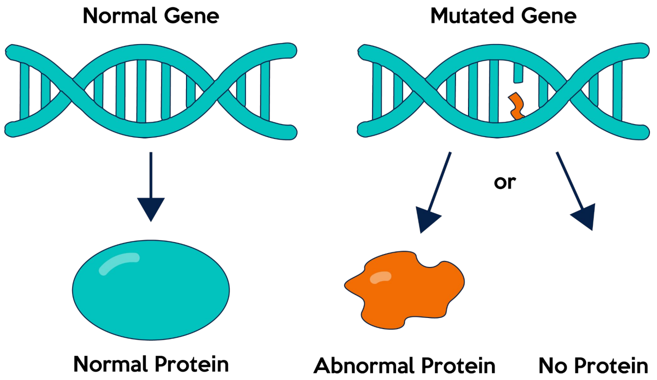
SC5010 Mini-Project

Guan Jia Sheng (U2040851A)
Nguyen Ngoc Minh Truc (U1940862C)
Lim Yong Yee (U2040881F)

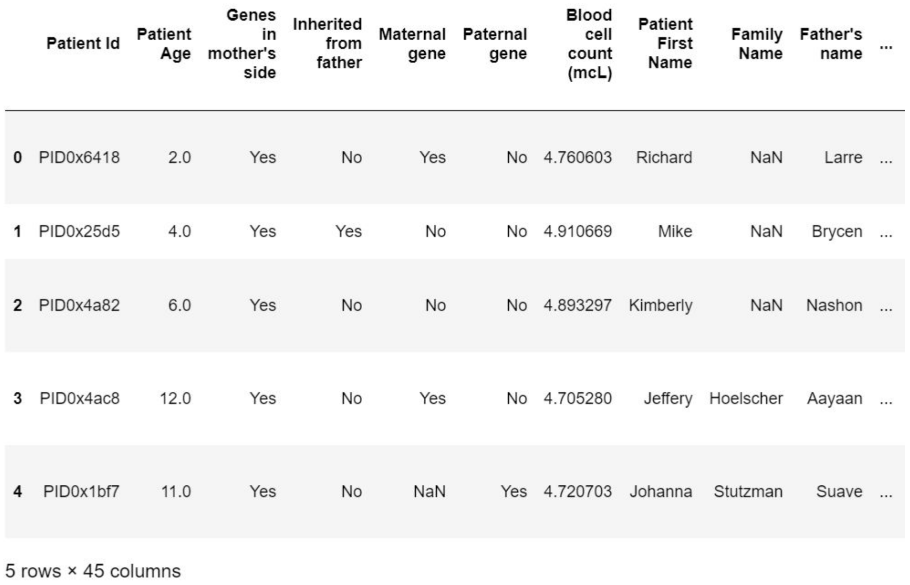
Genetic Disorders

- Genetic disorders are a leading cause of pediatric and infant deaths
- Mutations in our DNA cause changes in proteins and lead to genetic disorders
- Testing for genetic disorders require expensive sequencing
- Dataset: Genetic Disorders dataset from kaggle
- Aim: Predict genetic disorders based on readily available data



Genetic Disorder Dataset

- **Original: 22083 rows x 45 columns**
- 2 response variables:
  - 3 Genetic Disorders
    - Mitochondrial
    - Single-gene
    - Multifactorial
  - Each with 3 Disorder Subclasses
    - 9 subclasses

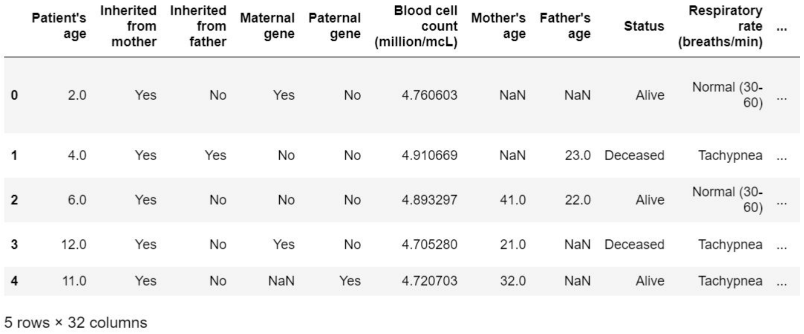


	Patient Id	Patient Age	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	Blood cell count (mcL)	Patient First Name	Family Name	Father's name	...
0	PID0x6418	2.0	Yes	No	Yes	No	4.760603	Richard	NaN	Larre	...
1	PID0x25d5	4.0	Yes	Yes	No	No	4.910669	Mike	NaN	Brycen	...
2	PID0x4a82	6.0	Yes	No	No	No	4.893297	Kimberly	NaN	Nashon	...
3	PID0x4ac8	12.0	Yes	No	Yes	No	4.705280	Jeffery	Hoelscher	Aayaan	...
4	PID0x1bf7	11.0	Yes	No	NaN	Yes	4.720703	Johanna	Stutzman	Suave	...

