Software Requirements Specification

for

Data Pulse

Version 1.0

Prepared by Group # 50

Sir Syed University of Engineering & Technology

Supervised By: Mr. Sarfaraz Abdul Sattar Natha

3rd March 2024

Team

Member Name	Primary Responsibility
Abdul Moiz Chishti **(GL)** 2020F-BSE-022	Documenting, Development of Model Comparison module
Shaheer Khan Qureshi 2020F-BSE-002	Documenting, Research, and development of models Comparison Reports
Syed Abdul Aleem 2020F-BSE-042	Documenting, Development of reports Generation Module

Table of Contents

1. Introduction

1.1 Purpose

The project aims to develop an automated platform that simplifies the data profiling and machine learning model selection process, facilitating efficient extraction of insights from data and selection of the most suitable machine learning model for users' datasets.

1.2 Scope

The scope includes creating an end-to-end solution for automated data profiling, analysis, comparison of multiple machine learning models based on dataset characteristics, and provision for downloading the optimal trained model for deployment.

1.3 Definitions, Acronyms and Abbreviations

Data Profiling: Examination of datasets to understand their statistics and quality.
Machine Learning (ML): Algorithms that improve automatically through experience.
Model Selection: Choosing the best machine learning model for a given dataset.

1.4 Acronyms and Abbreviations

ML: Machine Learning
AI: Artificial Intelligence
UI: User Interface

2. Project Planning and Management

2.1 SWOT Analysis

Strengths:

User-friendly interface simplifying data analysis for non-technical users; rapid development and deployment of data applications; integration capabilities with Python's vast data science ecosystem.

Weaknesses:

Dependence on Python may limit adoption among users unfamiliar with it; potential performance issues with large datasets; reliance on external libraries for advanced analytics features.

Opportunities:

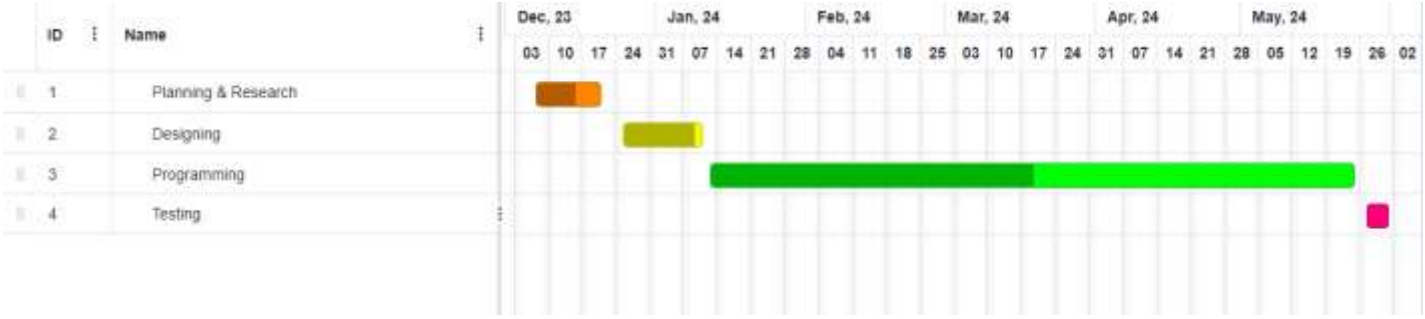
Growing demand for accessible data analysis tools; potential to expand into educational sectors and small businesses; opportunities for customization and extension by the community.

