UNLOCKING HEALTH: EARLY GENETIC TESTING FOR DISORDER DETECTION

Oct 31st, 2023

Anusha Guruprasad – ag84104n@pace.edu

CS – 667 : Practical Data Science

EXECUTIVE SUMMARY SLIDE

- The world's population is growing too fast, and many people don't have enough healthcare, food, and places to live. This is causing more genetic disorders.
- Not enough people know about these diseases, and we don't do enough genetic testing, so these diseases are becoming more common.
- Many children are dying from these diseases, so it's crucial to do genetic testing when someone is pregnant.
- The dataset has information about children with genetic disorders, and we can use it to predict what kind of genetic disorder they have.

PROJECT PLAN

Deliverable	Details	Due Date	Status
Data & EDA	Create final assignment deck skeleton and fill in the deck up to the end of the EDA section	10/31/23	Completed
Methods, Findings, and Recommendations	Fill in the Methods, Findings, and Recommendations sections of the deck	11/7/23	In Progress
Final presentation	Send in final completed deck and present it	11/21/23	Not started

DATA

DATA DETAILS

- Data Source: Kaggle
- Sample Size : 22083 rows and 43 columns
- Data removed during pre-processing - Patient ID, Patient First Name, Test 1, Test 2, Test 3, Test 4, Test 5, Family Name, Father's name, Location of Institute, Institute Name
- The data has 2 major prediction variables that need to be considered – multi-output classification (multiple predictions at the same time). The two target variables are Genetic disorder as well as the disorder subclass.