5 rows × 45 columns

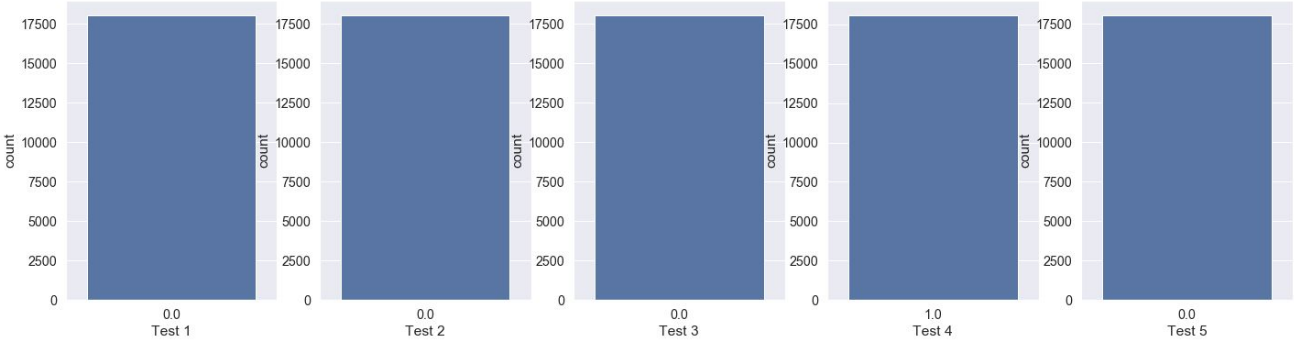
Data Cleaning

- Drop missing Disorder Subclass
- Fill missing classes based on subclasses
- Drop irrelevant columns:
  - Names
  - Locations
  - Test 1–5
- **Final: 19915 rows x 32 columns**



	Patient's age	Inherited from mother	Inherited from father	Maternal gene	Paternal gene	Blood cell count (million/mcL)	Mother's age	Father's age	Status	Respiratory rate (breaths/min)	...
0	2.0	Yes	No	Yes	No	4.760603	NaN	NaN	Alive	Normal (30-60)	...
1	4.0	Yes	Yes	No	No	4.910669	NaN	23.0	Deceased	Tachypnea	...
2	6.0	Yes	No	No	No	4.893297	41.0	22.0	Alive	Normal (30-60)	...
3	12.0	Yes	No	Yes	No	4.705280	21.0	NaN	Deceased	Tachypnea	...
4	11.0	Yes	No	NaN	Yes	4.720703	32.0	NaN	Alive	Tachypnea	...

