Sri Lanka Institute of information Technology

Fundamentals of Data Mining
IT3051

Mini-Project : Statement of Work Document
2025

Group Details

Weekend : FDM_MLB_G01

IT Number	Name	Email	Contact Number
IT23183018	Hirusha D G A D	it23183018@my.sliit.lk	077 2424 521
IT23191006	Cooray Y H	it23191006@my.sliit.lk	070 6080 877
IT23194908	Harischandra R A S R	it23194908@my.sliit.lk	076 2352 603
IT23224384	Hasangani W R A	it23224384@my.sliit.lk	077 1227 898

Submitted On : 2025-09-14

Table of Contents

Background

In today's world, the prevalence of genetic disorders has become a pressing concern in the healthcare sector. Genetic disorders arise primarily due to mutations in DNA, chromosomal abnormalities, or inherited genetic traits, and they often result in severe health complications ranging from lifelong disabilities to fatal conditions. Unlike other illnesses that may be prevented or cured through lifestyle modifications or medication, genetic disorders pose unique challenges as they are deeply rooted in an individual's biological makeup. With global population growth, lifestyle-related risk factors, and increasing environmental exposures, studies indicate a rising trend in the number of patients diagnosed with genetic disorders across the globe.

One of the greatest challenges in addressing genetic disorders is the lack of **early detection and awareness**. Many individuals remain undiagnosed until symptoms have progressed to advanced stages, where treatment options are either limited, invasive, or less effective. Such delays not only reduce the patient's quality of life but also result in increased psychological and financial burdens for families. On a larger scale, healthcare systems are impacted heavily due to long-term treatment requirements, specialized medical care, and rising costs associated with managing these conditions. The absence of accessible and affordable genetic testing in certain regions further aggravates this issue, leaving many patients undiagnosed until it is too late.

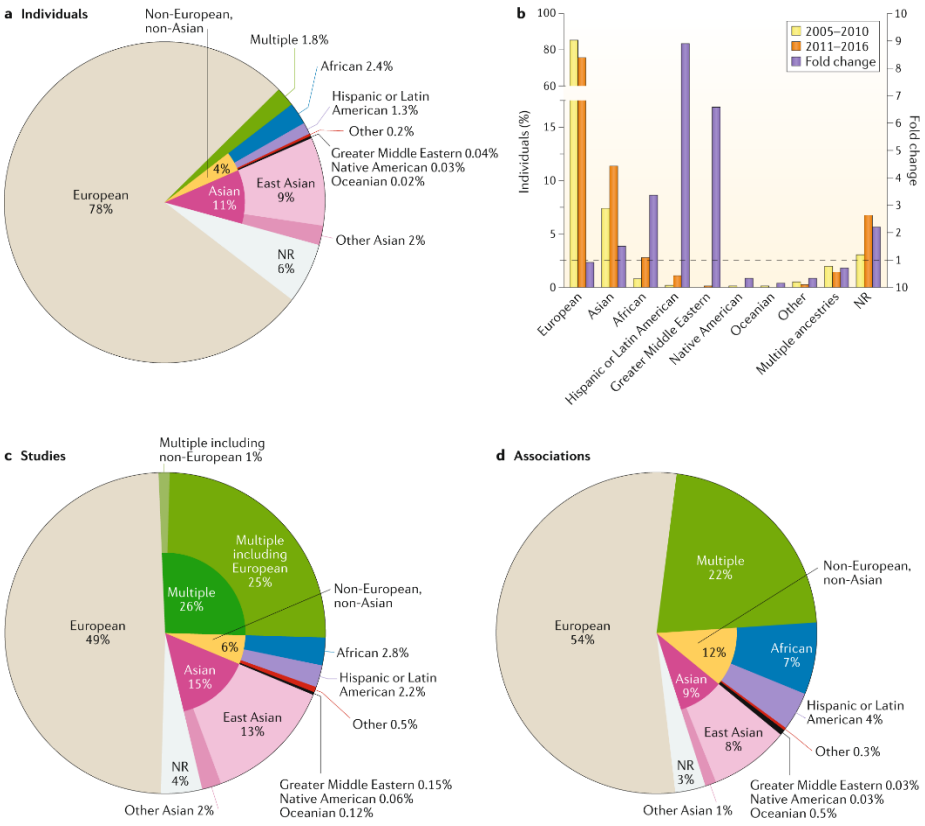


Figure 1 - Genomics of disease risk in globally diverse populations

In recent years, technological advancements in **data mining and machine learning** have provided promising avenues to address these challenges. By analyzing large volumes of medical, genetic, and demographic data, it becomes possible to identify hidden patterns and correlations that are often overlooked in traditional clinical diagnosis. Machine learning models can be trained to predict the likelihood of a genetic disorder, classify its type, and even provide supporting evidence for decision-making. These predictive systems empower medical professionals by enabling them to recommend timely genetic testing, suggest preventive measures, and design personalized treatment plans tailored to each individual's unique genetic profile.