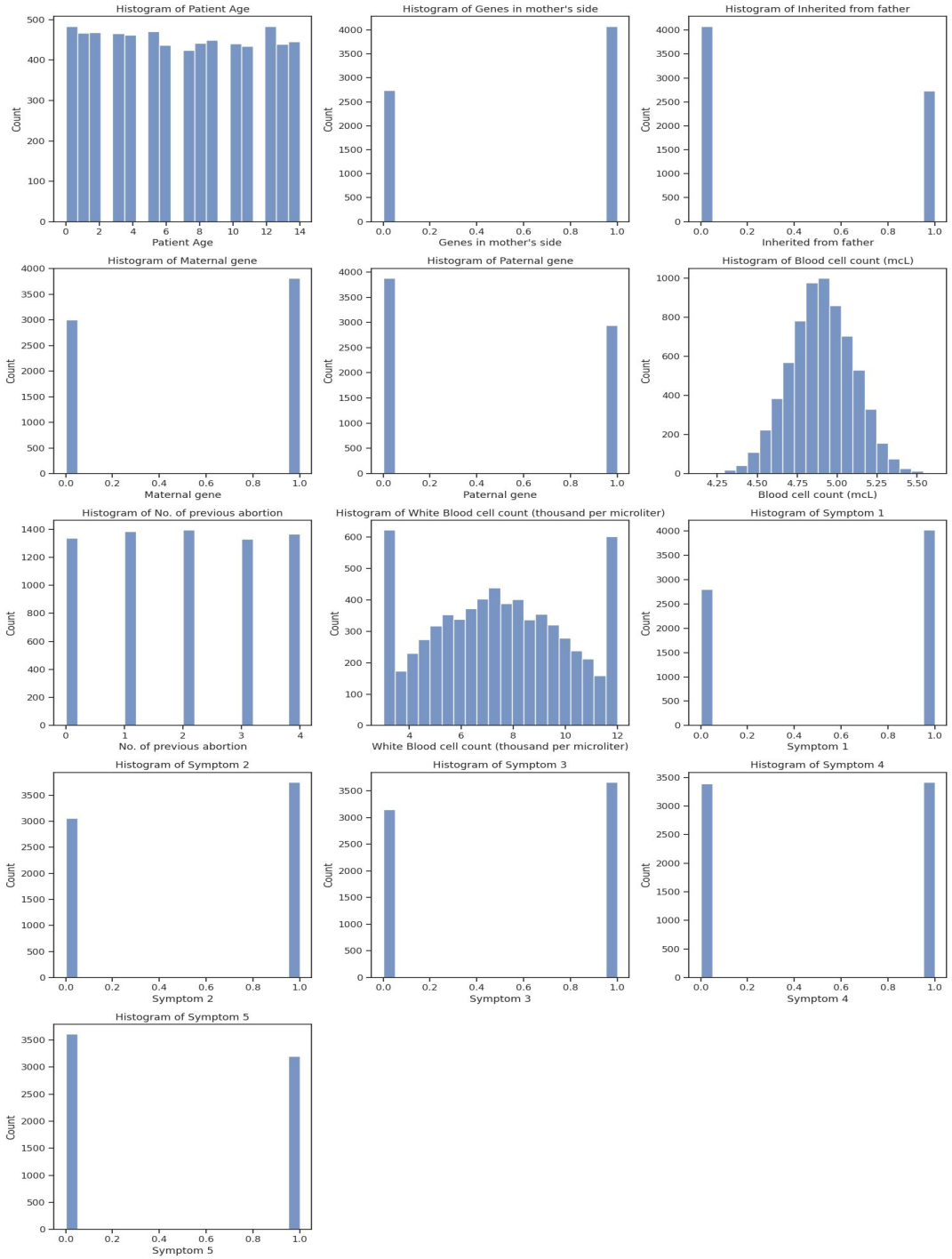
ASSUMPTION

- It is assumed that inherited disorders are more significantly influenced by maternal genes than paternal genes. This is because females have 2 ‘X’ chromosomes whereas males have one ‘X’ and one ‘Y’ chromosome.
- The mitochondria, often referred to as the cell's powerhouse, contain their own DNA. Mitochondrial DNA is inherited exclusively from the mother.

