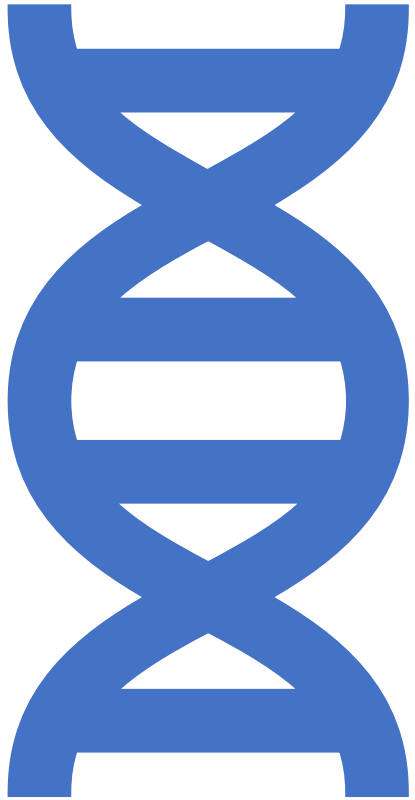


Genetic Disorder Prediction

By

Priyadarshani Kamble

Problem Statement:

- As per reports, because of the unsustainable increase in population and a lack of access to adequate health care, food, and shelter, the number of genetic disorder ailments have increased.
- Hereditary illnesses are becoming more common due to a lack of understanding about the need for genetic testing.
- Genetic disorders can cause serious health conditions and congenital disabilities in the baby. Studies suggest that chromosomal abnormalities occur in about 1 of 200 live births.
- Often kids die because of these illnesses, thus genetic testing during pregnancy is critical.

Context :
Genetic disorders occur when a mutation affects your genes. There are many types of Genetic disorders.

Single-gene inheritance diseases:

- Single gene inheritance diseases are diseases that occur because one defective gene is present. They are known as monogenetic disorders.

Multifactorial genetic inheritance disorders:

- Multifactorial conditions tend to run in families. This is because they are partly caused by genes. Your risk for a multifactorial trait or condition depends on how close you are to a family member with the trait or condition.

Mitochondrial genetic inheritance disorders:

- Mitochondrial genetic inheritance disorders are caused by mutations in the DNA of mitochondria, small particles within cells. This DNA is unique in that it is not located on the chromosomes in the cell nucleus. Mitochondrial DNA is always inherited from the female parent since egg cells (unlike sperm cells) keep their mitochondrial DNA during the process of fertilization.

Chromosome abnormalities:

- Chromosome abnormalities usually result from a problem with cell division and arise because of duplications or absences of entire chromosomes or pieces of chromosomes.

Genetic Disorder Classification

Mitochondrial genetic inheritance disorders

- Hereditary optic atrophy
- Barth syndrome
- Co-enzyme Q10 deficiency
- Myoclonic epilepsy with ragged red fibers (MERRF)
- MELAS syndrome, a rare form of dementia
- Kearns-Sayre syndrome
- Pearson syndrome
- Neuropathy, ataxia, retinitis pigmentosa (NARP)

Single-gene inheritance diseases

- Cystic fibrosis
- Sickle-cell anemia
- Polycystic kidney disease types 1 and 2
- Tay-Sachs disease etc.

Multifactorial genetic inheritance disorders

- Cancers of the breast, ovaries, bowel, prostate, and skin
- High blood pressure and high cholesterol
- Diabetes
- Alzheimer disease
- Schizophrenia
- Bipolar disorder
- Arthritis
- Osteoporosis

Chromosome abnormalities

- Down syndrome
- Cri-du-chat syndrome
- Klinefelter syndrome
- Patau syndrome (trisomy 13)
- Edwards syndrome (trisomy 18)
- Turner syndrome

