Gene Galaxy : Genetic Disorder Analysis

Tahirah Mohi-ud-din
Department of Computer Science
KLE Technological University
Belagavi, India
02fe22bcs165@kletech.ac.in

Abhay Patil
Department of Computer Science
KLE Technological University
Belagavi, India
02fe22bcs002@kletech.ac.in

Rohan Kulkarni
Department of Computer Science
KLE Technological University
Belagavi, India
02fe22bcs083@kletech.ac.in

Sangamesh
Department of Computer Science
KLE Technological University
Belagavi, India
02fe22bcs103@kletech.ac.in

***Abstract*—** Genetic disorders arise from mutations or alterations in the genome, leading to significant changes in the structure or function of organisms. Such mutations may lead to fatal diseases like cancer, Alzheimer's. The exploratory data analysis (EDA) project is designed to explore a dataset on genetic disorders so as to discover significant patterns and insights. These include higher gene paternal over inheritance than maternal side and the possible protective effects of folic acid against cardiomyopathy and birth defects. Respiratory rate analysis showed that dominant genetic disorders were present in certain categories. Gender analysis demonstrated that there was a higher prevalence of genetic disorders amongst males. Previous abortion statistics revealed a remarkable occurrence of maternally inherited mitochondrial genetic diseases. A correlation between maternal age and genetic disorders has shown that children born by women with ages from 27-40 years were at high risk of having one. We have carried out many other analyses which have been presented separately. Healthcare professionals and policymakers can use these findings during early detection and intervention strategies, based on this EDA's findings. In future work, machine learning may be used to improve early diagnosis as well as personalized treatment plans for patients suffering from this condition. This EDA gives a primer on how specific trends are observable in different types of genetic disorders

Keywords—Genetic disorder, Genome mutation, Exploratory data analysis.

Introduction :

Genetic disorders are due to mutations or modifications in the genome, leading to significant changes in organisms' structure or functions. Since genes are made up of deoxyribonucleic acid (DNA) and they contain essential biological information, alteration of DNA sequences through mutation may lead to different types of genetic disorders. These problems may appear as single gene inheritance disorders, chromosomal disorders, mitochondrial genetic inheritance disorders or complex multifactorial genetic inheritance disorders.

Motivation:

Due to its complexity and size it is difficult to analyze genomic data manually resulting into errors. Genomics which is a specialized branch of bioinformatics studies genomes by understanding their structures, identifying abnormalities and exploring their implications. Because of these many features it has become important for an overall assessment on genomic data that may help identify the underlying mechanisms behind the various genetic diseases such as cancer, diabetes and Alzheimer's disease.

However, accurately analyzing genomic data especially when it comes to predicting genetic disorder poses still some challenges.

a) Objective

The aim of this project is to perform Exploratory Data Analysis (EDA) on genomic data with a focus on data cleaning and pre-processing stages. The main goals are:

(1) Cleaning and pre-processing of genomic data for later analysis

(2) Study basic characteristics and trends in the dataset

(3) Formulate hypotheses from first look at the data.

b) Contribution

This research contributes towards comprehension of genetic information through Exploratory Data Analysis (EDA), all the way to hypothesis formulation. This involves cleansing and preparing genomic data for analysis through it, followed by presenting come-upon remarks that build hypotheses.

I. RELATED WORK

This section examines the related literature on our proposed research approach.

One of such research is [1] predicting genetic disorder and types of disorder using chain classifier. Ali Raza, Furqan Rustum et al., in their study, raised the bar in prediction of genetic disorders by basing their new machine learning method on a chain classifier. This approach combines class probabilities from Extra Trees (ET) and Random Forest (RF) models to give rise to a new feature set and presents itself as an improvement over extant methods by exploiting a classifier chain. The Extreme Gradient Boosting (XGB) model emerged as the best performer with 92% α-evaluation score and 84% macro accuracy score against current state-of-the-art approaches. It achieved superior performance compared to existing works in both accuracy and computational efficiency.