EXPLORATORY DATA ANALYTICS

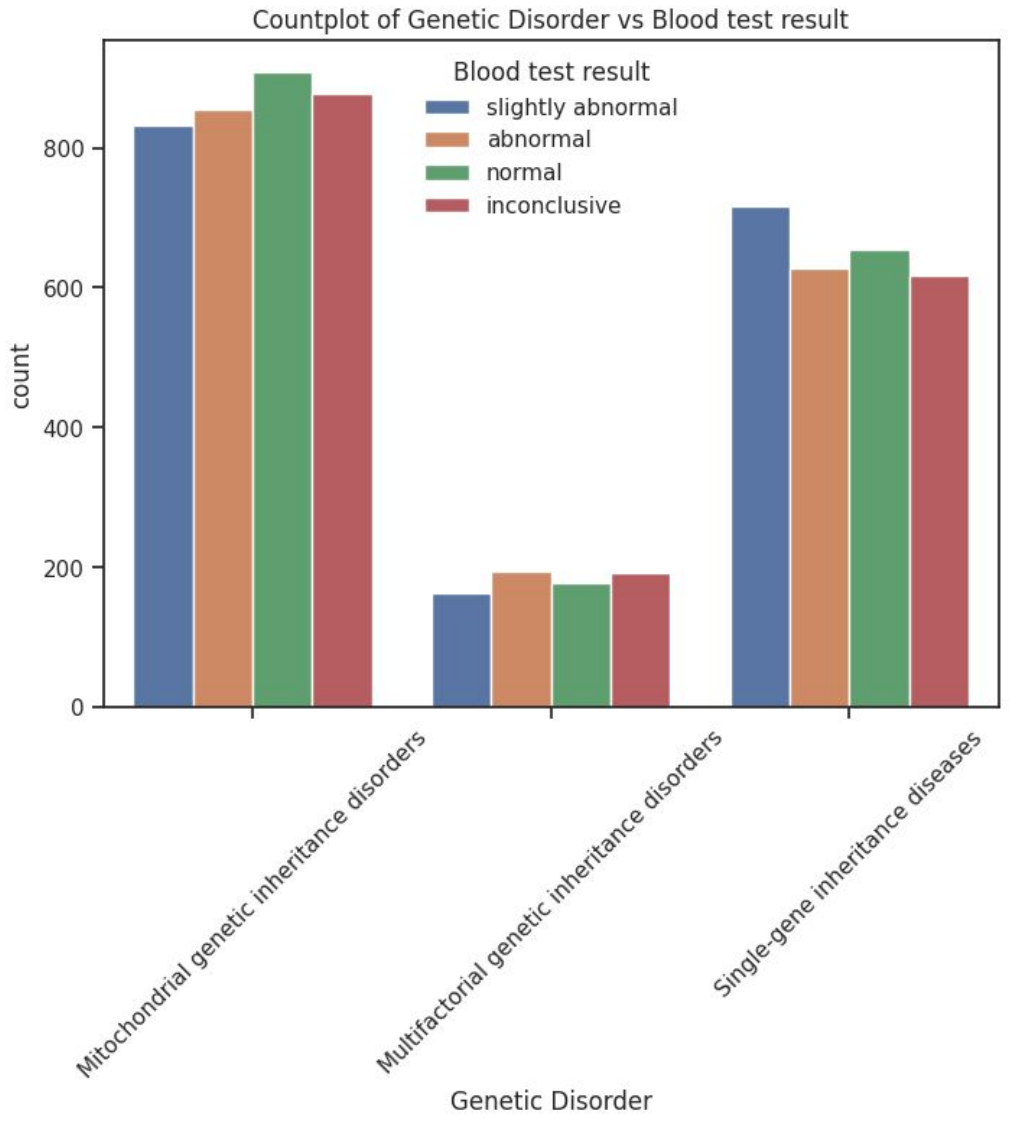
GRAPHS FOR NUMERICAL DATA

- Distribution of the data or the different types of values each column holds.
- Example: Mother side gene – Present or not.
- This helps us understand what kind of data we have.
- Examine how the data is spread out and check if there are any columns where the information is uneven or lopsided. In other words, see if some columns have much more or much less data compared to others.