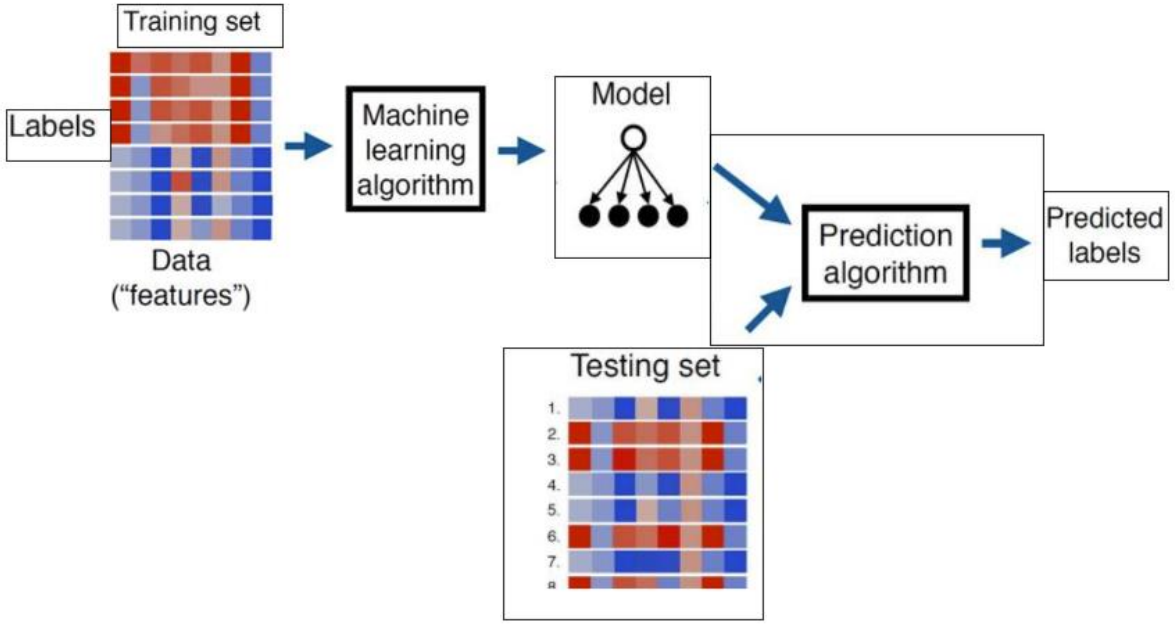
Machine Learning

Machine learning methods have been applied to a huge variety of problems in genomics and genetics.

A machine learning algorithm is provided with a dataset with possible genetic disorders.

The algorithm processes this labeled data and stores a model

The model learns and predict labels for each record. If the learning was successful, then all or most of the predicted labels will be correct



The Data

We will analyze the dataset provided on Kaggle to predict the genetic disorder.
dataset contains medical information about children who have genetic disorders.

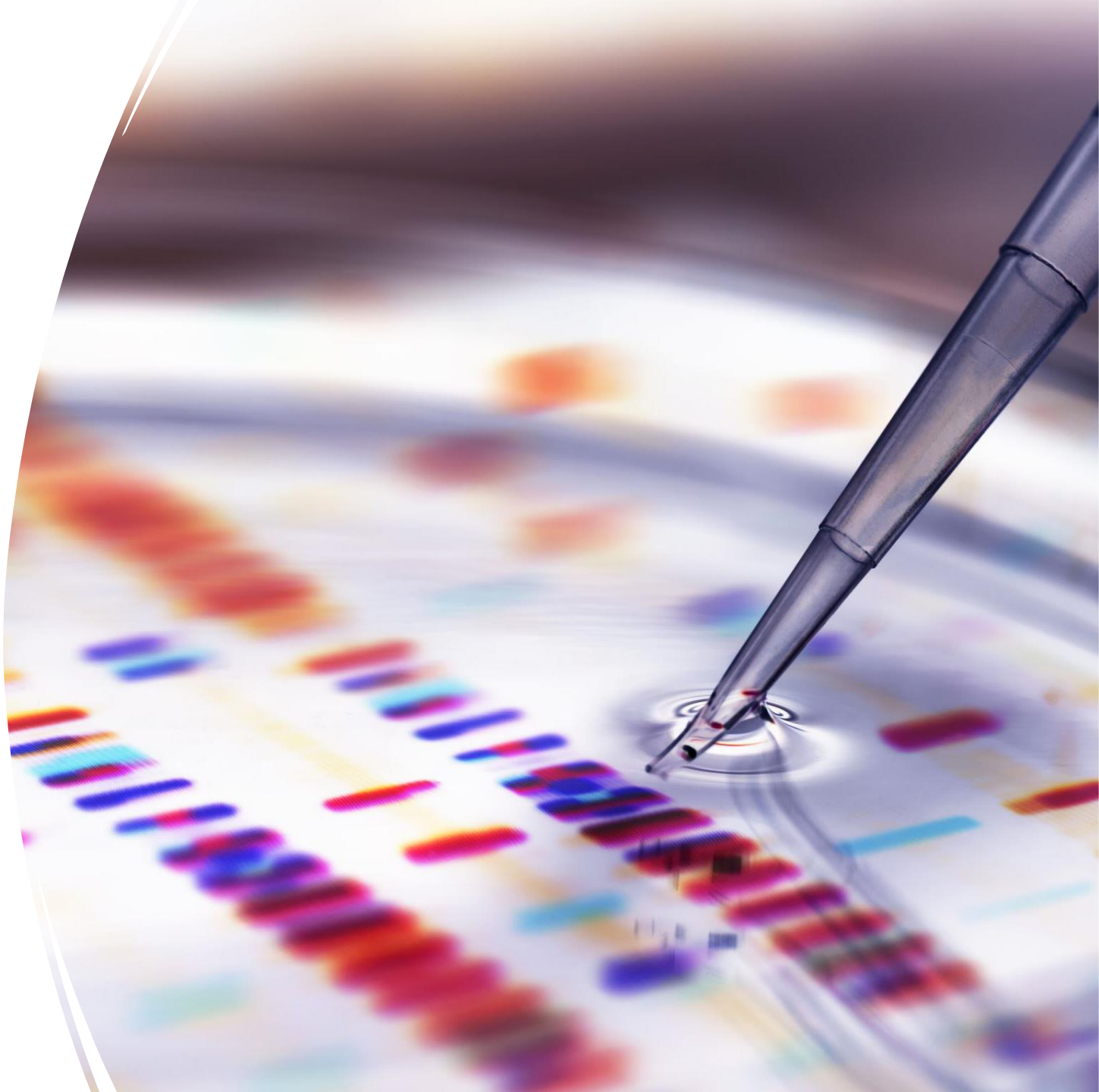
Column name	Column description
Patient Id	Represents the unique identification number of a patient
Patient Age	Represents the age of a patient
Genes in mother's side	Represents a gene defect in a patient's mother
Inherited from father	Represents a gene defect in a patient's father
Maternal gene	Represents a gene defect in the patient's maternal side of the family
Paternal gene	Represents a gene defect in a patient's paternal side of the family
Blood cell count (mcL)	Represents the blood cell count of a patient
Patient First Name	Represents a patient's first name
Family Name	Represents a patient's family name or surname
Father's name	Represents a patient's father's name
Mother's age	Represents a patient's mother's name
Father's age	Represents a patient's father's age
Institute Name	Represents the medical institute where a patient was born
Location of Institute	Represents the location of the medical institute
Status	Represents whether a patient is deceased
Respiratory Rate (breaths/min)	Represents a patient's respiratory breathing rate
Heart Rate (rates/min)	Represents a patient's heart rate
