Executive Summary: Blood Report Analytics

Project Overview

This project aimed to develop a data-driven solution for a diagnostic lab chain to gain insights from their blood test reports. The key deliverables included generating a realistic dataset, performing exploratory data analysis (EDA), developing visualization dashboards, and implementing predictive analytics models.

Dataset Generation

A synthetic dataset of 10,000 (for development, 1M+ for production) anonymized blood reports was generated, simulating patient demographics, various test values (e.g., Haemoglobin, TLC, CRP), and diagnoses. This dataset serves as the foundation for all subsequent analyses.

Exploratory Data Analysis (EDA) Insights

* **Summary Statistics**: Provided a clear understanding of the central tendency and spread of various blood parameters, identifying typical ranges and variations.

* **Outlier Detection**: Identified unusual test results, which could indicate potential data entry errors or rare medical conditions. For example, outliers in Haemoglobin levels were detected using the Interquartile Range (IQR) method.

* **Correlation Matrix**: Revealed relationships between different blood parameters. For instance, a strong positive correlation was observed between Haemoglobin and HCT, as expected physiologically.

Visualization Dashboards

An interactive dashboard was developed using Plotly Dash, offering visualizations for:

* **Infection Trends Over Time**: Displays the daily count of bacterial and viral infections, allowing for the identification of temporal patterns.

* **Anemia Distribution by Age and Gender**: Illustrates the prevalence of anemia across different age groups and genders, highlighting demographic health trends.

* **CRP Trends**: Visualizes C-reactive protein levels over time, which can help monitor inflammation spikes.

Predictive Analytics

A machine learning model (RandomForestClassifier) was implemented to predict diagnoses based on blood test values. The model achieved a high accuracy of **0.994** on the test set, with strong precision, recall, and F1-scores across different diagnosis categories. This demonstrates the potential for automated diagnostic assistance and risk stratification.

Conclusion

This project successfully laid the groundwork for a data-driven approach to managing and understanding diagnostic lab data. The generated dataset, comprehensive EDA, interactive dashboards, and accurate predictive model provide valuable tools for quality monitoring, public health insights, operational efficiency, and patient health analytics.

Key Visualizations

Daily Infection Cases Over Time

Anemia Cases Distribution by Age and Gender

CRP Levels Over Time