Capstone project

Of Genomes and Genetics

Domain Background

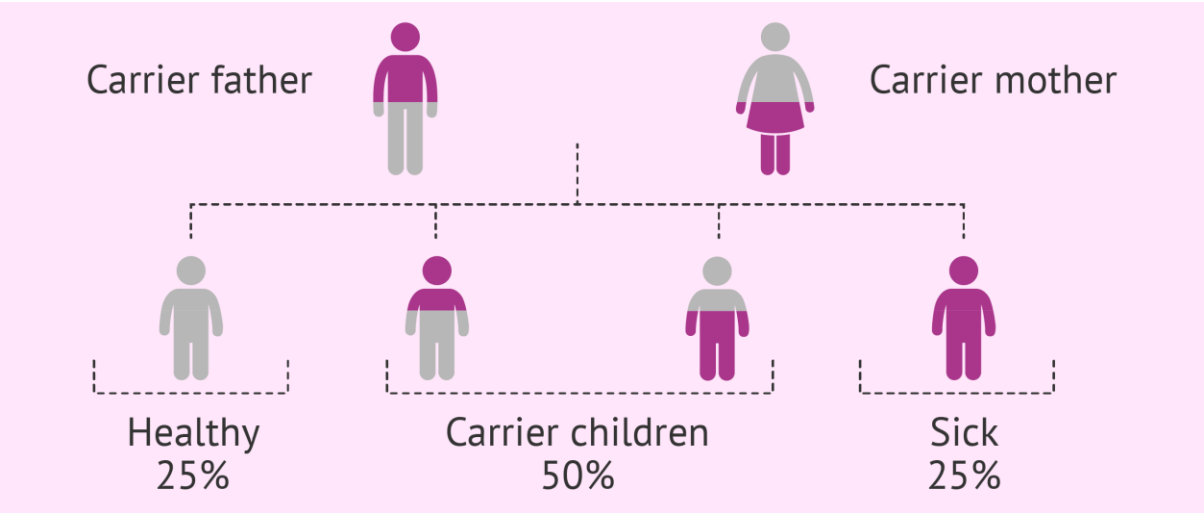
There is evidence that suggests the frequency of mutations (genetic conditions) in humans is increasing over time as our population continues to grow. Therefore, it is important that we work to understand and analyse these mutations to improve outcomes for patients through genetic screening tests and patient diagnosis.

This dataset contains information on children with a known genetic condition. Genetic conditions can be inherited (passed down) from either the mother, father or both parents. They can also appear spontaneously in the child without either parent having the condition (De Novo mutation). In this dataset only inherited conditions are considered.

There are five main inheritance patterns:

- Autosomal dominant
- Autosomal recessive
- X-linked dominant
- X-linked recessive
- Mitochondrial

Example of Autosomal recessive inheritance:



(https://www.invitra.com/en/wp-content/uploads/2014/06/autosomal-recessive-inheritance-pattern.png)

Business perspective

The goal is to be able to predict the genetic disorder and disorder subclass a child has, based on whether the parents are carriers of the condition and the child's symptoms. This could help medical professionals with their diagnosis of genetic and inherited conditions in patients. This information would in turn assist them in treating these patients.

Data description, sources and quality

Dataset is from Kaggle which sourced it from HackerEarth Machine Learning challenge. Original source unknown.

Dataset contains a training set of 22083 records and 45 columns. Test dataset contains 9465 records and 45 columns. There is also a sample_submission file with 5 rows and 3 columns. There is a reasonable amount of missing data.

The columns in this dataset are defined as follows:

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

Genetic conditions present in the dataset:

- **Mitochondrial genetic inheritance disorders**

Disease is caused by changes in the mitochondrial or nuclear DNA. Mitochondrial changes are only inherited from the mother. Nuclear DNA changes are inherited through standard inheritance pathways from either parent.

- **Single gene inheritance disorders**

When a disease is caused by a mutation in a single gene. Different mutations in the same gene can lead to the same disease.

- **Multifactorial genetic inheritance disorders**

Conditions caused by many contributing factors. They are influenced by multiple genes (polygenic) as well as lifestyle and environmental factors.

Subclasses of genetic conditions in the dataset:

- Leigh syndrome
- Mitochondrial myopathy
- Cystic fibrosis
- Tay-Sachs
- Diabetes
- Hemochromatosis
- Leber's hereditary optic neuropathy
- Alzheimer's
- Cancer

Proposed solution – evaluation metrics

From notebooks previous results of 0.603-0.954 achieved: which metric used??

Model used	Result achieved
Random forest	0.603
Boosting (LGBM)	0.686
KNN	0.788
SVM (default hyperparameters)	0.949
SVM (rbf kernel and C = 100)	0.954

Appears that SVM model could be a good option and try to optimize further. Could also try optimizing KNN model. I would like to try a stacking model and see how the meta classifier does. Other options?

Project design – data preprocessing methods to use.

- There is a lot of missing data and there are outliers present.
- Where features highly correlated – remove 1 of the features.
- Feature selection
- Remove columns that don't provide information due to too many missing values or other reasons.
- Investigate outliers – remove or keep?
- String columns to numeric – one hot encoding or encode to numeric categories.
- Handling NaNs - NaNs in many columns - impute with median or mode.
- Normalization of features.
- Use cross-validation.

References

Dataset:

https://www.kaggle.com/datasets/aryarishabh/of-genomes-and-genetics-hackerearth-ml-challenge

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

Previous models:

https://www.kaggle.com/datasets/aryarishabh/of-genomes-and-genetics-hackerearth-ml-challenge/code?select=train.csv

Domain knowledge:

https://www.ncbi.nlm.nih.gov/books/NBK115561/

https://www.genomicseducation.hee.nhs.uk/documents/inheritance-and-genetic-conditions/

https://pubmed.ncbi.nlm.nih.gov/31424543/

https://rarediseases.info.nih.gov/diseases/7048/mitochondrial-genetic-disorders

https://www.ncbi.nlm.nih.gov/books/NBK132154/

https://medlineplus.gov/genetics/understanding/mutationsanddisorders/complexdisorders/

Video explaining where mutations come from – 1st minute good

https://nz.video.search.yahoo.com/search/video?fr=mcafee&ei=UTF-8&p=de+novo+mutations&type=E211US1144G0#id=3&vid=606a0df890348cf56b3c37267b38ac29&action=click