Test 1 - Test 5	Represents different (masked) tests that were conducted on a patient
Parental consent	Represents whether a patient's parents approved the treatment plan
Follow-up	Represents a patient's level of risk (how intense their condition is)
Gender	Represents a patient's gender
Birth asphyxia	Represents whether a patient suffered from birth asphyxia
Autopsy shows birth defect (if applicable)	Represents whether a patient's autopsy showed any birth defects
Place of birth	Represents whether a patient was born in a medical institute or home
Folic acid details (peri-conceptional)	Represents the periconceptional folic acid supplementation details of a patient
H/O serious maternal illness	Represents an unexpected outcome of labor and delivery that resulted in significant short or long-term consequences to a patient's mother
H/O radiation exposure (x-ray)	Represents whether a patient has any radiation exposure history
H/O substance abuse	Represents whether a parent has a history of drug addiction
Assisted conception IVF/ART	Represents the type of treatment used for infertility
History of anomalies in previous pregnancies	Represents whether the mother had any anomalies in her previous pregnancies
No. of previous abortion	Represents the number of abortions that a mother had
Birth defects	Represents whether a patient has birth defects
White Blood cell count (thousand per microliter)	Represents a patient's white blood cell count
Blood test result	Represents a patient's blood test results
Symptom 1 - Symptom 5	Represents (masked) different types of symptoms that a patient had
Genetic Disorder	Represents the genetic disorder that a patient has
Disorder Subclass	Represents the subclass of the disorder