Beyond supporting clinicians, such systems also contribute to broader societal benefits. Families can make informed decisions about healthcare and lifestyle, governments and health organizations can allocate resources more efficiently, and overall healthcare expenditure can be reduced by focusing on preventive measures rather than prolonged treatments. **Moreover, in countries like Sri Lanka, where genetic awareness is still growing, predictive systems can serve as educational and awareness tools that encourage proactive healthcare practices.**

With this context, our project aims to design and develop a **Genetic Disorder Predicting System** that leverages advanced machine learning algorithms to classify and predict genetic disorders based on patient data. The proposed solution is intended not only to assist clinicians in early detection but also to provide an accessible platform for healthcare institutions and families. By integrating predictive analytics into medical practice, the system seeks to minimize the risks associated with late detection, reduce healthcare costs, and most importantly, improve patient outcomes and quality of life.

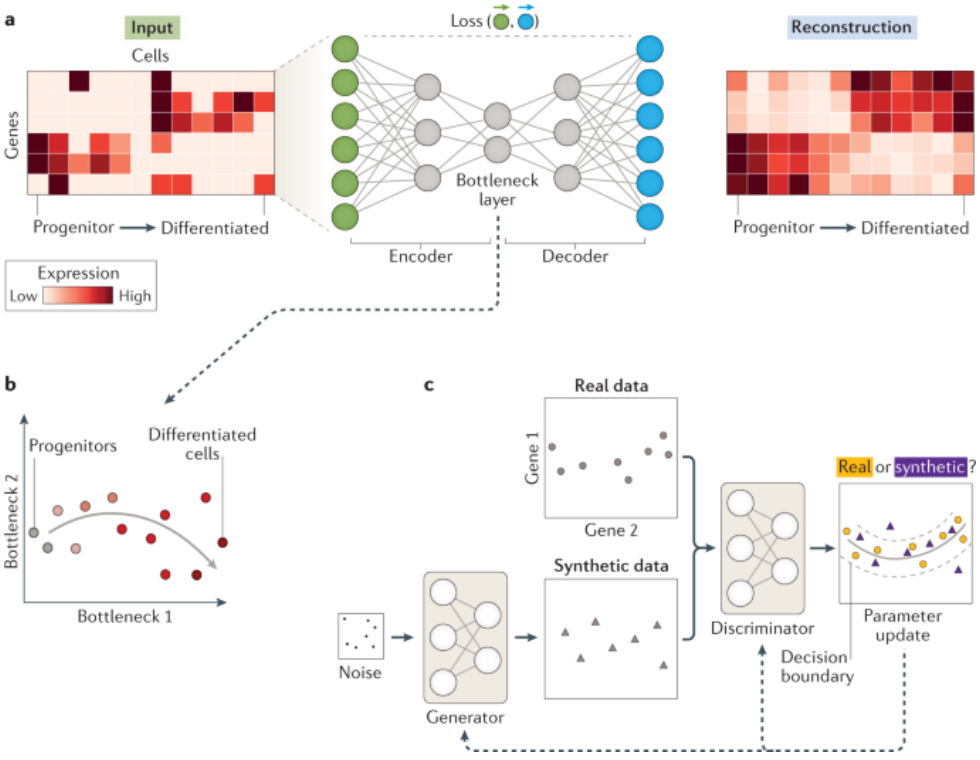


Figure 2 - Obtaining genetics insights from deep learning via explainable artificial intelligence

- **Problem** – The increasing prevalence of genetic disorders poses a major challenge in healthcare. Many patients remain undiagnosed until late stages, leading to reduced treatment effectiveness, higher healthcare costs, and a lower quality of life. The lack of early detection tools makes it difficult for clinicians to identify at-risk patients and recommend timely preventive actions. There is a pressing need for a predictive system that can analyze patient data and provide reliable insights into potential genetic disorders.