The second one is [2] Supervised machine learning empowered multifactorial genetic inheritance disease prediction. In this work, Taher M. Ghazal, Hussam Al Hamadi et al., designed an impact of machine learning on determining various diseases in medical or biomedical subjects because it can recognize what has been seen before by its training process. Using two types of ML techniques: SVMs as well as KNNs for diagnosing diabetes, dementia and cancers resulting from polygenic inheritance disorders were applied in the proposed model. The proposed model SVM achieves

the highest testing prediction classification accuracy of 92.5% as compared with the proposed model of KNN.

[3] The future of genetic disease studies: assembling an updated multidisciplinary toolbox. Swetha Ramadesikan, Jennifer Lee & Ruben Claudio in this piece suggest that the application of a modernized multidisciplinary toolkit involving sophisticated methodologies and several methods to investigate diseases will enable further advances in how we understand and manage genetically-inherited conditions.

[4] Machine learning-based genetic diagnosis models for hereditary hearing loss by the GjB2, SLC26A4 and MT-RNR1 variants. In this study by Xiaomei Luo, Fengmei Li, Kaicheng Hong, Jiansheng Chen, Xiaohe Chen and Hao Wu has constructed predictive models for the clinical HHL gene loci based on various types of ML algorithms along with traditional risk assessment algorithms. After that, they compared them to the predictive performance of human experts. In terms of the comprehensive evaluation of ML algorithms in genetic diagnosis, the SVM performed bestment with an accuracy capacity of 81.4%, and a higher balanced performance value comparing EI and GRS.

II. DATASET DESCRIPTION AND PREPROCESSING

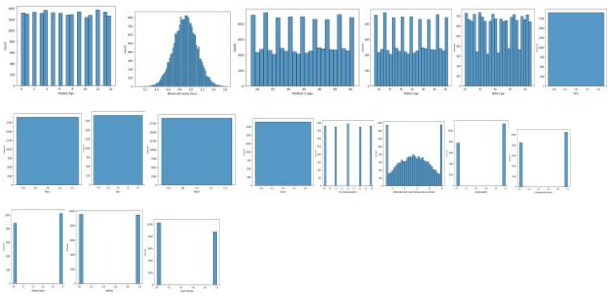
In this section, we provide a description of the dataset used in our analysis and detail the preprocessing steps applied to prepare the data for exploratory data analysis (EDA).

This dataset appeared as a prediction challenge on Kaggle few years ago.
Source URL:
[https://www.kaggle.com/datasets/aibuzz/predict-the-genetic-disorders-datasetof-genomes]

The dataset has 21011 data (Tuples) and 45 features (Attributes). Our target variables are genetic disorder & disorder subclass.

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	Represents different (masked) tests that were conducted on a patient
Test 2	Represents different (masked) tests that were conducted on a patient
Test 3	Represents different (masked) tests that were conducted on a patient
Test 4	Represents different (masked) tests that were conducted on a patient
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	Represents (masked) different types of symptoms that a patient had
Symptom 2	Represents (masked) different types of symptoms that a patient had
Symptom 3	Represents (masked) different types of symptoms that a patient had
Symptom 4	Represents (masked) different types of symptoms that a patient had
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

A. *Data Distribution*



Handling Missing values

We are identifying missing values by examining each column of the dataset for null or NaN values using python's pandas library . In our dataset the attributes like patient Id , Genes in mother's side , Paternal gene , Blood cell count (mcl), Patient First Name, Father's name , Location of institute , status are free from missing or null values . Rest attributes contain missing values & the description of each attribute with missing values is given below :