Source : https://www.hackerearth.com/challenges/competitive/hackerearth-machine-learning-challenge-genetic-testing/

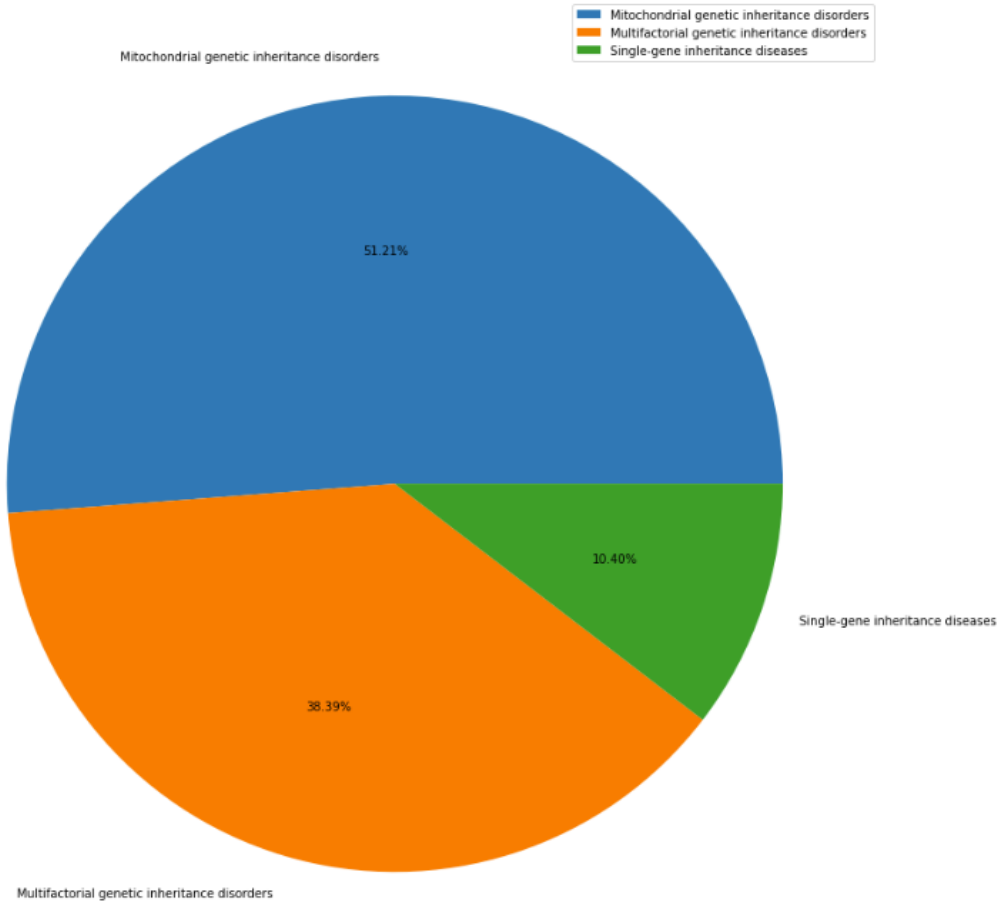
Data Wrangling

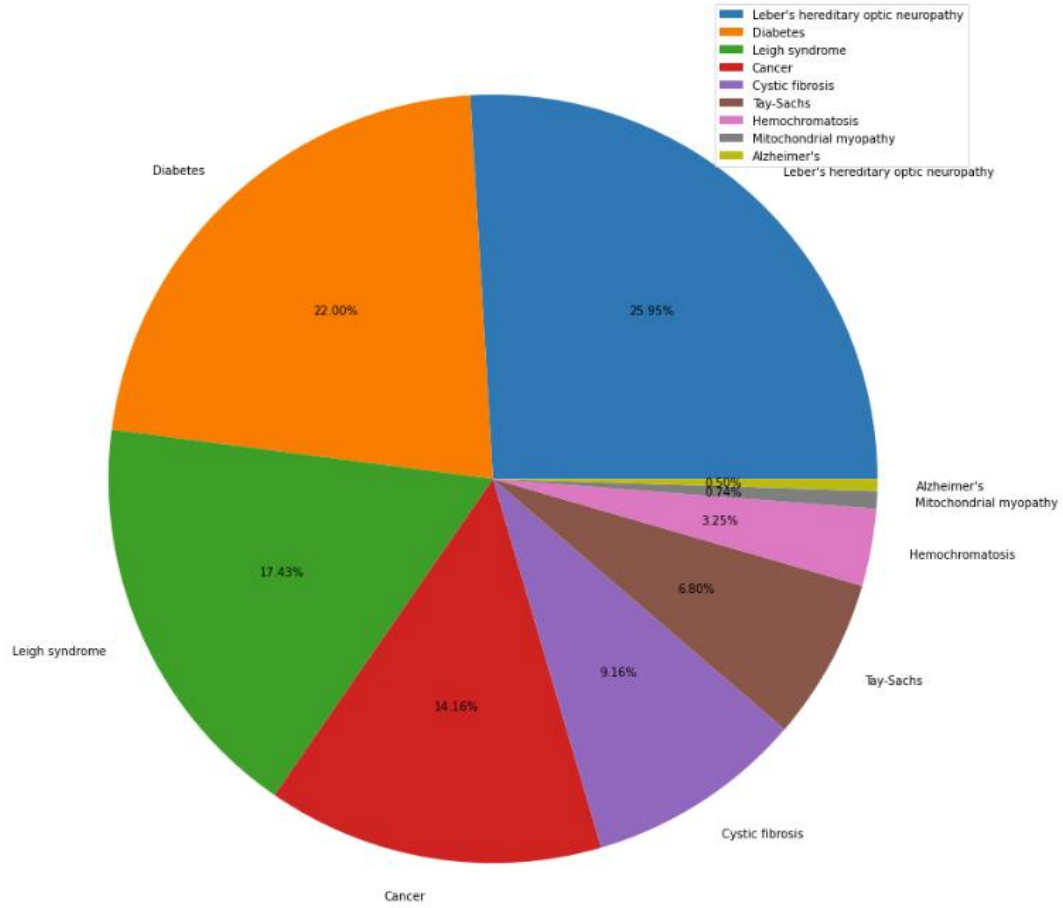
- Original Dataset contains 22K records and 45 variables.
- Dropped the variables that are not very useful for our prediction
- There are variables data representing Nan or incorrect values. e.g. values like 'Not applicable', 'None', '-', 'No Record' which can be replaced with Nan.
- Renamed Columns for simplicity
- Filled missing values with 'missing' for categorical variables
- Filled missing values with mean value for numeric variables
- The Target variables Genetic Disorder, Disorder Subclass, have many rows with null values. Drop these as they are of not any use
- After implementing above steps, Dataset rows reduced to 18047