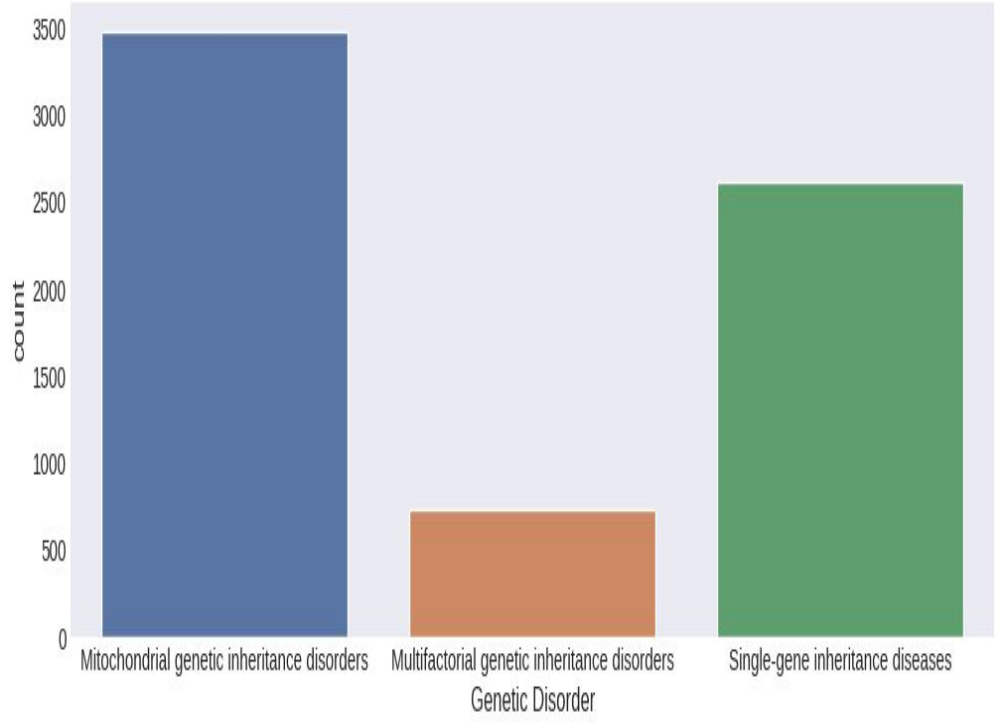


BLOOD TEST HOLDS LITTLE TO NO RELATION

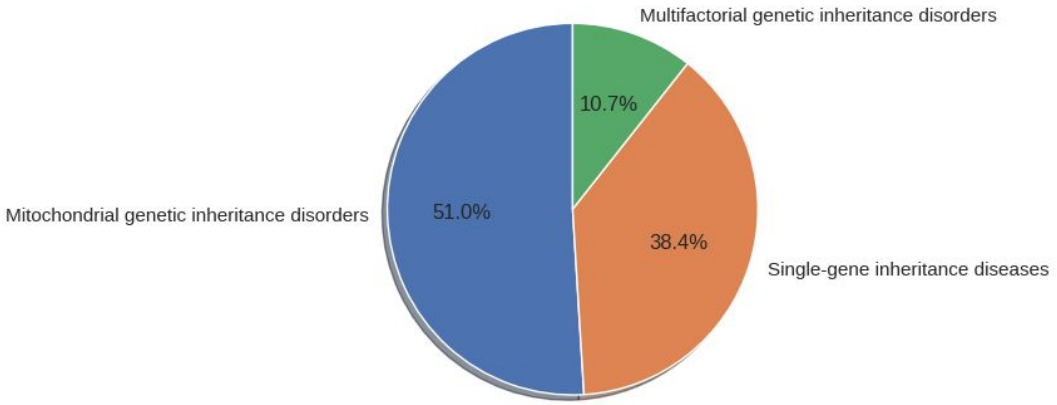
- Among the 3 disorders we can see blood test does not actually detect the existence of it.
- Running general tests would not be able to detect such disorders.
- Hence, we need to run this at early stages.



UNDERSTANDING THE FIRST TARGET/GENE



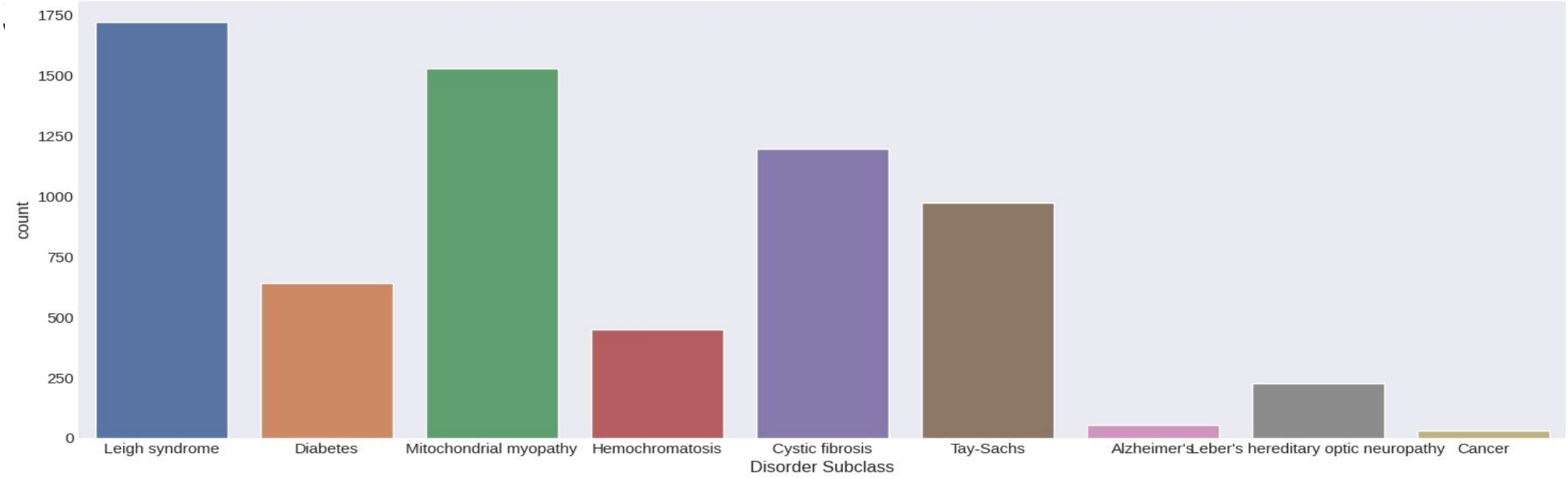
- Mitochondrial genetic inheritance disorders are more frequent than Multifactorial genetic inheritance disorders.