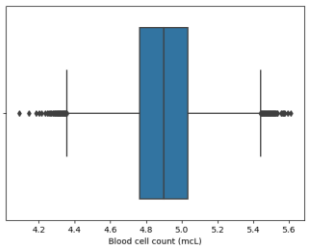
Attributes	Missing Values
Patient Id	1072
Patient Age	2440
Genes in mother's side	1072
Inherited from father	1359
Maternal gene	3766
Paternal gene	1072
Blood cell count (mcL)	1072
Patient First Name	1072
Family Name	10312
Father's name	1072
Mother's age	6790
Father's age	6761
Institute Name	5932
Location of Institute	1072
Status	1072
Respiratory Rate (breaths/min)	3131
Heart Rate (rates/min	3097
Test 1	3091
Test 2	3125
Test 3	3113
Test 4	3121
Test 5	3144
Parental consent	3092
Follow-up	3142
Gender	3135
Birth asphyxia	3130
Autopsy shows birth defect (if applicable)	5236
Place of birth	3090
Folic acid details (peri-conceptional)	3085
H/O serious maternal illness	3124
H/O radiation exposure (x-ray)	3119
H/O substance abuse	3162
Assisted conception IVF/ART	3076
History of anomalies in previous pregnancies	3138
No. of previous abortion	3126
Birth defects	3124
White Blood cell count (thousand per microliter)	3118
Blood test result	3106
Symptom 1	3128
Symptom 2	3184
Symptom 3	3075
Symptom 4	3096
Symptom 5	3127
Genetic Disorder	3121
Disorder Subclass	3140

The strategies for handling the missing values are:

For handling numerical missing values we impute them with the median of their respective columns to preserve data integrity and minimize information loss.

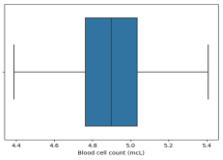
For handling categorical missing values we impute them with the mode of their respective columns to preserve data integrity & minimize information loss.

B. *Handling Outliers*



Our Dataset has only one attribute (Blood cell count (mcl)) with outliers. we are removing outliers by finding the whisker function that is by determining lower whisker & upper whisker.
Lower whisker = Q1-1.5*IQR
Upper whisker=Q3+1.5*IQR
After finding the Lower whisker & Upper whisker, any value that is below the lower whisker is replaced with the

lower whisker & any value that is above the Upper whisker is replaced by Upper whisker. This helps to ensure that extreme values don't skew our analysis.



C. *Renaming Attributes*

To improve clarity and consistency in the dataset, the following attributes were renamed as: Symptom 1 with Cardiomyopathy, Symptom 2 with Developmental Delays, Symptom 3 with Metabolic Issues, Symptom 4 with Asthma & Symptom 5 with Cystic fibrosis.

Attributes	Renamed Attributes	Attribute Description
Symptom 1	Cardiomyopathy	Refers to problem with you heart muscle
Symptom 2	Developmental Delays	Delays in reaching milestones such as walking or talking
Symptom 3	Metabolic Issues	Represents abnormal chemical reaction that occurs in the body to disrupt this reaction
Symptom 4	Asthma	Represents the respiratory issues that a patient had
Symptom 5	Cystic fibrosis	Represents the persistent lung infection, digestive issues and mucus build-up issues that a patient had.

We are also renaming the data values of attribute Genetic disorder as 'Mitochondrial genetic inheritance disorders: Mitochondrial GI disorders', 'Multifactorial genetic inheritance disorders: Multifactorial GI disorders', 'Single-gene inheritance diseases: Single-gene inheritance diseases' to improve clarity & consistency.

D. *Standardization/Normalization/String ProcessingR*

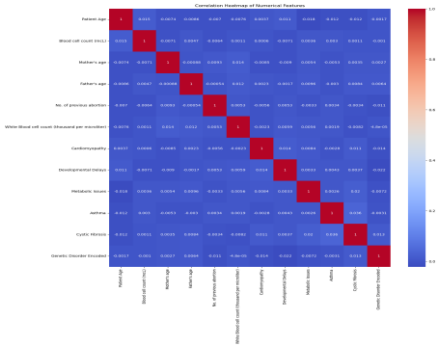
This is a crucial step in analysis. Here the important features (attributes) are selected and unimportant and irrelevant features are dropped. In our dataset there are several features (attributes) that do not contribute to gene disorder prediction.

The features (Attributes) like 'patient Id', 'patient first name', 'family name', 'father's name', 'institute name', 'location of institute', 'place of birth', 'parental consent' are dropped due to their low or no contribution in predicting the genetic disorders . The Data features (attributes) 'test 1','test 2','test 3','test 4', 'test 5' and 'autopsy shows birth defect (if applicable)' are dropped due to lower feature importance values (they have same values overall the column).