Exploratory Data Analysis

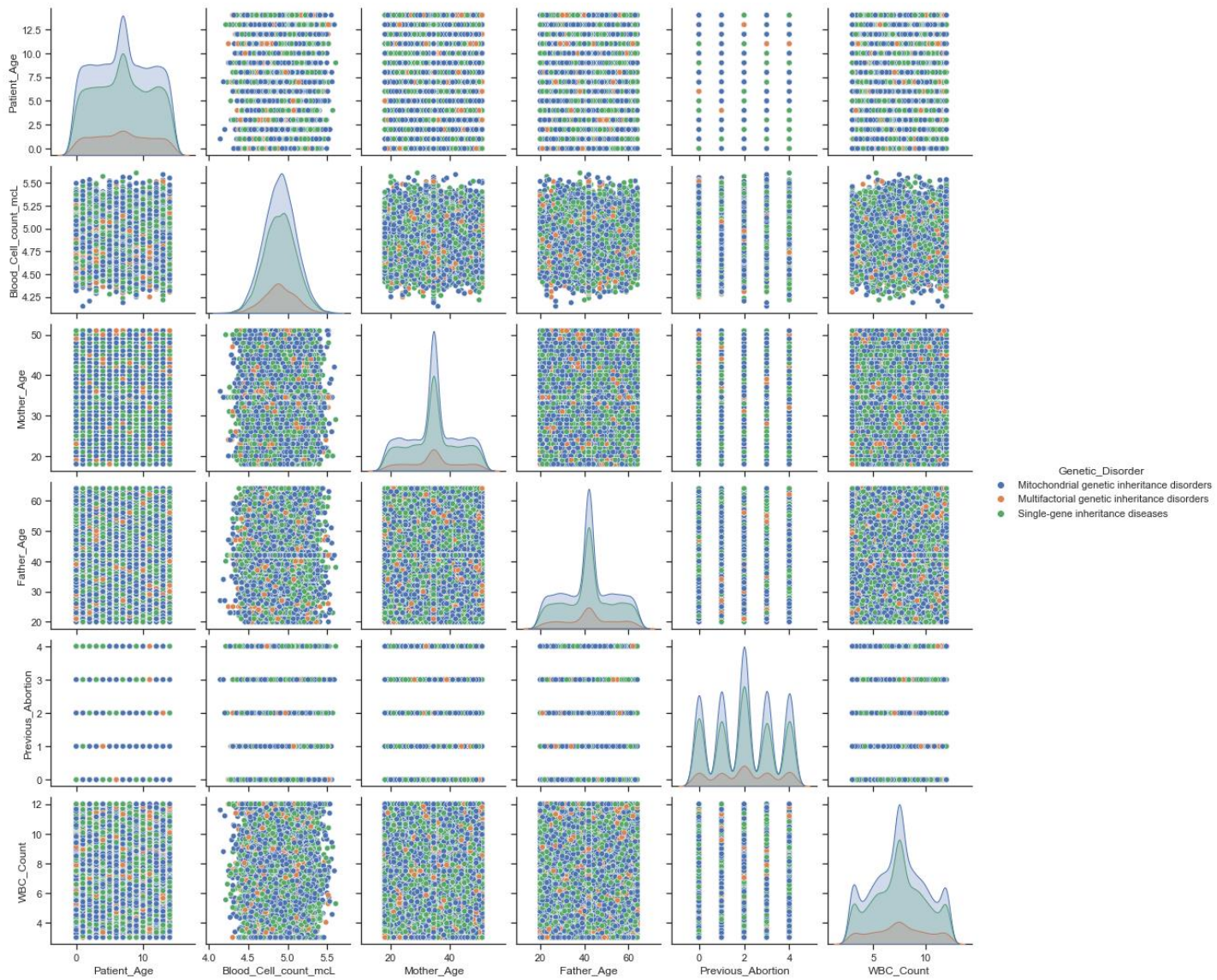


Genetic Disorder and Disorder subclass distribution

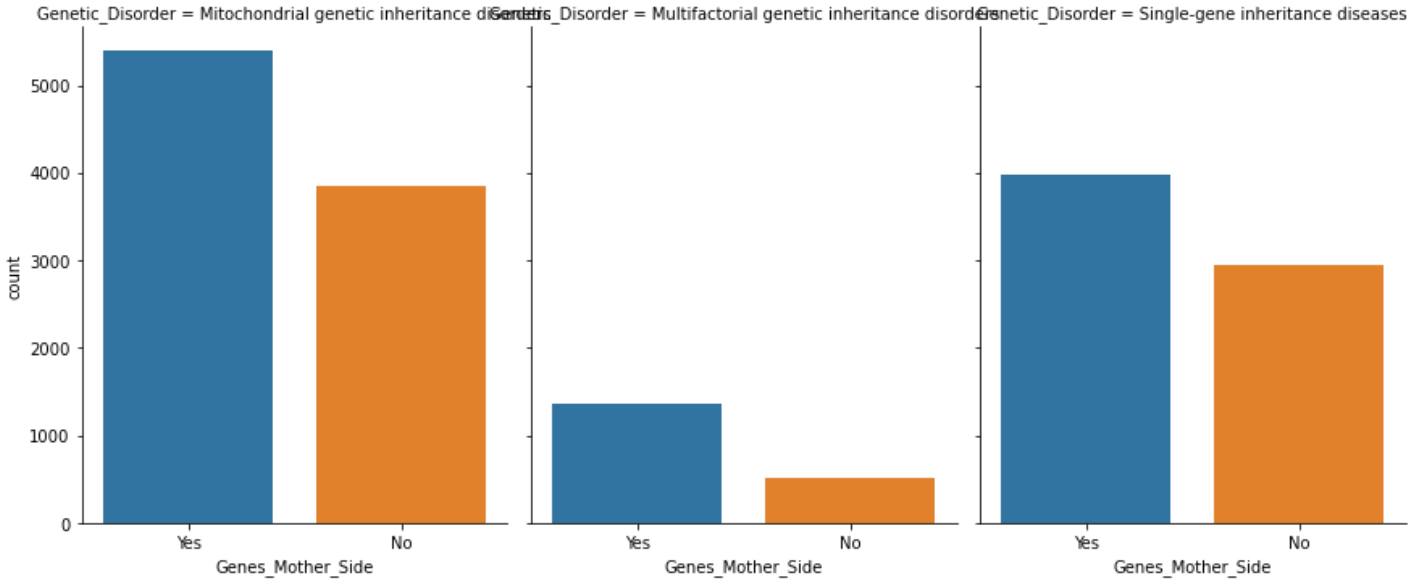




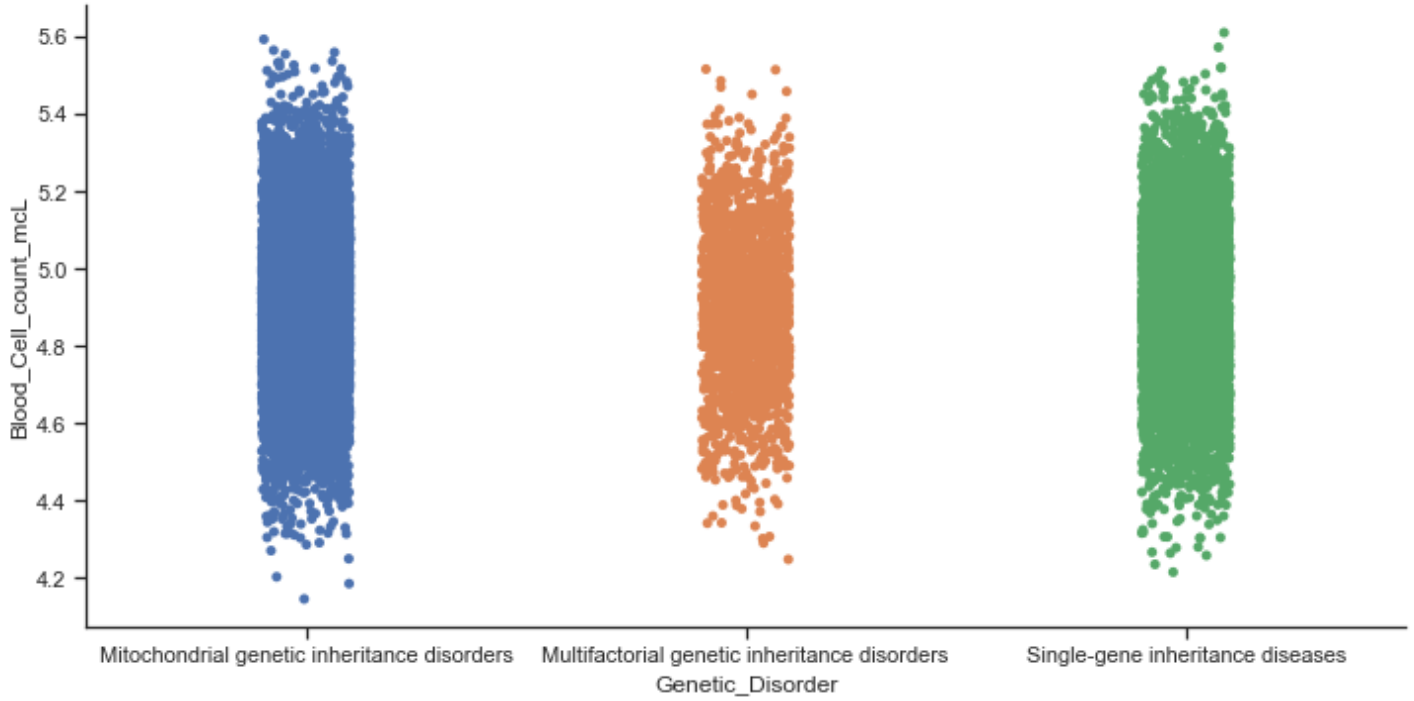
Pair plot shows relationship between variables



Genetic Disorder count vs Mother's Side distribution



Genetic Disorder vs Blood cell count distribution



Data Preprocessing

- Before feeding the data to model, we need to converted the categorical column into a numerical one using One-Hot-Encoding and label encoder
- The dataset has been separated in a Train dataset (12632 samples) and a Test dataset (5415 samples).
- The data was scaled before feeding into the respective models.