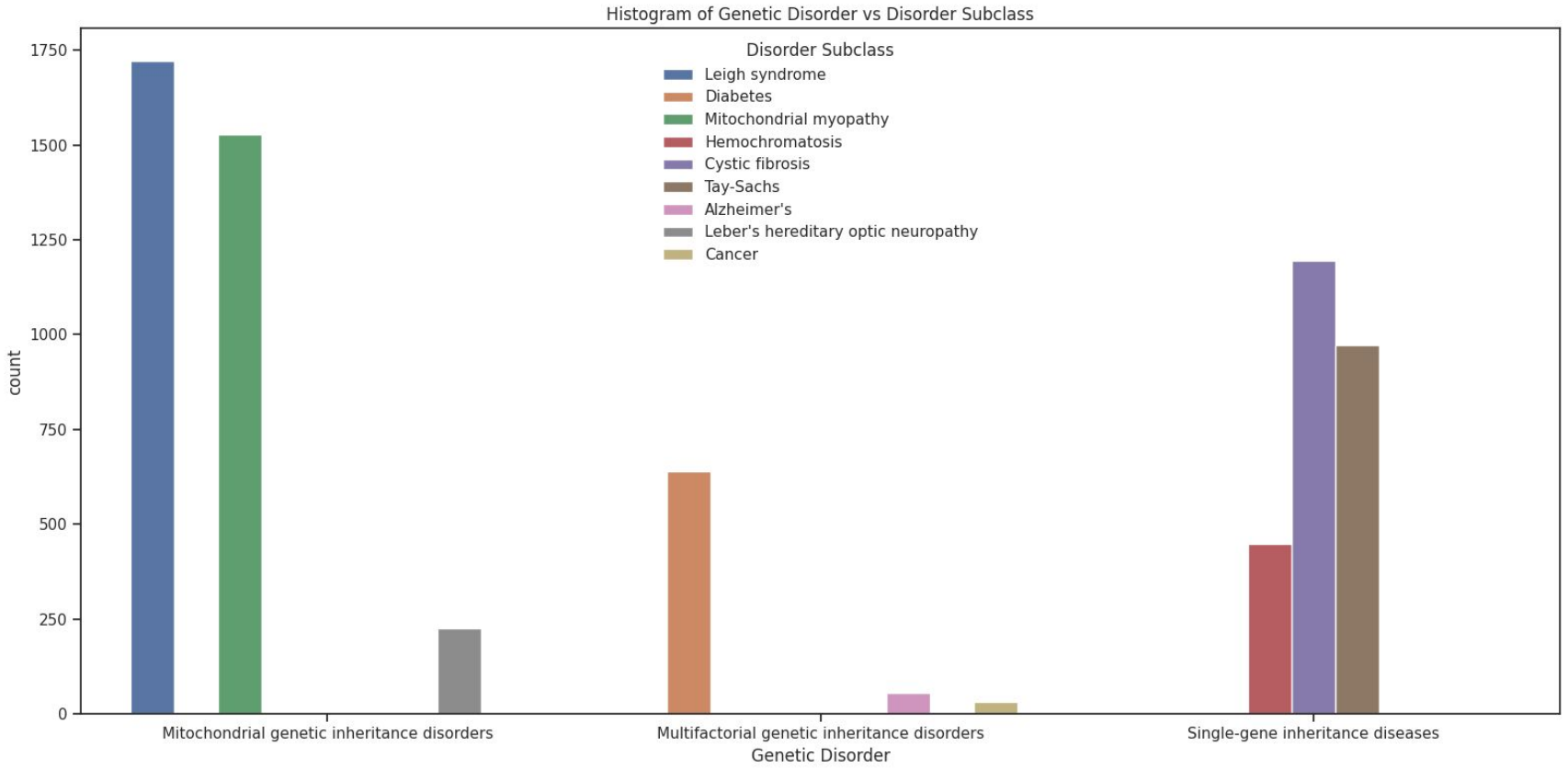
Note: the two graphs represent percentage and numeric representation

UNDERSTANDING THE 2ND TARGET/DISORDER

- The distribution of subclass of disorders. Leigh Syndrome is the most common subclass.
- Now how do I understand which disorder it belongs to?



EXPLORING HOW THE TWO MAIN CHARACTERISTICS WE WANT TO PREDICT ARE RELATED TO EACH OTHER.



MODELING METHOD SECTION

OUTCOME VARIABLE

- Trying to predict the ‘Disorder Subclass’ can help understand the major reason for a disorder and its subclass being detected in a child.
- The columns chosen are all related to the parent’s hereditary information, the child’s medical history, and symptoms data.
- Below is the detailed list of all the columns or feature’s that are being considered for the modeling.
- Columns Categories :
  - Patient data : 'Patient Age’, 'Status’, 'Gender’, 'Place of birth', 'Follow-up’, ‘Parental consent’.

