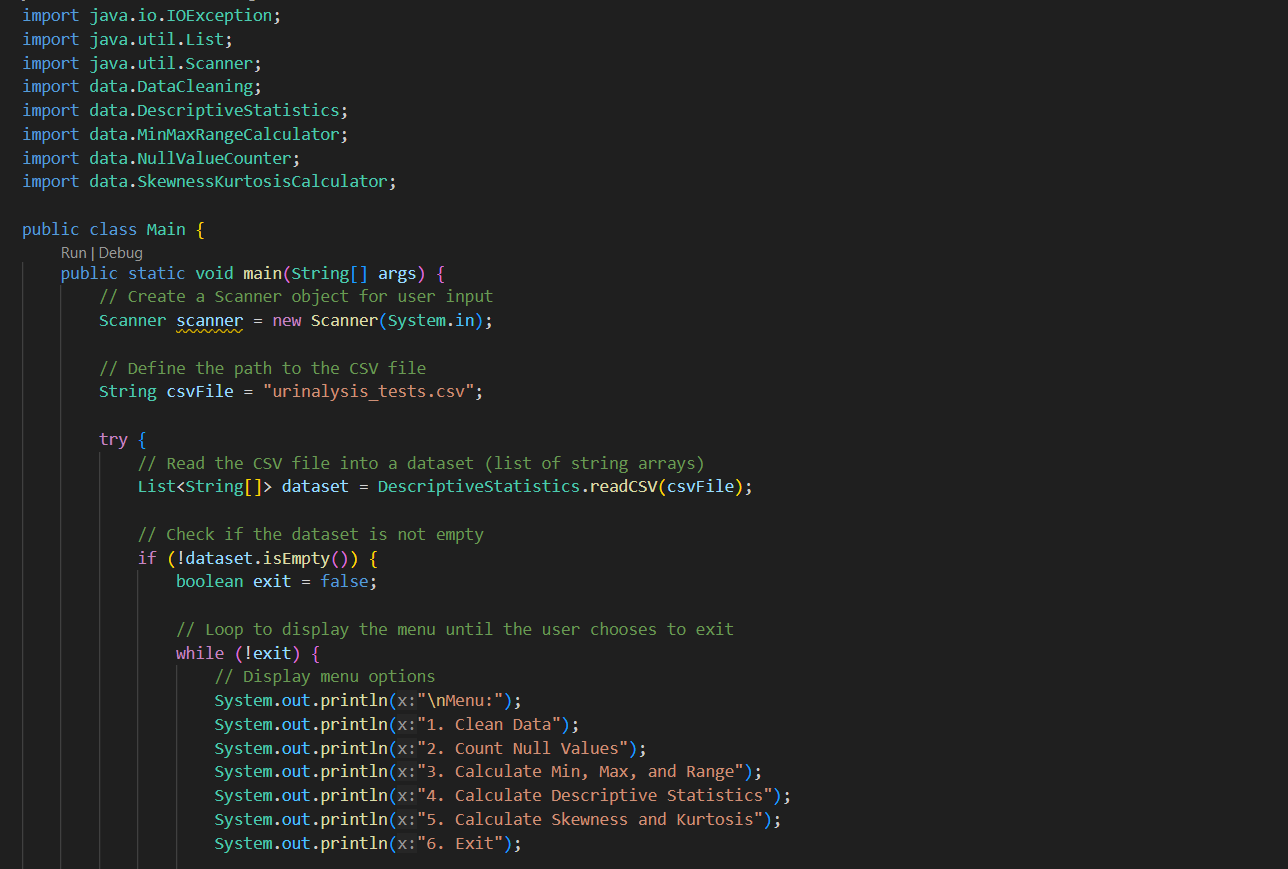
The dataset consists of the following features:

1. **Age :** The age of the patient, with some patients being months old (in which case, the age is pre-processed by dividing it by 100, e.g., 8 months old would be represented as 0.08).
2. **Gender :** The gender of the patient, either male or female.
3. **Color :** The color of the patient's urine.
4. **Transparency :** The transparency of the patient's urine.
5. **Glucose :** The presence of glucose in the patient's urine, which can be an indicator of certain health conditions.
6. **Protein :** The presence of protein in the patient's urine, which is examined to assess kidney function and detect potential issues.
7. **pH :** The pH level of the patient's urine, measuring its acidity or alkalinity.
8. **Specific Gravity :** The specific gravity of the patient's urine, which is a measure of the concentration of particles in the urine compared to water.
9. **WBC (White Blood Cells) :** The count of white blood cells (leukocytes) in the patient's urine, which are a crucial part of the immune system.
10. **RBC (Red Blood Cells) :** The count of red blood cells in the patient's urine, which are responsible for carrying oxygen throughout the body.
11. **Epithelial Cells :** The presence of epithelial cells in the patient's urine, which line the surfaces and cavities of the body, including the urinary tract.
12. **Mucous Threads :** The presence of mucous threads (strands of mucus) in the patient's urine.
13. **Amorphous Urates :** The presence of amorphous urates (non-crystalline formations consisting of uric acid) in the patient's urine.
14. **Bacteria :** The presence of bacteria in the patient's urine.
15. **Diagnosis :** The diagnosis of whether the patient has a UTI or not, represented as either NEGATIVE or POSITIVE.

Main File :

The **Main** class performs the following tasks:

1. **User Interaction**: Uses a menu-driven approach to interact with the user, allowing them to select different data analysis operations.
2. **Dataset Handling**: Reads the urinalysis test results from a CSV file and processes the data based on user input.
3. **Data Analysis Operations**: Offers multiple operations, including data cleaning, null value counting, min-max range calculation, descriptive statistics computation, and skewness and kurtosis calculation.

A screen shot of a computer program

Description automatically generated

***A screenshot of a computer

Description automatically generated***

***Output:***



Data Cleaning:

Data cleaning is a crucial step in any data analysis or machine learning project, as it helps to ensure the quality and reliability of the data. In Java, there are several libraries and frameworks available for data cleaning. Those libraries provide various methods and functionalities for dealing with common data cleaning tasks, such as handling missing values, removing duplicates, and formatting data. Here after Data Cleaning, the cleaned data got stored into new csv File.

A screen shot of a computer code

Description automatically generated

A screen shot of a computer program

Description automatically generatedA screen shot of a computer code

Description automatically generated

***Output:***

A black screen with text

Description automatically generated

Null Value Count:

Null value counting is an important step in data cleaning and preprocessing, as it helps identify columns or features that contain missing or null values. In Java, you can use various approaches to count null values in a dataset, depending on the data structure and the libraries or frameworks you're using.

A screen shot of a computer code

Description automatically generated

A screen shot of a computer code

Description automatically generated

***Output:***



Minmax Range Calculator:

The Min-Max Range Calculator is a utility that calculates the minimum and maximum values for each column or feature in a dataset. This information is crucial for understanding the range of values present in the data, which can be valuable for various purposes, such as data normalization, feature scaling, identifying potential outliers, and setting appropriate thresholds or boundaries for data processing tasks.

In Java, we can implement a Min-Max Range Calculator using different approaches, depending on the data structure and libraries or frameworks you're working with.

A screen shot of a computer program

Description automatically generated

A screen shot of a computer program

Description automatically generatedA screen shot of a computer program

Description automatically generated

***Output :***

A screenshot of a computer

Description automatically generated

Descriptive Statistics:

Descriptive statistics play a crucial role in understanding the characteristics of a dataset. They provide summary measures that describe the central tendency, dispersion, and distribution of the data. These measures offer valuable insights into the nature of the variables, which can inform various stages of the data analysis and modelling process.

In the context of this study, descriptive statistics were calculated to gain a comprehensive understanding of the urinary tract infection (UTI) dataset. This analysis involved computing various statistical measures, including:

**Measures of Central Tendency**: These measures provide information about the central or typical values in the dataset. Calculations were performed to determine the mean, median, and mode for each numerical variable. The mean represents the arithmetic average of the values, while the median is the middle value in the sorted list. The mode identifies the value(s) that occur most frequently in the data.

**Measures of Dispersion**: These measures quantify the spread or variability of the data around the central values. The variance and standard deviation were calculated to assess the extent to which the data points deviate from the mean. A higher variance or standard deviation indicates greater dispersion in the data.