- **Client** – Healthcare institutions, genetic testing centers, and clinicians who aim to improve early detection and prevention of genetic disorders. These stakeholders face challenges in managing rising cases of genetic illnesses and require data-driven solutions to better allocate resources, guide medical decisions, and reduce the impact of late diagnosis on patients and families.

- **Solution** – Develop a **Genetic Disorder Predicting System** that applies data mining and machine learning techniques to classify and predict the likelihood of genetic disorders. By processing patient demographic, clinical, and genetic information, the system will provide accurate predictions of both disorder categories and sub-classes. This solution will serve as a supportive decision-making tool for medical professionals.

- **Goal** – Create a machine learning model that can reliably predict the presence and type of a genetic disorder for a given patient. The ultimate goal is to enable healthcare providers to take preventive measures, recommend further testing, and design early intervention plans. By doing so, the system will help reduce healthcare costs, improve patient survival rates, and enhance overall quality of care.

- **Dataset (Selected and Approved)** – For this project, we have selected the publicly available **"Of Genomes and Genetics" dataset** from Kaggle. The dataset contains detailed patient records, including demographic attributes, medical test results, symptoms, and genetic information, along with the target labels **`Genetic Disorder`** and **`Disorder Subclass`**.

  OF Genomes and Genetics from Kaggle

Scope of Work

This project consists of five main layers, which are essential to designing and developing the Genetic Disorder Prediction System:

1. User Interface Layer

This layer represents the **frontend interface** where end-users, such as clinicians, genetic counselors, or healthcare staff, will interact with the system. The UI will allow users to input patient information including demographic details, test results, and genetic history. The interface will be designed with simplicity and clarity to ensure user-friendliness, even for non-technical users. A questionnaire-style form will be implemented for easy data entry, along with options to view predictive outcomes and reports.

2. Data Wrangling and Data Cleansing Layer

The dataset contains a large number of attributes, many of which have missing, inconsistent, or irrelevant values. This layer will handle preprocessing tasks such as:

- Removing identifiers that do not contribute to prediction (e.g., Patient ID, names, institute name).
- Handling missing values through imputation or removal.
- Encoding categorical data into numerical formats suitable for machine learning algorithms.
- Normalizing and scaling test results to maintain consistency across different attributes.
  This ensures that the input data is reliable, accurate, and suitable for further analysis.

3. Data Mining Layer

This layer focuses on extracting **patterns and insights** from the dataset using data mining techniques. It involves:

- Identifying correlations between symptoms, genetic traits, and disorders.
- Applying **association rule mining** (Apriori, FP-Growth) to uncover hidden relationships, such as risk factors linked to specific disorders.
- Performing exploratory analysis to determine which attributes have the strongest influence on the target outcome.
  This stage builds the foundation for predictive modeling by generating meaningful knowledge from raw data.

4. Model Building and Analysis Layer

In this layer, machine learning algorithms will be applied to build predictive models. The project will train and evaluate classification models (e.g., Decision Tree, Random Forest, Logistic Regression, Naïve Bayes, Neural Networks) to predict both the **Genetic Disorder (category)** and **Disorder Subclass (specific disease type)**.

Additionally, **clustering methods (KMeans)** will be explored to group patients based on their test results and symptoms, providing unsupervised insights. Each model will be analyzed using performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix, ensuring the most suitable model is selected for deployment.

5. Data Visualization Layer

The visualization layer is responsible for presenting results in a clear and interpretable format. Users will be able to view:

- Predicted disorder categories and subclasses.
- Clustered patient groups based on symptoms and tests.
- Association rules highlighting risk patterns.
- Graphical dashboards displaying patient-level predictions, trends, and model confidence. This layer enhances decision-making by allowing healthcare providers to quickly interpret the system's outputs and take appropriate medical actions.

Activities

- **Identify the Real-World Problem & Confirm Dataset Fit**

Select the Kaggle "Of Genomes and Genetics – HackerEarth ML" dataset and validate that it supports the project's aims (predicting both the broad `Genetic Disorder` and the specific `Disorder Subclass`). Clearly define success criteria (e.g., F1-score per class, clinical interpretability) and outline stakeholders (clinicians, genetic counselors).

