Parshvanath Charitable Trust's

A. P. SHAH INSTITUTE OF TECHNOLOGY, THANE

(All Programs Accredited by NBA)

Department of Information Technology

Genetic Disorder Prediction

Shreya Mahajan 20104001

Contents

- Introduction
- Objectives
- Scope
- Summarizing the dataset.
- Visualizing the dataset.
- Algorithms Details
- Result
- References

1. Introduction

Problem Identified:

- Genetic Disorder Classification: Accurate classification and diagnosis of genetic disorders are critical for patient care, but it can be challenging due to the complexity and diversity of genetic disorders.
- Data Preprocessing: Genetic disorder data is often noisy, incomplete, and inconsistent, making it difficult to apply machine learning models effectively.

Solution Proposed:

- Genetic Disorder Classification Model: We propose building a machine learning model that can classify genetic disorders based on patient data, including genetic, clinical, and demographic information.
- Predictive Analytics: The project aims to provide accurate predictions of genetic disorders, which can assist healthcare professionals in early diagnosis and personalized treatment.
- User-Friendly Interface: We plan to create a user-friendly web interface where medical practitioners can input patient data, and the system will provide predictions and recommendations.

2. Objectives

1. To develop a machine learning model for accurate classification of various genetic disorders based on patient data.
2. To create a data preprocessing pipeline that addresses data quality issues, including missing value imputation, categorical data handling, and feature engineering.
3. To provide a user-friendly web interface for medical practitioners to input patient data and receive predictions and recommendations for genetic disorders.
4. To ensure scalability and generalisability of the model to handle a wide range of genetic disorders.
5. To address ethical considerations related to patient data privacy and consent, ensuring responsible data usage.
6. To impact healthcare by improving the accuracy and efficiency of genetic disorder diagnosis, leading to better patient outcomes and advancements in genetic medicine.

3. Scope

1. Can be applied in clinical settings to predict the status of individuals with a wide range of genetic disorders, ensuring timely and accurate diagnoses.
2. Can be used by healthcare professionals, including doctors, genetic counselors, and nurses, to enhance patient care and decision-making.
3. Can serve as an educational tool for medical students, genetic counselors, and researchers to understand the complexities of genetic disorders and their diagnoses.
4. Can be beneficial for patients and families, providing them with a clearer understanding of their genetic health and enabling them to make informed decisions regarding family planning and genetic testing.
5. Can aid in rare disease diagnosis, where accurate and early detection is crucial due to the limited availability of treatments.

4. Summarizing the dataset

- Link: https://www.kaggle.com/datasets/aryarishabh/of-genomes-and-genetics-hackerearth-ml-challenge?rvi=1

The dataset folder contains the following files:

train.csv: 22083 x 45

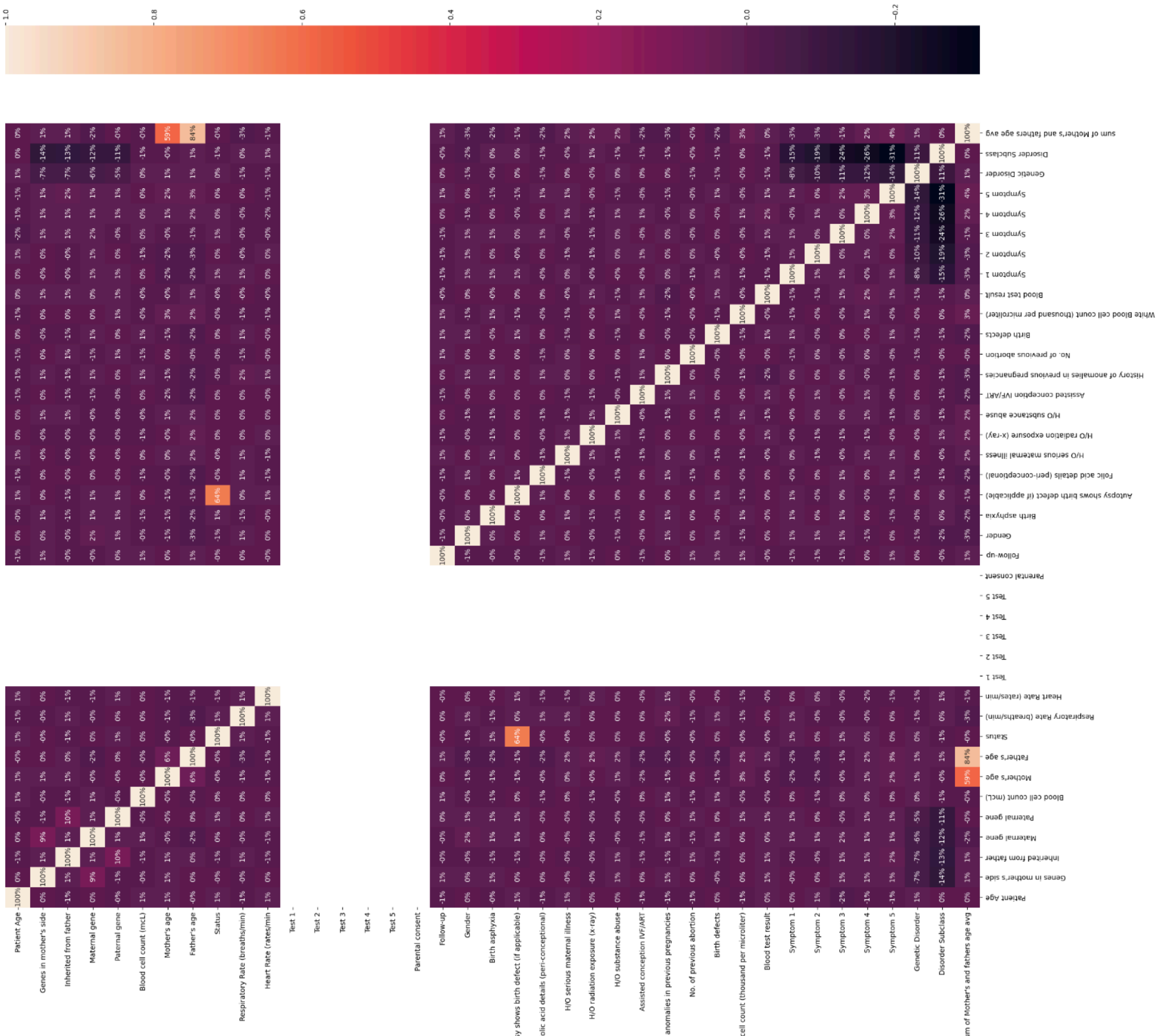
test.csv: 9465 x 43

The dataset "Of Genomes and Genetics - Hackerearth ML Challenge" available on Kaggle provides a comprehensive collection of genetic and clinical data, making it a valuable resource for genetic disorder prediction and analysis.

Features: The dataset includes a diverse set of features, such as patient age, gender, genetic information, clinical test results, family history, and more.

Target Variable: The primary target variable in this dataset is likely related to the genetic disorder status, indicating whether a patient has a genetic disorder or not.

5. Visualizing the dataset.



6. Algorithm

K-Nearest Neighbors (KNN) is a simple yet effective machine learning algorithm used for classification and regression tasks.