Threats:

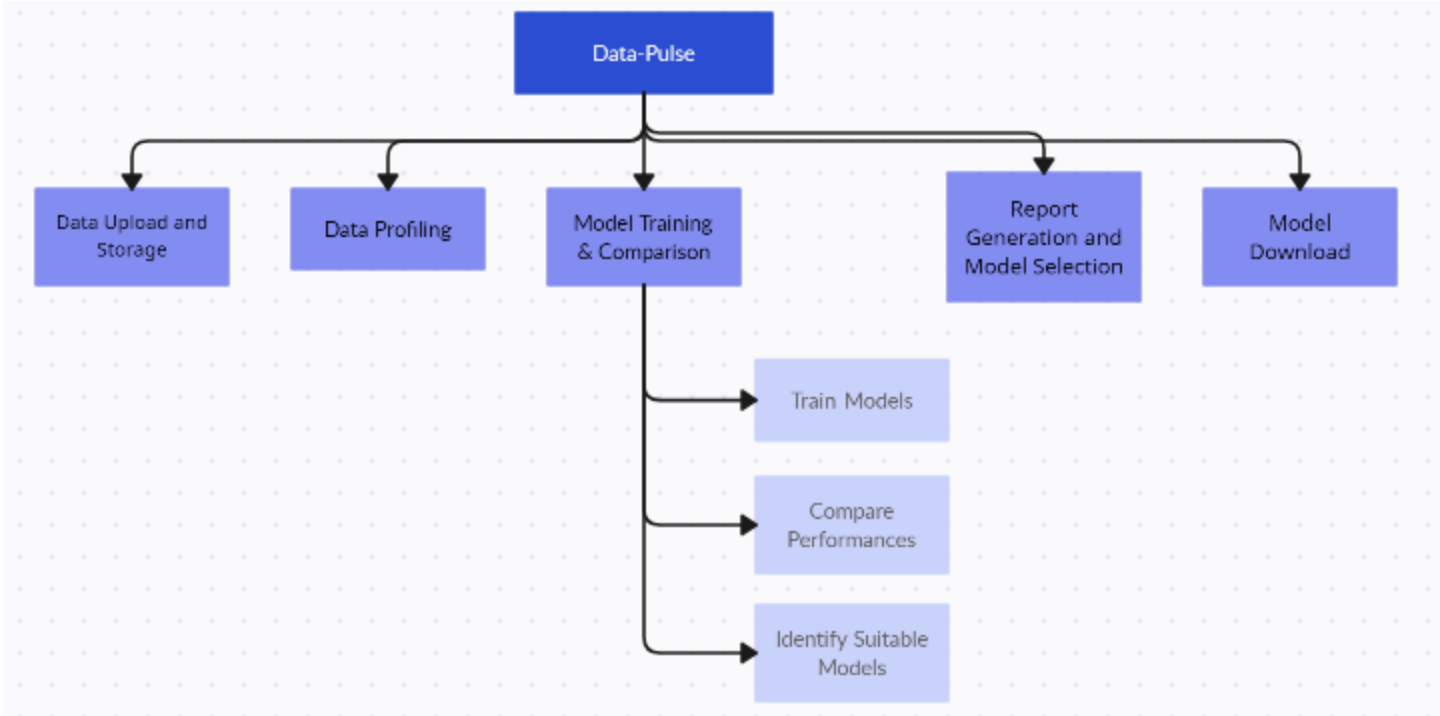
Competition from established data analysis platforms; changes in technology or user needs that could render the app obsolete; security vulnerabilities inherent in web applications.

This analysis highlights the app's potential to democratize data analysis while also pointing out the challenges it may face in a rapidly evolving tech landscape.

2.2 Gantt Chart



2.3 Work Breakdown Structure [WBS]


Data-Pulse
Data Upload and Storage
Data Profiling
Model Training & Comparison
Report Generation and Model Selection
Model Download
Train Models
Compare Performances
Identify Suitable Models

3. Overall Description

3.1 Product Perspective

An innovative system designed to modernize data analysis and machine learning model selection, integrating seamlessly with existing data management ecosystems to enhance decision-making and operational efficiency.

3.2 Product Function

The platform automates data profiling, provides comparative analysis of machine learning models based on specific data characteristics, and enables the download of the optimal model for direct application.

3.3 Operating Environment

The system is developed for compatibility across various operating systems and platforms, ensuring accessibility for a wide range of users.