- Patient Health condition : 'Blood cell count (mcL)', 'Respiratory Rate (breaths/min)', 'Heart Rate (rates/min)', 'Autopsy shows birth defect (if applicable)', 'Folic acid details (peri-conceptional)', 'H/O serious maternal illness', 'H/O radiation exposure (x-ray)', 'H/O substance abuse', 'White Blood cell count (thousand per microliter)', 'Blood test result', 'Symptom 1', 'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5', 'Genetic Disorder', 'Disorder Subclass'.
- Parental Data: 'Genes in mother's side', 'Inherited from father', 'Maternal gene', 'Paternal gene', 'Assisted conception IVF/ART', 'History of anomalies in previous pregnancies', 'No. of previous abortion', 'Birth defects'mother's_age_binned', 'father's_age_binned'.

- All the above columns represent and link the patient medical data along with the parent medical history which is vital in identifying any genetic or genomic disorders.

MODEL TYPE: MULTI OUTPUT CLASSIFIER

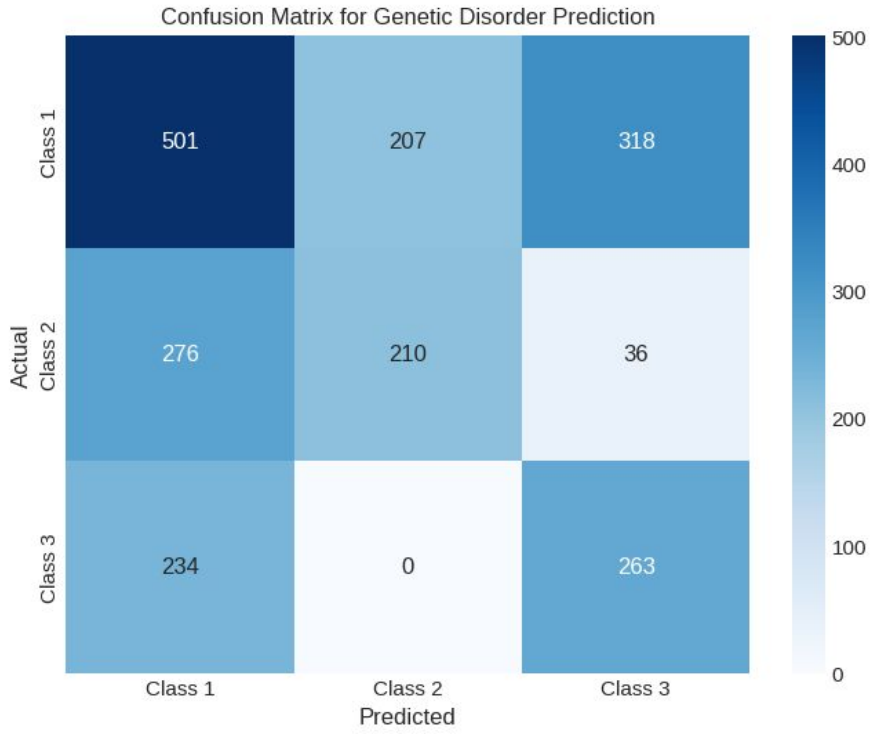
- As the data consists of two major columns that need to be focused on detection : Genetic Disorder and Disorder Subclass.
- In non-technical terms, MultiOutputClassifier is a tool that helps us handle a scenario where we have multiple things to predict (multiple output variables) instead of just one. Imagine you're trying to predict not only whether a patient has a specific genetic disorder but also the subclass of that disorder. Instead of building two separate models for these predictions, you can use MultiOutputClassifier to create a single model that considers both predictions simultaneously.

- Here's an analogy: Think of a student taking two different exams, say, one for Biology and another for Chemistry. Instead of studying separately for each exam, the student decides to use a special study strategy that helps them prepare for both subjects at the same time. This strategy is similar to how MultiOutputClassifier works – it helps the model learn and make predictions for multiple things at once.
- For a technical Understanding of the model: Click here to move to the appendix slide

FINDINGS

UNDERSTANDING THE OUTCOME :

- Assisted conception IVF/ART, H/O substance abuse, Birth defects, Inherited from the father, and History of anomalies in previous pregnancies are the main columns or features that need to be considered as they have weightage on the Genetic disorder and subclass detection according to the model.
- Many records have similar data and can be misinterpreted by the model while going through it during the understanding stage (fitting stage).
- Understanding the underlying issue :
  - The number of correctly identified records are the diagonal values – 501 for class 1, 210 for class 2, and 263 for class 3.
  - Adding all that divided by the total records gives accuracy.