A screen shot of a computer code

Description automatically generatedA screen shot of a computer code

Description automatically generatedA screen shot of a computer code

Description automatically generated

A screen shot of a computer program

Description automatically generatedA screen shot of a computer program

Description automatically generatedA screen shot of a computer code

Description automatically generatedA screen shot of a computer program

Description automatically generated

***Output:***

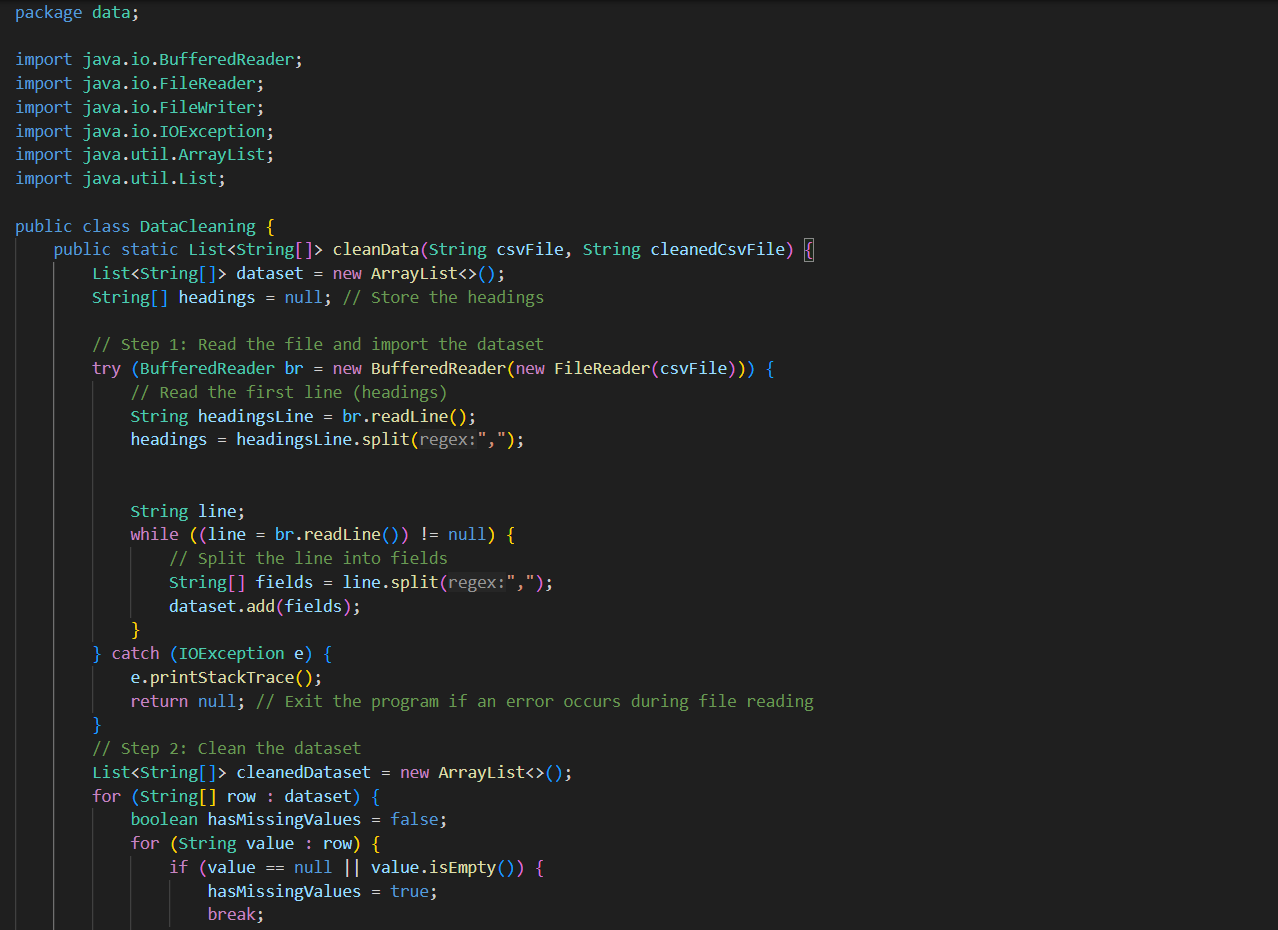
A screenshot of a computer

Description automatically generatedA black rectangle with white dots

Description automatically generated

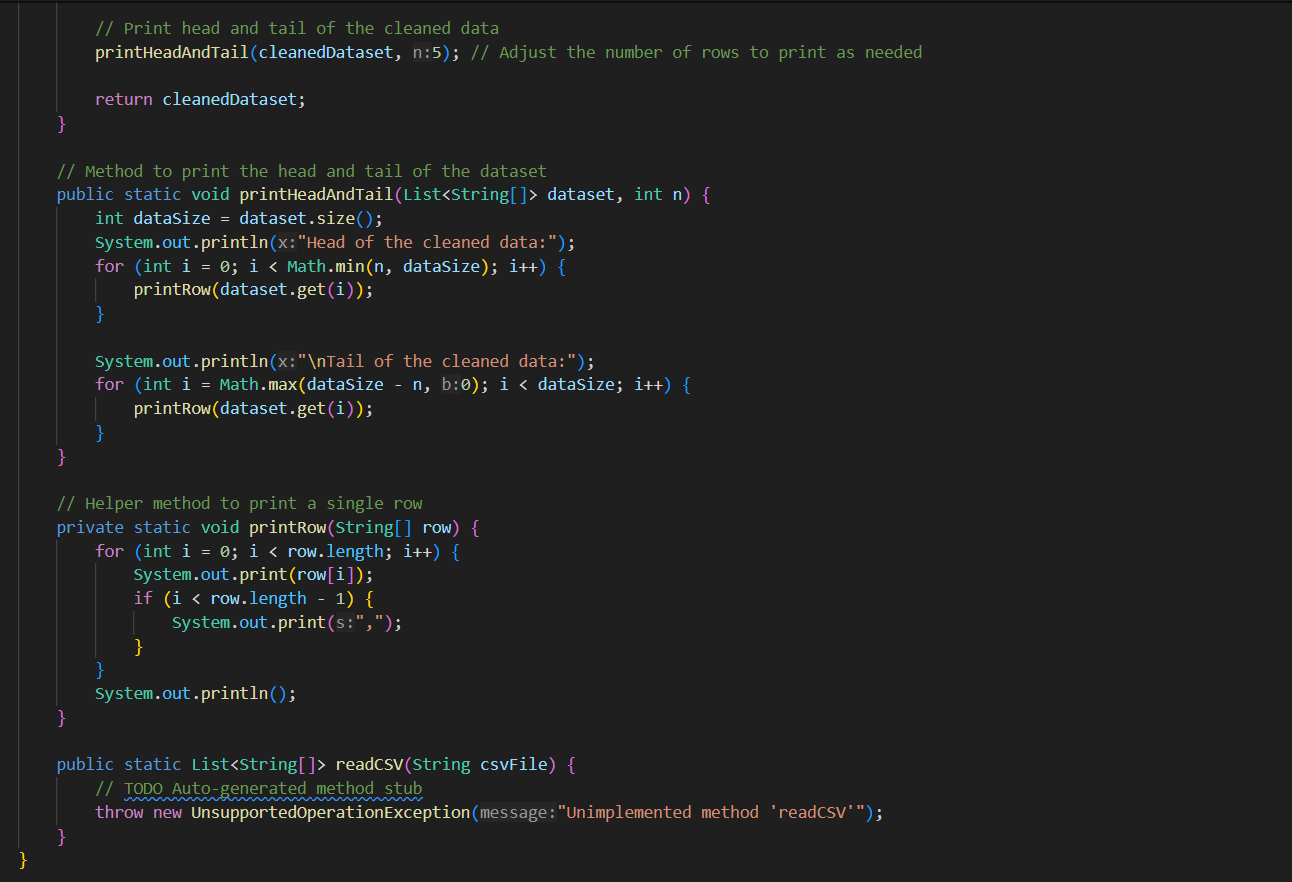
Skewness Kurtosis Calculator *:*

The **Skewness Kurtosis Calculator** is a class or module designed to compute the skewness and kurtosis of numerical columns in a dataset. These statistical measures provide insights into the distribution and shape of the data, which are essential for understanding the characteristics and potential anomalies within the dataset.

******

A screen shot of a computer code

Description automatically generated



***Output :***

***A black rectangle with white dots

Description automatically generated***

Conclusion :

The comprehensive data analysis pipeline presented in this report, including data loading, cleaning, exploration, and advanced statistical analysis, lays a strong foundation for the researchers' goal of optimizing UTI diagnosis using machine learning and artificial neural network techniques. The thorough data preprocessing steps, such as handling missing values and calculating min-max ranges, ensure data quality and reliability for further analysis.

The generation of descriptive statistics and the calculation of skewness and kurtosis provide valuable insights into the central tendency, spread, distribution, and shape of the data. These insights are crucial for effective feature engineering, data transformation, and identifying potential outliers, all of which can significantly impact model performance.

By following best practices in data analysis and leveraging the insights gained from this initial exploration, the researchers can develop accurate and reliable machine learning models that effectively address the challenge of reducing misdiagnoses in UTI diagnosis. However, it is important to consider that the provided dataset covers a specific time period and geographic location, and additional factors or external data sources may be necessary to enhance the robustness and generalizability of the developed models.