Capstone Project Report

Project Title: Predictive Analysis of Pediatric Appendicitis

Course Name: Integrative Data Analytics

Team Members: Habiba Momen, Roger Morocho

Instructor: Dr. MD Kibria

Submission Date: May 2025

1. Introduction

1.1 Background and Context

Appendicitis is one of the most common surgical emergencies in children, and early diagnosis is crucial to prevent complications such as perforation, infection, and prolonged hospitalization. Pediatric appendicitis presents a diagnostic challenge due to the variability of symptoms across age groups. To support better decision-making and reduce the need for unnecessary surgeries, there is a need for accurate, data-driven diagnostic support systems.

This capstone project focuses on the use of real-world clinical data to understand patterns in pediatric appendicitis diagnosis. The dataset contains a mix of demographic, clinical, laboratory, and imaging data that can be used to analyze the factors that most strongly predict appendicitis.

1.2 Objectives of the Project

- To identify the most relevant predictors of appendicitis using clinical and diagnostic data.
- To apply exploratory data analysis (EDA) and dimensionality reduction techniques like Principal Component Analysis (PCA) to uncover patterns.
- To provide actionable insights that can assist medical professionals in early diagnosis and treatment planning.

2. Methodology

2.1 Dataset Description

- **Source**: The dataset is titled *Rogensburg Pediatric Appendicitis Dataset*.
- **Format**: CSV file imported via Python.
- **Size**: Over 100 records, each containing around 30 features.
- **Type of Data**: A mix of numerical and categorical variables, including:
    - Demographics: Age, Sex, Height, Weight, BMI
    - Clinical Scores: Alvarado Score, Pediatric Appendicitis Score

- Diagnostics: Results from ultrasound (Appendix_on_US, Free_fluid, Coprostasis), and final diagnosis confirmation

2.2 Analytical Techniques

- **Data Cleaning & Preprocessing**: Removal of duplicates and unnecessary columns; handling missing values using imputation or deletion; encoding categorical data.
- **Exploratory Data Analysis (EDA)**: Use of statistical summaries, histograms, violin plots, heatmaps, and boxplots to explore relationships.
- **Principal Component Analysis (PCA)**: Used to reduce dimensionality, identify redundancy, and retain essential information for classification.
- **Correlation & Feature Importance**: Heatmaps and bar plots used to identify features most associated with appendicitis.

2.3 Tools and Technologies

- **Programming Language**: Python
- **IDE**: Jupyter Notebook
- **Libraries Used**: pandas, numpy, matplotlib, seaborn, sklearn

3. Analysis and Results

3.1 Data Preprocessing

- **Initial Inspection**: The dataset had columns like "Unnamed: 0" and text-based flags (e.g., 'yes', 'no') which were removed or encoded.
- **Handling Missing Values**: Missing values were present in features like Bowel_Wall_Thickening and Free_fluid. Rows with critical missing values were dropped; others were filled using forward fill or most frequent values.
- **Outliers**: Detected through boxplots (especially in Height, Weight, BMI). Some extreme values were considered for removal.

3.2 Exploratory Data Analysis (EDA)

- **Distribution Analysis**: Variables like Age and BMI were visualized using histograms and KDE plots. A higher incidence of appendicitis was found in mid-teenage years.
- **Categorical Analysis**: Violin and box plots showed that children diagnosed with appendicitis had significantly higher Alvarado and Pediatric Appendicitis Scores.
- **Correlation Matrix**: Strong positive correlation found between outcome (Appendicitis diagnosis) and features such as:
    - Alvarado_Score
    - Paedriatic_Appendicitis_Score
    - Appendix_on_US
- **Sex Differences**: Males and females showed slightly different clinical score distributions.

3.3 Principal Component Analysis (PCA)

- **Purpose**: To reduce the feature space, identify major variance drivers, and assist in visualization.
- **Steps**:
    - Standardization of numeric data.
    - Covariance matrix creation.
    - Eigen decomposition to determine explained variance.
- **Scree Plot**: First 2–3 components explain the majority of the variance.
- **Insights from PCA**:
    - Features such as Alvarado_Score, Weight, Free_fluid, and Paedriatic_Appendicitis_Score contributed heavily to the first few principal components.
    - PCA plots helped visualize clear separation between positive and negative cases of appendicitis.

- **Outcome**: PCA confirmed that certain variables carry redundant information and dimensionality can be reduced without significant information loss.

4. Discussion

The analysis confirmed that predictive scores (Alvarado and Pediatric Appendicitis Score), along with imaging results (e.g., Appendix_on_US), are strong indicators of appendicitis in pediatric patients. The combination of EDA and PCA provided clarity on which features are most informative and which can be dropped for simplification.

Clinical relevance is high — such insights can help doctors make informed diagnostic decisions and reduce the need for invasive tests. By reducing dimensionality via PCA, we also prepared the dataset for future machine learning models, ensuring faster and more accurate performance.

5. References

- Dataset: Rogensburg Pediatric Appendicitis Dataset
- Libraries:
    - pandas for data manipulation
    - numpy for numerical operations
    - matplotlib, seaborn for visualization
    - sklearn.decomposition.PCA for Principal Component Analysis
- Academic Reference:
    - Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics.
    - EDA methodologies from Towards Data Science articles and Kaggle kernels