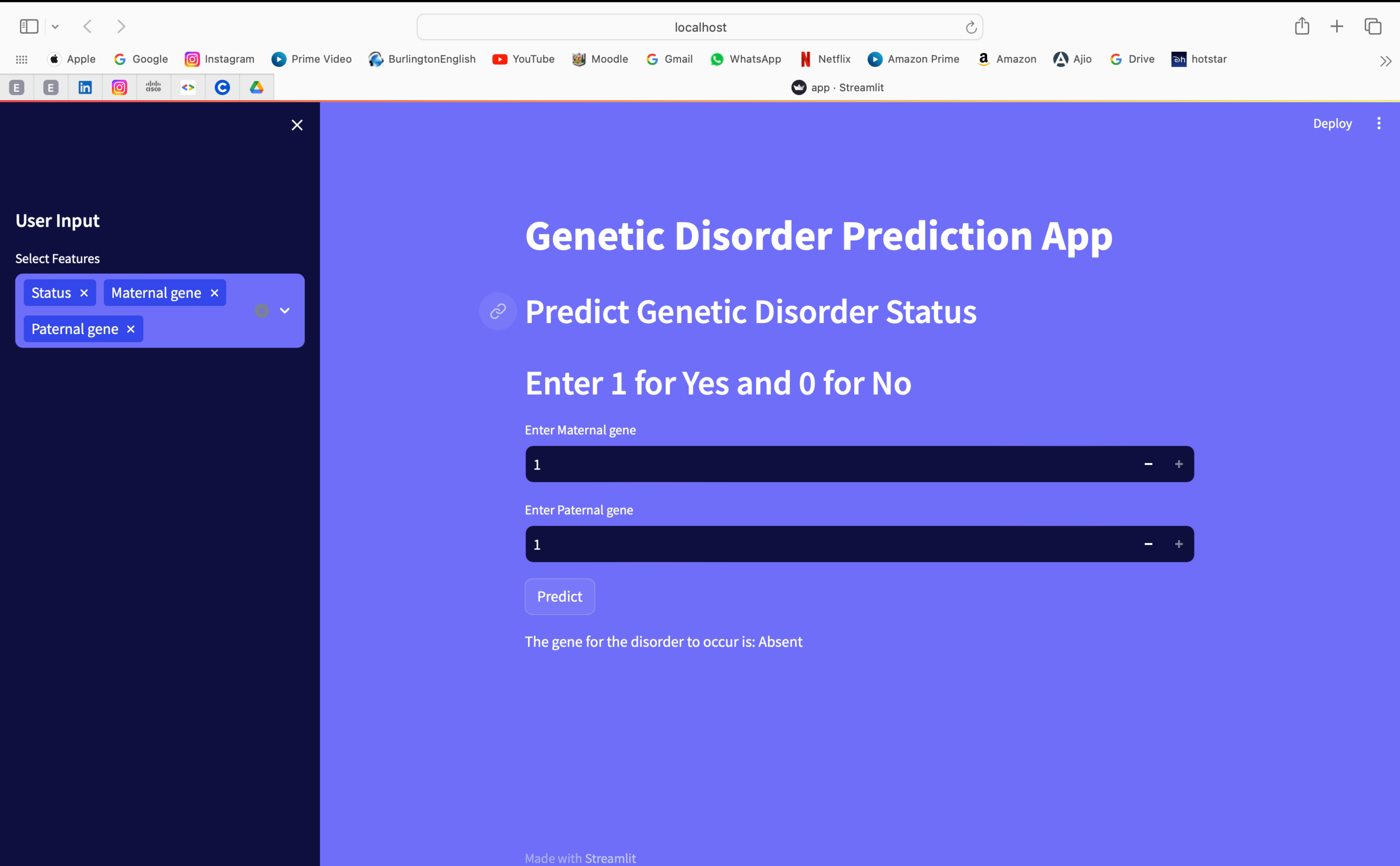
In KNN, the prediction for a given data point is determined by the class (in classification) or the value (in regression) of its nearest neighbors in the training dataset.

The "K" represents the number of nearest neighbors considered, and it's a hyperparameter chosen by the user. KNN relies on distance metrics, often Euclidean distance, to measure the proximity between data points.

When making a prediction, the algorithm selects the majority class (in classification) or calculates the mean value (in regression) of the K nearest data points. KNN is non-parametric and instance-based, meaning it doesn't build an explicit model but stores the training data for predictions.

It's straightforward to implement and interpret, making it a valuable tool in various machine learning applications.

7.Result



User Input

Select Features

Status × Maternal gene × Paternal gene ×

Genetic Disorder Prediction App

Predict Genetic Disorder Status

Enter 1 for Yes and 0 for No

Enter Maternal gene: 1

Enter Paternal gene: 1

Predict

The gene for the disorder to occur is: Absent

Made with Streamlit

Thank You...!!