- **Data Preparation, Model Construction, and Training**

1. **Data Audit**: profile missingness, duplicates, and leakage; flag personally identifiable fields.
2. **Cleaning & Preprocessing**:
    - Drop identifiers that do not add predictive power (e.g., Patient Id, names, institute).
    - Impute missing values (median/most-frequent/"Unknown"), encode categoricals (one-hot/ordinal), and scale numeric lab tests for algorithms like SVM/KMeans.
    - Address **class imbalance** across subclasses via class weights/SMOTE.
    - Split data into train/validation/test folds.

3. **Feature Engineering**: symptom counts, grouped risk factors (e.g., IVF exposure, maternal illness), and interaction features among tests/symptoms.
4. **Training**: develop baseline and advanced models (e.g., Logistic Regression, Random Forest, XGBoost/LightGBM, MLP), with hyperparameter search and cross-validation.

- **Knowledge Discovery (Unsupervised & Pattern Mining)**

In parallel with supervised modeling:

- **Clustering (KMeans)** on standardized test results and symptoms to reveal natural patient groupings and compare with labeled disorders.
- **Association Rule Mining** (Apriori/FP-Growth) on categorical factors (maternal illness, exposure, symptoms) to extract interpretable rules that can support clinician decisions (e.g., "{Radiation exposure, Symptom_3} ⇒ {Mitochondrial}")

- **Evaluate the Model**

Use stratified validation and a held-out test set. Report accuracy, macro/micro F1, class-wise precision/recall, confusion matrices (both for `Genetic Disorder` and `Disorder Subclass`), ROC-AUC (one-vs-rest), and calibration. Perform error analysis (misclassified subclasses), ablation studies (feature groups), and explainability (feature importance/SHAP)

- **Make Predictions & Generate Clinician-Friendly Outputs**

Produce patient-level predictions for both targets with confidence scores, highlight top contributing factors/rules, and categorize risk (High/Medium/Low) for triage.

- **Front-end Development & Deployment**

Build a simple web app where users can:

- Enter patient details via a guided questionnaire.
- View predicted disorder & subclass with explanations.
- Explore clusters (visualized) and relevant association rules.
- Download a brief "clinical note" PDF.
  Frontend (React or Streamlit pages) with a Python API (FastAPI/Streamlit)

Approach

Our project begins with the careful selection of a suitable dataset. We chose the publicly available **"Of Genomes and Genetics – HackerEarth ML"** dataset from Kaggle, as it provides a comprehensive set of patient records containing demographic details, genetic information, medical histories, and test results, along with target labels for both **Genetic Disorder** and **Disorder Subclass**.

After examining the dataset, the following stepwise approach will be implemented:

1. Data Preprocessing

- **Feature Selection and Dimensionality Reduction**: Remove identifiers (e.g., Patient ID, names, institute names) and irrelevant attributes that do not contribute to predictions.
- **Handling missing values**: Apply appropriate imputation methods (mean/median for numerical, mode/"Unknown" for categorical).
- **Encoding Ctegorical Features**: Convert attributes like maternal/paternal genes, symptoms, and consent fields into machine-readable formats using one-hot or label encoding.
- **Normalization and Scaling**: Standardize test results and numerical features to ensure consistency, particularly for clustering algorithms.
- **Dataset Splitting**: Divide into training, validation, and test subsets for reliable model assessment.

2. Model Building

- **Classification Models**: Build and train multiple models such as Decision Tree, Random Forest, Logistic Regression, Naïve Bayes, and Multi-layer Perceptron (MLP). These models will predict both the broader genetic disorder categories and finer disorder subclasses.
- **Clustering Models**: Apply KMeans to uncover hidden patterns among patients based on symptoms and test results, comparing unsupervised groupings with labeled outcomes.
- **Association Rule Mining**: Use Apriori and FP-Growth algorithms to discover rules connecting maternal health factors, exposures, and symptoms with specific disorders.

3. Evaluation and Comparison

- Evaluate classification models using accuracy, precision, recall, F1-score, and confusion matrices.
- Compare models across both the Genetic Disorder and Disorder Subclass predictions to identify the most reliable approach.
- Assess clustering by measuring cluster quality (silhouette score) and comparing against known disorder categories.
- Validate association rules based on support, confidence, and lift to ensure practical interpretability.

4. Model Selection and Optimization

- Select the best-performing model(s) based on balanced accuracy and interpretability.
- Tune hyperparameters using cross-validation techniques.
- Apply methods to handle class imbalance (SMOTE, class weighting) to ensure fair predictions across all disorder subclasses.