We have done encoding for disorder subclass. The Disorder subclass contain the values 'leber's hereditary optic neuropathy', 'Diabetes', 'Leigh syndrome', 'cancer', 'cystic fibrosis', 'Tay-sachs', 'hemochromatosis', 'Mitochondrial',' Alzheimer's' and are mapped by the values 0,1,2,3,4,5,6,7,8.We have also done encoding for Genetic disorder .The genetic disorder contains the values as Mitochondrial genetic inheritance disorder , Multifactorial genetic inheritance disorders & single-gene inheritance disorder and are mapped with the values 0,1,2.

Genetic Disorder	Disorder Subclass	Genetic Disorder Encoded	Disorder Subclass Encoded
Mitochondrial genetic inheritance disorders	Leber's hereditary optic neuropathy	0	5
Mitochondrial genetic inheritance disorders	Cystic fibrosis	0	2
Multifactorial genetic inheritance disorders	Diabetes	1	3
Mitochondrial genetic inheritance	Leigh syndrome	0	6

E. *Correlation Matrix*



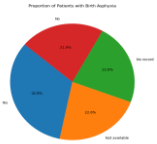
Correlation matrix is used to understand the relationship between different numerical variables. .

The correlation between the features(attributes) can be Positive, negative or zero. The value of one feature tends to increases as the value of one feature increases in positive correlation coefficient. And the value of one feature tends to decrease as the value of one feature increases in negative correlation coefficient. For example, there is a negative correlation between 'Patient age', and 'Mother's age' (-0.0074). It means that there is a weak tendency for mothers to be younger than their patients.

- Exploratory data analysis

D. *Univariate Analysis*

Univariate Analysis is the analysis where we focuses on understanding one variable at a time .Based on this analysis we are identifying any patterns , trends or anomalies in that particular variable.

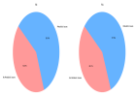


The above given pie chart shows the proportions of patients born with birth asphyxia .This analysis helps in understanding the frequency of patients born with birth asphyxia.

F. *Bivariate Analysis*

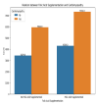
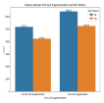
Bivariate analysis is the analysis where we focus on two variables at a time. Based on the analysis we are identifying any patterns, trends or anomalies between those variables.

The above Bar graph shows the distribution of genetic disorder based on whether they are inherited from mother's side or father's side. This graph shows that there is a higher percentage of genes that are inherited from fathers compared to mothers.

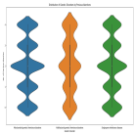


Identify applicable funding agency here. If none, delete this text box.

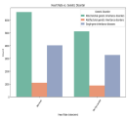
The above graph is double pie chart , it will show the count of offsprings with and without metabolic issues based on presence or absence of serious maternal illness.In the above pie chart the pie chart on the left shows that there are 58.4% of children born to mothers with a histroy of metabolic illness also have metabolic illness & the remaining 41.6% do not have metabolic issues. And the pie chart on the right shows that there are 57.9% of children born to mothers without a history of metabolic illness have metabolic illness & the remaining do not have metabolic illness.


(a)

(b)

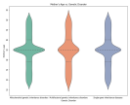
The above given bar plot (a) will show the count of offspring with or without cardic anomalies based on the level of the folic acid intake. This graph suggests that the follic acid supplements may be associated with a lower risk of developing cardiomyopathy.
The above given bar plot (b) will show the count of offsprings with or without Birth defects based on the level of the folic acid intake. This graph suggests that there are lesser birth defects among women who take follic acid supplements compared to those who did not take folic acid supplements.



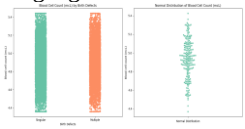
The above given violin graph shows the Distribution of genetic disorder by previous abortions . By analyzing the graph we can infer that there is significant Commonness of Mitochondrial Genetic inheritance disorder followed by single-gene inheritance diseases & Multifactorial genetic inheritance disorders.



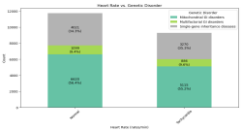
The above given Bar graph will show the distribution of different genetic disorder across various heart rate (normal,
Tachycardia) . By analyzing the counts we infer if certain genetic disorders are dominant in specific heart rates.