Model Selection

- Random Forest shows better accuracy and F1 score

Machine Learning Model	Accuracy	F1 Score
Logistic Regression	50.4	34.57
Decision tree	42.25	42.55
Random Forest	48.72	40.89
KNeighbors	44.82	39.73
SVM	48.25	34.15
Gradient Boosting	50.1	37.43
XGBClassifier	47.59	43.18
LGBMClassifier	48.66	40.71

For all the models under study, to avoid over-fitting, we optimized the corresponding hyper-parameters by a 5-fold cross-validation on the Train set. We then evaluated on the Test set the models trained on the entire Train set.

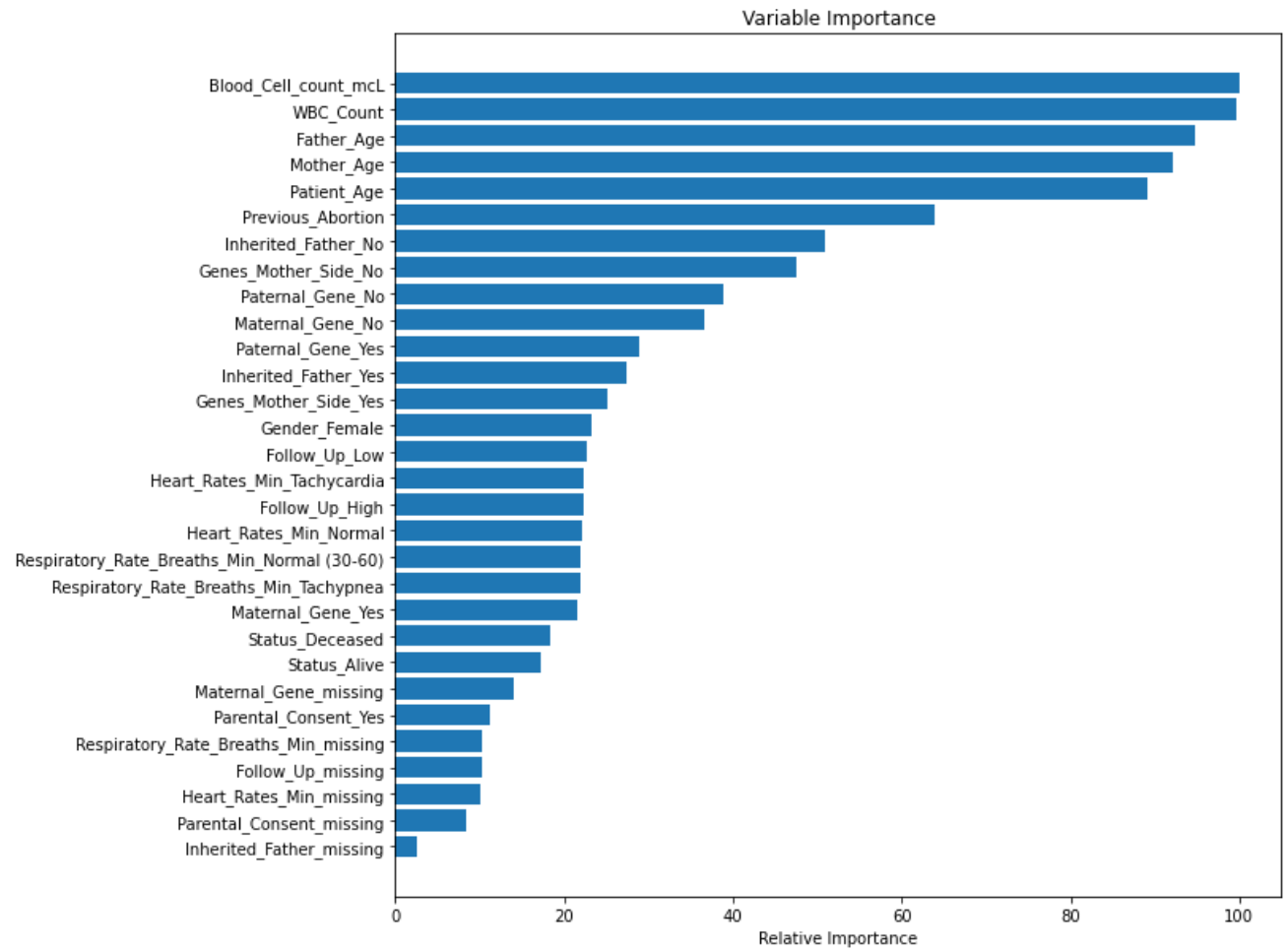
Imbalance Data Handling

- There is lot of imbalance amongst the Genetic Disorder classes
  - Mitochondrial 9241
  - Single gene 6929
  - Multifactorial 1877
- Oversampling is one of the most widely used techniques to deal with imbalance classes. Using SMOTE method, and class weight adjusted to balanced , f1 score improved to 42.65