5. Interface and System Integration

- Integrate the final model into a web application to provide predictions in a user-friendly format.
- The interface will allow users to input patient data via a questionnaire, and in return, receive predictions, confidence scores, visualizations of clusters, and explanatory association rules.
- Backend services will be developed using Python frameworks (Flask/Streamlit/FastAPI), while the frontend will be built with React or Streamlit for simplicity.

Deliverables

The primary objective of this project is to design and implement a predictive system capable of identifying genetic disorders at both the category and subclass levels. The following deliverables will be produced as part of the project outcome:

1. **Cleaned and Processed Dataset**
    - A refined dataset created from the original Kaggle source, with missing values handled, irrelevant attributes removed, and categorical and numerical features preprocessed for machine learning tasks.
2. **Exploratory Data Analysis (EDA) Report**
    - A comprehensive report summarizing the dataset characteristics, including class distributions, correlations, missing data analysis, and feature importance.
    - Visualizations (graphs and charts) to illustrate patterns, trends, and potential risk factors for genetic disorders.
3. **Predictive Machine Learning Models**
    - Classification models trained to predict both the broader *Genetic Disorder* categories and the finer *Disorder Subclasses*.
    - Clustering models (e.g., KMeans) to identify patient groupings based on symptoms and test results.
    - Association rule mining outputs (Apriori/FP-Growth) to highlight interpretable patterns and risk factors.
4. **Model Evaluation and Comparison**
    - Documentation of model performance metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC curves.
    - Selection of the most suitable model(s) for deployment, supported by comparative analysis.
5. **User-Friendly Web Application**
    - A functional web interface that allows healthcare professionals to input patient details and receive predictive outputs.
    - Features will include:
        - Predictions of disorder category and subclass with confidence scores.
        - Visualization dashboards showing patient clusters and discovered association rules.
        - Downloadable reports summarizing patient-level predictions.
6. **Final Project Report and Documentation**
    - A detailed written report covering background, scope, methodology, results, and conclusions.
    - Documentation of preprocessing steps, algorithms applied, evaluation results, and system design.
    - User guide for operating the web application.

Assumptions

To ensure the successful development and deployment of the Genetic Disorder Prediction System, the following assumptions have been made:

1. **Data Quality Assumption**

    - The publicly available Kaggle dataset used for this project is assumed to be accurate, complete, and representative of real-world patient data.
    - The records provided are expected to reflect valid genetic and clinical patterns that can support model training.

2. **Independence Assumption**

    - Each patient record is assumed to be independent of others, meaning no duplicate or linked entries that might bias the model's learning process.

3. **No Hidden Variables Assumption**

    - It is assumed that all necessary and relevant features required to predict genetic disorders are present in the dataset, and no critical variables are missing or hidden.

4. **Consistency of Data Distribution**

    - The training and testing datasets are assumed to follow a similar distribution, ensuring that models trained on the training set can generalize well to unseen patient records.

5. **Feature Relevance Assumption**

    - All selected features (demographics, test results, symptoms, and genetic information) are assumed to hold relevance to predicting genetic disorders, even if their contribution varies in significance.

6. **Noise in Data Assumption**

    - The dataset may contain noisy or inconsistent entries; however, it is assumed that preprocessing and data cleansing will adequately minimize these effects, allowing machine learning models to extract useful information.