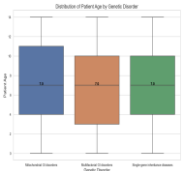
The above given violin graph shows the distribution of Mother's age by Genetic disorder . By analyzing the graph , we can infer there is a high chances of an offsprings to be genetically disabled if the mother's age lives in the range between (27-40) & the Mitochondrial genetic inheritence has high frequency followed by single gene inheritence disease & Multifactorial genetic disorder.



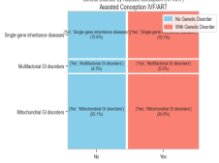
The above graph shows that there is a normal distribution of birth defects over blood cell count which means Blood cell count has no effect on birth defects.



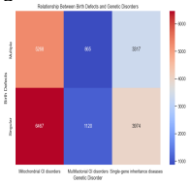
The above analysis shows that there is almost similar percentage of each genetic disorder in both normal heart rates & tachycardic (fast heart rates). Which shows that having a normal or fast heart rate doesn't seem to make a big difference in how often these genetic disorders occur.



The above given box plot shows the distribution of genetic disorder with patient age. By analysis the graph we can infer that the age group from (0-14) are affected by genetic disorder.



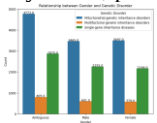
The above given graph shows that there is a significant difference in the areas of the mosaic tiles between categories would suggest that assisted conception methods may influence the likelihood of genetic disorders.



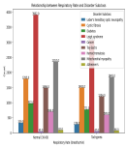
The above given graph shows the relationship between the birth defects & genetic disorder. It indicates If certain birth defects are more commonly associated with specific genetic disorders, it suggests potential links between these conditions.

G. Multivariate Analysis

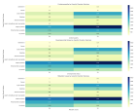
Multivariate Analysis is the analysis where we focus on more than two variables at a time. Based on the analysis we are identifying any trends, patterns or anomalies between those variables.



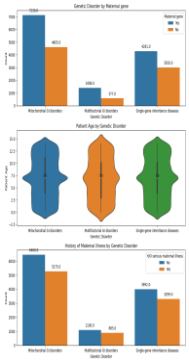
The above given bar graph will show how the gentic disorder diseases vary with gender . By analyzing the counts, we can infer that there is a notable prevalence of gentic disorder among male patients followed by females & ambiguous.



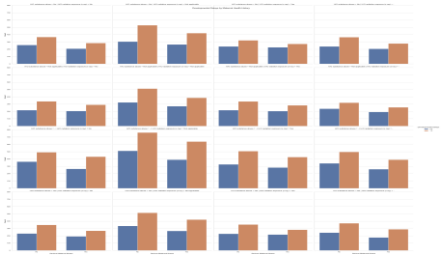
The above given Bar graph will show the distribution of different genetic disorder across various respiratory rates
(normal(30-60) , tachypnea) . This graph suggests , by analyzing the counts , we can infer whether certain genetic disorders are Dominant in specific respiratory categories.



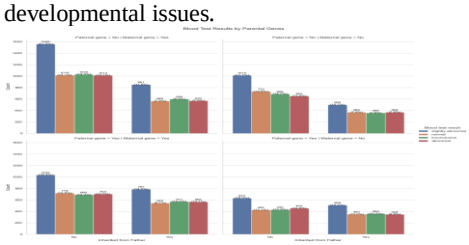
The above heatmaps provide a clear visualization how different genetic disorder subclasses are linked with specific health issues.



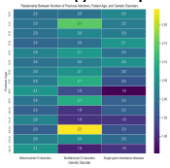
The above bar graph, violin graph & bar graph shows the distribution of genetic disorder by maternal gene, patient age & history of maternal illness respectively. All the three graphs infer that there are high chances of mitochondrial genetic disorder followed by single-gene inheritance & multifactorial genetic disorder.



The above graph shows if certain combinations of maternal health factors show higher incidences of developmental delays, it could highlight the importance of maternal health in preventing developmental issues.



