Predicting Genetic Disorders in India Using Machine Learning and Geospatial Analysis

1. Introduction

This project aims to develop a machine learning model to predict genetic disorders based on family history, medical data, and geographical factors. The study will integrate existing dataset attributes with additional location-based data acquired through RTI requests. The objective is to analyze disorder prevalence across different regions in India and identify high-risk zones.

2. Objectives

Develop a machine learning model to predict genetic disorders.

Integrate geographical data to analyze regional disorder prevalence.

Identify high-risk zones for genetic disorders.

Deploy the model as an API or web tool for research and policy use.

3. Methodology

Data Collection – Use an existing dataset and acquire additional geographic data through RTI requests.

Data Preprocessing – Handle missing values, encode categorical variables, and scale numerical features.

Feature Engineering – Select relevant features and derive new ones for improved predictions.

Model Development – Train and evaluate machine learning models (Random Forest, XGBoost, SVM, etc.).

Geospatial Analysis – Integrate geographic attributes and analyze disorder distribution across regions.

Model Optimization – Tune hyperparameters and address class imbalance.

Deployment – Develop an API or web-based tool for real-time predictions.

4. Expected Outcomes

A machine learning model capable of predicting genetic disorders with high accuracy.

Insights into regional trends and high-risk zones for genetic disorders.

A publicly accessible API or web tool for predictions and analysis.

Research findings that can aid policymakers and medical professionals.

5. Timeline

Weeks 1-2: Data exploration, cleaning, and preprocessing.

Weeks 3-4: Feature engineering, model selection, and baseline model training.

Weeks 5-6: RTI request for geographic data and data integration.

Weeks 7-8: Advanced modeling with geographic attributes and zonal analysis.

Weeks 9-10: Model deployment, documentation, and final reporting.

6. Tools and Technologies

Programming & Data Processing: Python, Pandas, NumPy, Scikit-learn

Machine Learning Models: Random Forest, XGBoost, SVM, Logistic Regression

Geospatial Analysis: Geopandas, Folium, QGIS

Model Deployment: Flask, FastAPI, Streamlit

Visualization & Reporting: Matplotlib, Seaborn, Plotly, Jupyter Notebook

7. Challenges and Mitigation Strategies

Data Availability: Acquire additional geographic data via RTI requests and open datasets.

Class Imbalance: Use techniques like SMOTE and class-weighted models.

Feature Selection: Apply SHAP analysis and domain expert validation.

Model Overfitting: Implement cross-validation and regularization techniques.

Deployment Complexity: Use lightweight frameworks (Flask, FastAPI) for easy integration.