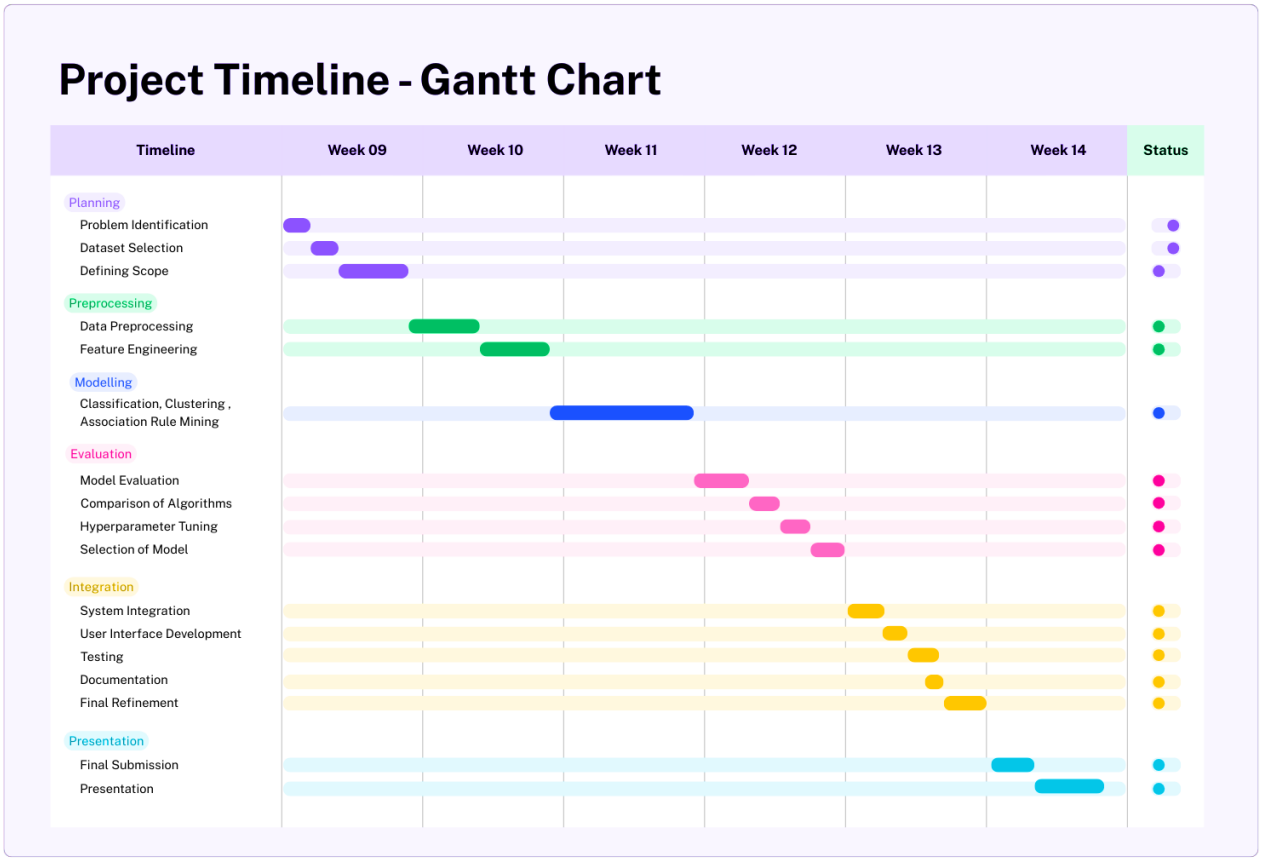
7. **User Input Assumption**

    - In the deployed system, it is assumed that healthcare professionals will provide correct and complete input when entering patient data into the user interface for prediction.

8. **Ethical Usage Assumption**

   - The developed system will be used strictly as a decision-support tool by healthcare institutions and not as a replacement for certified medical diagnosis or professional consultation.

Project Plan & Timeline



Project Team, Roles and Responsibilities

Member IT Number	Member Name	Member Role	Member Responsibilities
IT23183018	Hirusha D G A D	Team Leader Solution Developer Data Analyst ML Ops Engineer	✓ Coordinate team activities and ensure timely progress. ✓ Lead dataset exploration and requirement analysis. ✓ Design system architecture and taking finalized decisions. ✓ Oversee the implementation of core machine learning models and data pre-processing activities. ✓ Contribute to documentation and report preparation. ✓ Guide integration of models into the system.
IT23191006	Cooray Y H	Solution Developer Data Engineer	✓ Handle data preprocessing, cleaning, and transformation. ✓ Implement feature engineering and dimensionality reduction. ✓ Support training and optimization of classification and clustering models. ✓ Manage dataset handling and ensure reproducibility.
IT23194908	Harishchandra R A S R	Solution Developer UI-UX Developer	✓ Design and develop the user-friendly web interface. ✓ Ensure seamless interaction between frontend and backend. ✓ Implement visual dashboards for predictions, clusters, and association rules. ✓ Assist in system integration and testing.

IT23224384	Hasangani W R A	Solution Tester Documentation Lead	✓ Conduct testing and validation of predictive models. ✓ Evaluate model performance and compare results. ✓ Ensure quality assurance of the final web application. ✓ Lead preparation of project documentation and user guide.