Hyper Parameter Tunning

- Using RandomizedSearchCV, the hyperparameters are selected, using which the accuracy increased to 49.18 with F1 score : 41.69

Feature Importance



Takeaways

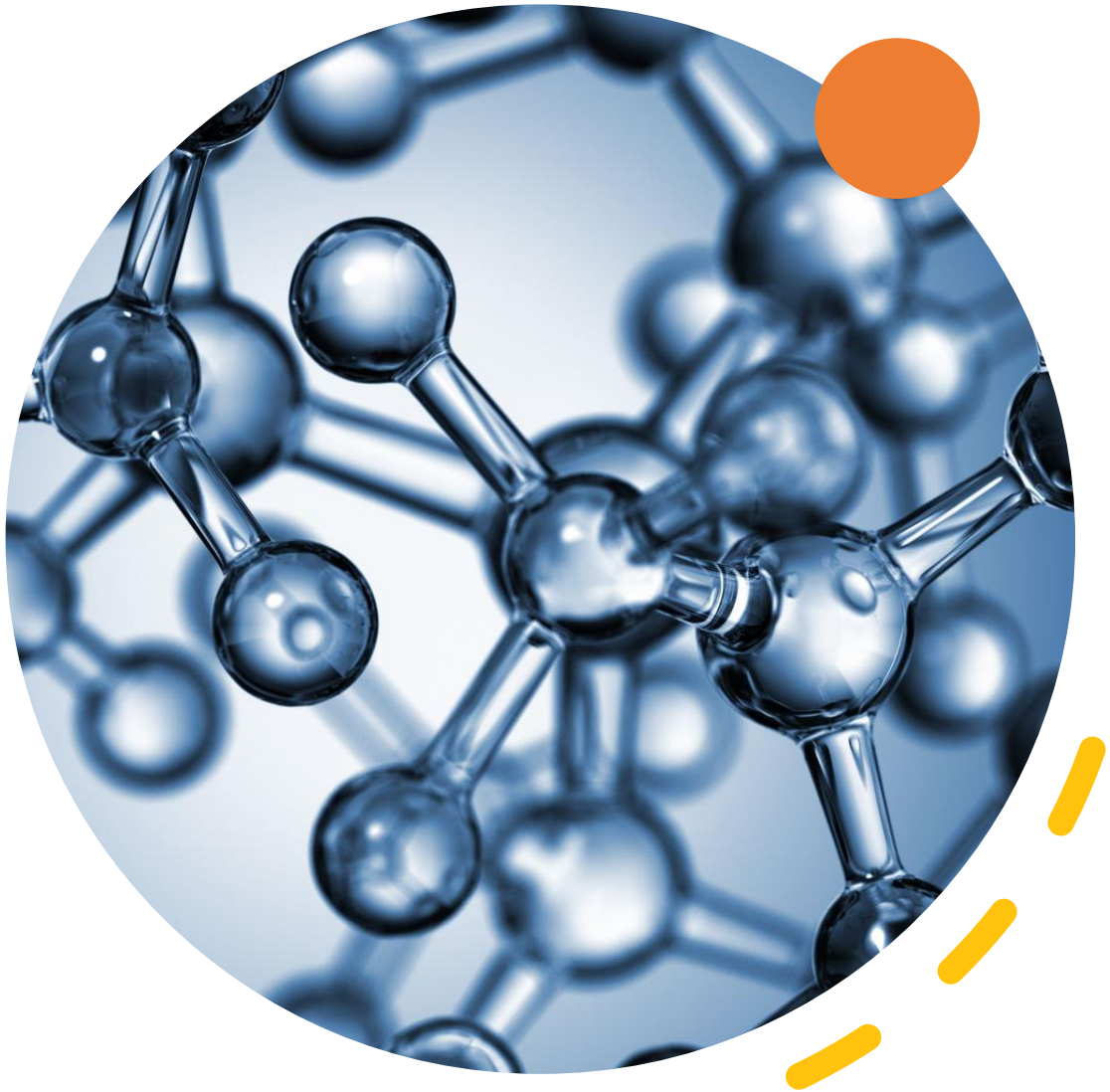
- The dataset is too small. It also with lot of missing values. After handling the missing values, we applied different models. Random Forest Classifier Accuracy : 48.72 RandomForestClassifier F1 : 40.89 Random Forest classifier showed better performance.
- As the dataset is imbalanced, using SMOTE Oversampling and handling the Class_weight, helped to improve the score. Using RandomizedSearchCV, the hyperparameters are selected, using which the accuracy increased to 49.18 with F1 score : 41.69
- Feature Importance graph shows that almost all of the features are important.

Future Research

- Expand the Data volume
- Get balanced and correct (less missing values) data
- Include more parameters

Thank You

- Dipanjan Sarkar for guidance during project



Questions?