ACCURACY VS RECALL:

- Recall: measures how many of the actual positive instances were correctly predicted by the model. For instance, in a medical test for a disease, the recall would tell you how many people with the disease were correctly identified as having it.
- Accuracy: refers to how often a model's predictions are correct among all predictions made. It's like checking the percentage of answers you got right on a test. For instance, if a model predicts 80 out of 100 instances correctly, its accuracy is 80%.
- For critical applications such as medical diagnostics or safety-related predictions, achieving higher recall values (closer to 1.0) is preferred to minimize false negatives and ensure that as many positive cases as possible are correctly identified.

```
Recall for Genetic disorder: 0.49259624876604147
```

- This is a low recall value. This must be increased as the data deals with medical diagnostics among infants.
- Now the accuracy and recall for the second target or predicted column :

```
accuracy_disorder_subclass
```

```
0.0112469437652811 74
```

```
Recall for disorder subclass: 0.3864491844416562
```

- From the above, we can see that the recall has dropped again. This is because once the Genetic disorder is detected then only will we move to its subcategory.
- In a business context, understanding these coefficients helps in identifying which features (or factors) are more influential in predicting genetic disorders. For instance, if a specific genetic marker or clinical attribute has a high coefficient, it implies that it significantly influences the likelihood of a particular genetic disorder. This insight can guide healthcare professionals or researchers in identifying crucial factors or markers for disease prediction or risk assessment, aiding in better diagnosis or intervention strategies.

RECOMMENDATIONS

ACTIONABLE :

- The analysis of this model centers around assessing the relevance of genetic testing at birth, aiming to address potential health issues before they escalate in later life stages, potentially leading to severe consequences.
- The data indicates noteworthy observations related to specific columns like 'Birth Defects' and 'History of Anomalies in Previous Pregnancies.' These factors seem to provide significant insights into potential health conditions.
- Additionally, the model highlights the necessity for more comprehensive details regarding the parents in understanding a child's genetic predispositions. The genetic makeup of a child is notably influenced by the genetic information inherited from their parents.
- Therefore, exploring and considering the genetic attributes of the parents emerges as a crucial aspect to better comprehend the potential health outcomes and predispositions of the child. This emphasizes the importance of investigating parental genetic information to gain a comprehensive understanding of the child's genetic composition and potential health risks.

NEXT STEPS:

- Employing hyperparameter tuning techniques can provide deeper insights into the significance of each feature within the model.
- This approach aims to refine the model's performance by optimizing its settings, potentially resulting in improved outcomes and a more nuanced understanding of the influence wielded by individual features.
- Alternatively, different modeling approaches allow us to delve deeper into the data dynamics.
- While supervised learning provides insights based on labeled information, leveraging clustering techniques offers an opportunity to unveil inherent patterns and relationships within the dataset without predefined labels.
- This enables a more organic discovery of underlying structures, potentially revealing subtleties and nuances that may not be immediately apparent through supervised methods.

APPENDIX

TECHNICAL UNDERSTANDING OF THE MODEL

- The code uses a powerful machine learning technique called XGBoost to predict multiple health conditions from certain information.
    1. Setting Up Model: It sets up an XGBoost model to understand and learn from the data patterns related to different health conditions.
    2. Data Division: The code divides the available data into two parts: one for teaching the model (training data) and another for testing its performance (validation data).
    3. Training and Prediction for Each Condition: It teaches the model to predict specific health problems (like different genetic disorders) using the training data. Then, it tests how accurately the model predicts these problems with the validation data.

4. Overall Performance Check: Finally, it evaluates how well the model performs in predicting all the health conditions together, considering both genetic disorders and their subclasses.

- The code helps measure the model's accuracy in predicting various health issues, assisting in understanding how reliable it is in identifying these conditions.

Click here to go back to the modeling section.

DETAILS:

- https://scikit-learn.org/stable/modules/generated/sklearn.multioutput.MultiOutputClassifier.html
- https://towardsdatascience.com/clean-efficient-data-pipelines-with-pythons-sklearn-2472de04c0ea
- https://towardsdatascience.com/essential-guide-to-multi-class-and-multi-output-algorithms-in-python-3041fea55214
- https://medium.com/@cactuscode/multioutput-multiclass-classification-b0737a0693ec