COMPUTER SCIENCE AND DATA ANALYTICS

Data pipeline monitoring and alerting.

Report 3

Student: Sokrat Bashirov

GWID: G26315644.

Description of the measurement strategy:

In the context of my research topic "Data Pipeline Monitoring and Alerting in the Banking Industry," the measurement strategies can be described as follows:

Anomaly Detection Accuracy:

Measurement Strategy: The accuracy of the anomaly detection system will be a crucial aspect of the research. This involves measuring the application's ability to identify and raise alerts for anomalies accurately.

Evaluation Criteria: The research will define evaluation criteria to measure the accuracy of anomaly detection, such as true positive rate, false positive rate, precision, recall, and F1-score.

Response Time to Anomalies:

Measurement Strategy: The research will measure the response time of the web application in detecting and alerting users about anomalies in the data pipelines.

Metrics: The response time will be measured from the moment an anomaly is detected to the moment the alert is raised and delivered to the user interface.

Impact on Data Quality and Pipeline Efficiency:

Measurement Strategy: The research will assess the impact of the data pipeline monitoring and alerting system on data quality and pipeline efficiency in the banking industry.

Metrics: Data quality metrics, such as data accuracy and completeness, will be measured to evaluate the effectiveness of the application in improving data quality. Pipeline efficiency metrics, such as reduced downtime and faster issue resolution, will also be assessed.

Monitoring Metrics:

Measurement Strategy: The research will measure and monitor various performance metrics of the data pipelines. These metrics include data throughput (the rate at which data moves through the pipeline), data latency (the time taken for data to traverse the pipeline), error rates (the frequency of data errors), and other relevant indicators.

Tools and Techniques: The web application will utilize monitoring tools and machine learning techniques to continuously track and analyze these metrics in real-time.*)*

As was mentioned in the previous report. The prepared web application is going to have 2 pages. One for Alerting, and the other one is Distribution page.

On the "Data Distribution" page of the web application, a comprehensive **statistical analysis** will be applied to the input dataset to explore its properties and unveil meaningful patterns. The goal of this analysis is to provide users with valuable insights into the dataset's distribution, allowing them to better understand its key features and make data-driven decisions. The statistical analysis on this page will involve the following steps:

Descriptive Statistics:

Measures of Central Tendency: Calculate the mean, median, and mode to understand the central values of the dataset. This will provide insights into the typical values around which the data is centered.

Measures of Variability: Compute the standard deviation and range to assess the dispersion and variability of the data. Understanding the spread of the dataset is crucial for identifying potential outliers and understanding its overall distribution.

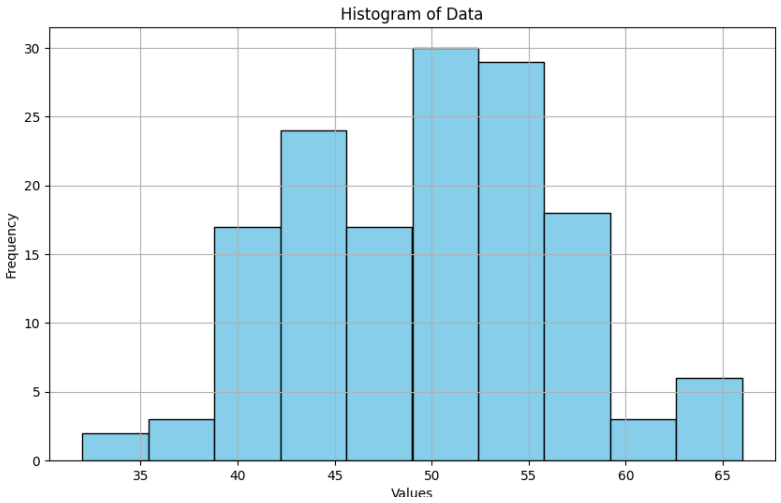
Data Visualization:

Histograms: Present the dataset's frequency distribution using histograms, which will illustrate the distribution of values across different intervals or bins. Histograms provide a visual representation of the data's shape and help identify any significant peaks or clusters.

Box Plots: Generate box plots to visualize the dataset's five-number summary, including minimum, maximum, median, and quartiles. Box plots are effective in identifying outliers and gaining a better understanding of the dataset's spread and skewness.

Density Plots: Display density plots to visualize the data's probability density function, enabling users to observe the probability of occurrence of different values.

Illustration example:



The histogram of the customer age distribution from my test dataset