3.4 Design and Implementation Constraints

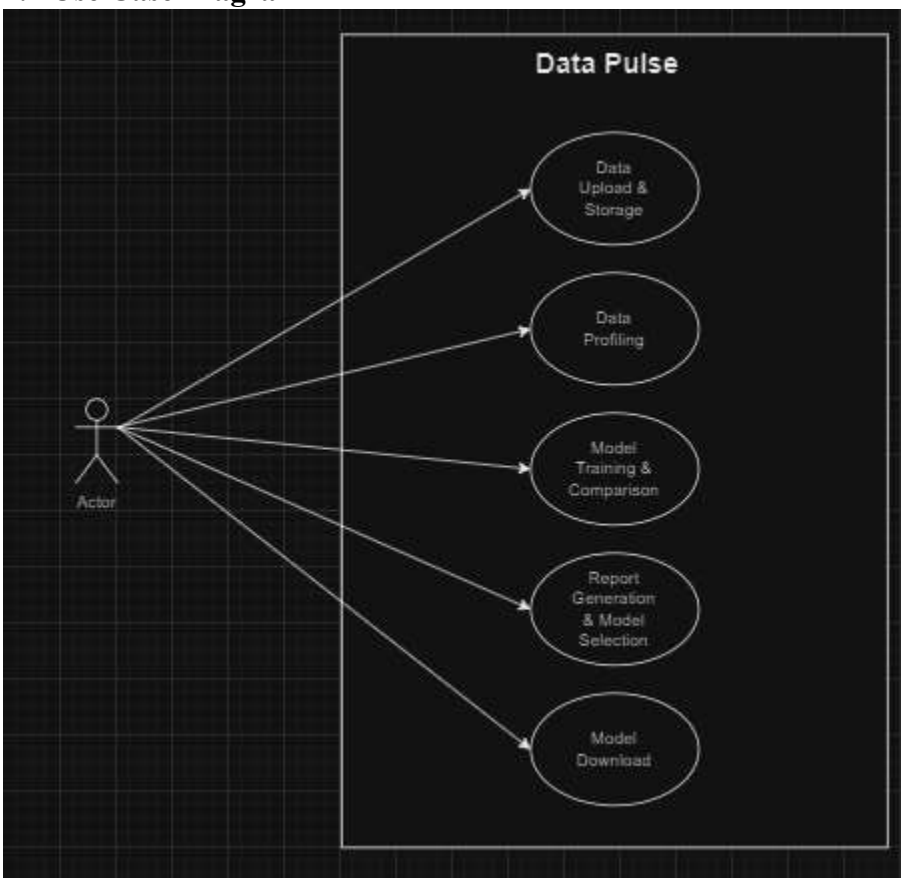
Challenges include handling diverse data types and sizes, ensuring the scalability of the system, and maintaining user-friendly interfaces while providing in-depth analysis capabilities.

3.5 Assumptions and Dependencies

The project assumes availability of the necessary technological infrastructure for development and deployment and depends on continuous updates to machine learning algorithms and data profiling techniques to remain effective.

4. Requirement Identifying Technique

4.1 Use Case Diagram



4.2 Use Case Description

The described Use Case Diagram for the "Automated Data Profiling and ML Model Comparison System" outlines a comprehensive system that facilitates several critical steps in machine learning (ML) model development and selection. This system is designed to automate processes from data upload to model deployment, making it a valuable tool for users looking to streamline their machine learning workflows. Below, each use case is described in detail:

Data Upload and Storage

Description: Users can upload their datasets through a web interface. This is the initial step in the machine learning workflow, where raw data is introduced into the system.
Actors Involved: Users.
Goal: To securely store datasets in the local storage system for further processing.

Data Profiling

Description: Once the data is uploaded, the system automatically profiles it. This process involves assessing the quality, structure, and various characteristics of the data that are crucial for effective model training and selection.
Actors Involved: Users (implicitly, as the profiling is automated).

Goal: To understand the dataset's attributes, such as missing values, data distribution, and potential data quality issues, which are essential for preprocessing and model training.