The above shows how inherited from father's side, maternal gene and paternal gene collectively influence blood test results. By analyzing the graph, we infer that in all the cases paternal genes have high frequency compared to maternal genes.



The above given correlated graph indicates that if certain age groups exhibit higher mean numbers of previous abortions for specific genetic disorders, it suggests a potential association between the number of previous abortions, patient age, and the occurrence of genetic disorders. Further-more analysis and investigation are necessary to understand the underlying factors driving these associations.

III. CONCLUSION

In this exploratory data analysis (EDA) project, we examined the dataset on genetic disorders to uncover significant patterns and insights.

Our analysis revealed several key findings:

1.Inheritence patterns: The analysis of distribution of genetic disorder based on whether they are inherited from father's side or mother's side. This analysis suggests that there is a higher percentage of genes that are inherited from father's side compared to mother's side.

2. Impact of Maternal Illness on Offspring's Metabolic Health**:** This analysis shows the count of offspring with or without cardic anomalies & Birth defects based on the level of the folic acid intake.This analysis suggests that the follic acid supplements may be associated with a lower risk of developing cardiomyopathy and birth defects.

3.Distribution of genetic disorder across various respiratory rates : This analysis suggests which genetic disorders are Dominant in specific respiratory categories.

4.Gender difference: This genetic analysis shows how the distribution of genetic disorder vary with gender. This analysis suggests that there is a notable prevalence of gentic disorder among male patients followed by females & ambiguous.

5.Distribution of genetic disorder by previous abortions:This analysis suggests that there is significant Commonness of Mitochondrial Genetic inheritance disorder followed by single-gene inheritance diseases & Multifactorial genetic inheritance disorders.

6.Distribution & correlation between different genetic disorders with specific issues: This analysis provides a clear visualization how different genetic disorder subclasses are linked with specific health issues.

7.Distribution of genetic disorder across various heart rates:This analysis suggests how certain genetic disorders are dominant in specific heart rates.

8.Mother's age & genetic disorder: This analysis suggest how there is a high chances of an offsprings to be genetically disabled if the mother's age lives in the range between (27-40).

9.Distribution of birth defects with blood cell count: This analysis shows that having a normal or fast heart rate doesn't seem to make a big difference in how often these genetic disorders occur.

10.Distribution of genetic disorder by patients age: By analysis the graph we can infer that the age group from (0-14) are affected by genetic disorder.

11.Relationship between birth defects & genetic disorder: By analyzing the graph, it indicates If certain birth defects are more commonly associated with specific genetic disorders, it suggests potential links between these conditions.

12.Significant association between the presence of specific genetic disorders and the age of the patient, maternal gene involvement, and the history of maternal illnesses: By analyzing the graph, it infers that there are high chances of mitochondrial genetic disorder followed by single-gene inheritance & multifactorial genetic disorder.

13.Development delay: The graph indicates if certain combinations of maternal health factors show higher incidences of developmental delays, it could highlight the importance of maternal health in preventing developmental issues.

14.Influence on blood test results: The analysis of this graph shows how inherited from father's side, maternal gene and paternal gene collectively influence blood test results. By analyzing the graph, we infer that in all the cases paternal genes have high frequency compared to maternal genes.

15.Correlation between genetic disorder , pervious abortions & patients age: The analysis of correlated graph indicates that if certain age groups exhibit higher mean numbers of previous abortions for specific genetic disorders, it suggests a potential association between the number of previous abortions, patient age, and the occurrence of genetic disorders.

The insights gained from this EDA can inform healthcare professionals and policymakers about the key areas to focus on for early detection and intervention strategies. For future work, applying machine learning techniques to predict the likelihood of

developing specific disorders could enhance early diagnosis and personalized treatment plans.

In conclusion, this EDA has provided a foundational understanding of the patterns and trends in genetic disorders within the dataset. These findings will provide a way for the future analysis.

REFERENCES

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9858679/
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9173933/
3. Frontiers | The Future of Genetic Disease Studies: Assembling an Updated Multidisciplinary Toolbox (frontiersin.org)
4. https://www.sciencedirect.com/science/article/pii/S2352396421001158