Model Training and Comparison

Description: The system trains multiple machine learning models using the profiled data. It then compares their performance based on predefined criteria, such as accuracy, precision, recall, or any other relevant metrics.
Actors Involved: Users (implicitly, as the training and comparison are automated).
Goal: To identify the most suitable model(s) that perform best on the given dataset according to the specified metrics.

Report Generation and Model Selection

Description: After comparing the models, the system generates a detailed report outlining each model's performance. This report helps users make informed decisions regarding the best model for their specific needs.
Actors Involved: Users.
Goal: To provide actionable insights into the performance of each trained model, facilitating an informed selection process.

Model Download

Description: Users can download the selected model for local deployment. This feature ensures that the integration of the machine learning model into various applications is straightforward, without the necessity for cloud services.
Actors Involved: Users.
Goal: To enable the easy and seamless deployment of the chosen model in the user's local environment or application, ensuring the model's benefits are readily accessible.

This architecture ensures that the system remains robust and scalable, whether deployed in a local or server-based environment. It provides a comprehensive solution for automating data profiling and machine learning model comparison, thus streamlining the process of developing and deploying machine learning models without relying on cloud storage.

5. Non- Functional Requirements

5.1 Performance Requirements

The system must handle datasets (up to 200 MB) efficiently, ensuring quick response times for data profiling and model comparison processes, supporting concurrent user sessions without degradation in performance.

5.2 Safety Requirements

The platform should incorporate error handling and validation mechanisms to prevent data loss or corruption. It must ensure the integrity of user data during processing and analysis.

5.3 Security Requirements

Data encryption in transit and at rest, user authentication, and authorization protocols are required to protect sensitive information and maintain privacy.

5.4 Software Quality Attributes

The system should be reliable, user-friendly, and adaptable, with a focus on scalability to handle growing data volumes and complexity.

5.5 Business Rules

The platform must adhere to data protection regulations and intellectual property laws, ensuring that data usage complies with legal standards.

5.6 Interoperability

It should be compatible with various data formats and external systems, allowing seamless integration and data exchange.

5.7 Extensibility

The design must accommodate future enhancements and integration of new machine learning models and data profiling techniques without significant overhauls.

5.8 Maintainability

Code should be well-documented and modular, simplifying updates, bug fixes, and customization.

5.9 Portability

The application should be deployable across different operating systems and cloud platforms without requiring major modifications.

5.10 Reusability

Components of the system, such as data processing modules and model evaluation algorithms, should be designed for reuse in other projects.

5.11 Installation

The installation process must be straightforward, with clear documentation for setting up the system in various environments.

6. Other Requirements

6.1 On-line User Documentation and help System Requirements

Our system would give online support to the user, by providing pdf manuals in soft forms, and notifications regarding our system as online assistance to the user.

6.2 Purchased Requirements

None

6.3 Licensing Requirements

None

6.4 Legal, copyright, and other Notices

Our project is not protected by copyright law.

6.5 Applicable Standards

- ISO/IEC 25010: For software quality requirements and evaluation (SQuaRE), covering usability, performance efficiency, compatibility, security, and more.
- ISO/IEC 27001: For information security management, ensuring data privacy, integrity, and availability.
- WCAG (Web Content Accessibility Guidelines): To ensure the app is accessible to users with disabilities.
- GDPR (General Data Protection Regulation): For apps used within or targeting users in the European Union, ensuring user data protection and privacy.

6.6 Reports (Feedback, Invoice, User, Usage, Balance Sheet, Executive Summary etc.)

In this section, we will take the review of our system by the user that whether our system is doing well or not.

7. References

https://arxiv.org/pdf/2012.12600.pdf
https://towardsdatascience.com/automating-scientific-data-analysis-part-1-c9979cd0817e
https://scripts.iucr.org/cgi-bin/paper?a56909
https://www.researchgate.net/publication/263671317_Automated_data_analysis
https://www.stitchdata.com/resources/automated-data-analytics/