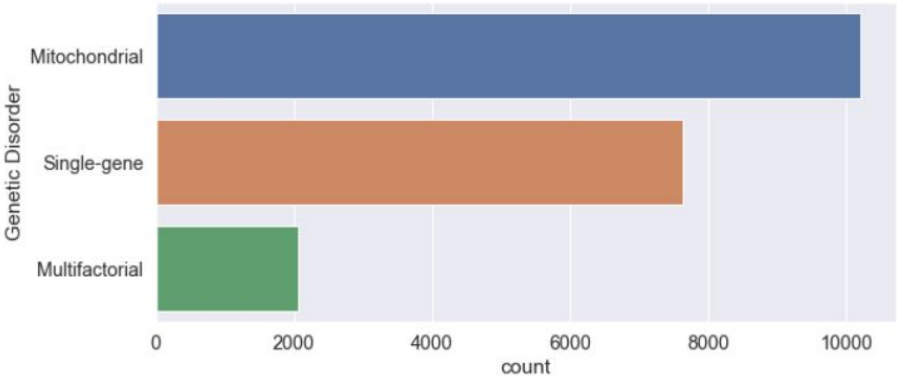
5 rows × 32 columns

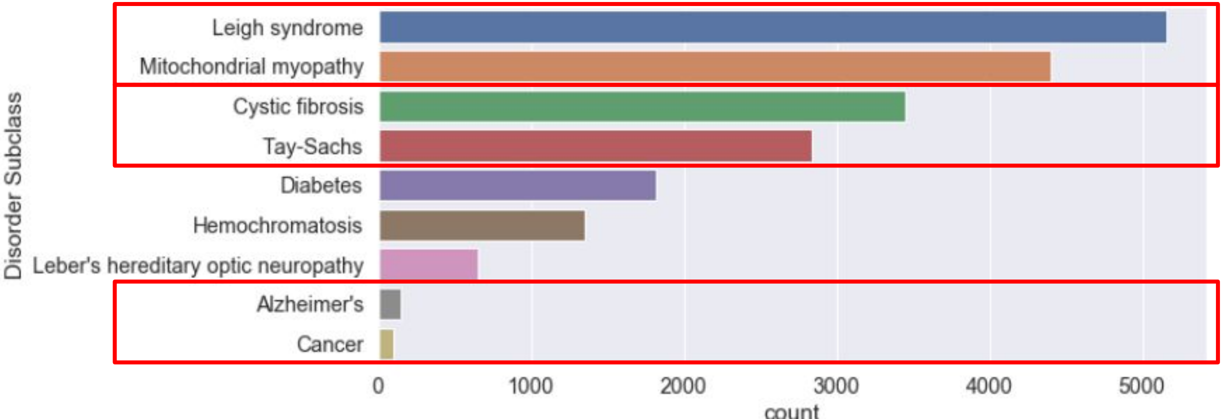


01

Exploratory Data Analysis

Genetic Disorder and Disorder Subclass





Mitochondrial

Single-gene

Multifactorial

Predictors

Patient observations

Patient's age
Gender
Blood test result
...

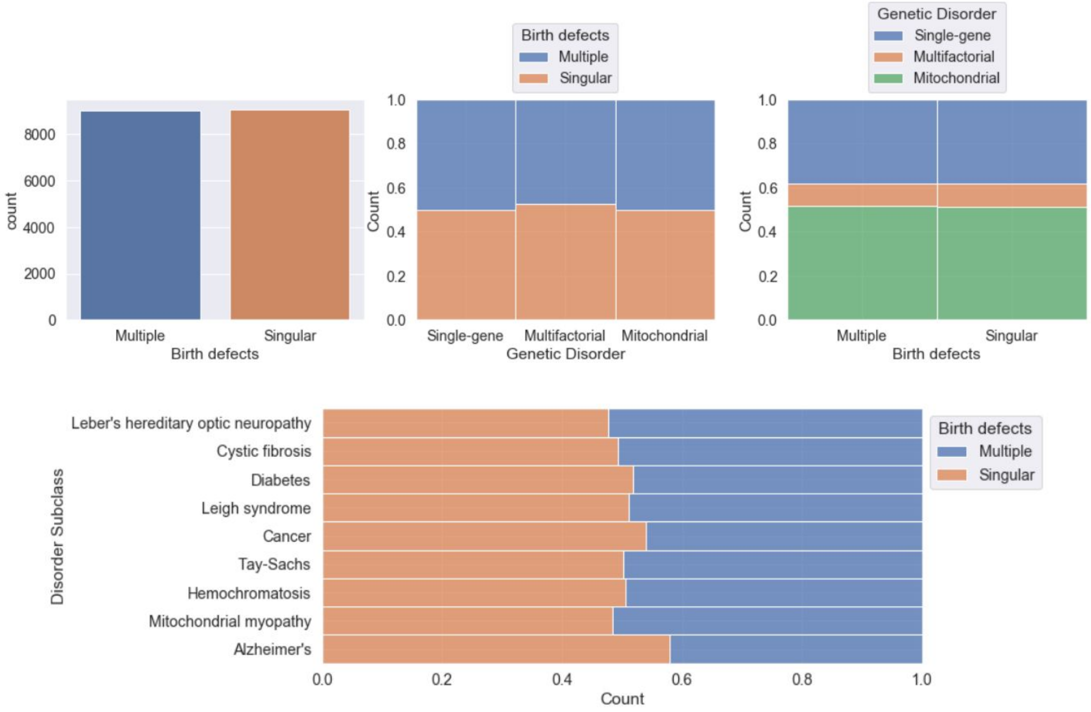
Inheritance factors

Inherited from mother
Maternal gene
Inherited from father
Paternal gene
...

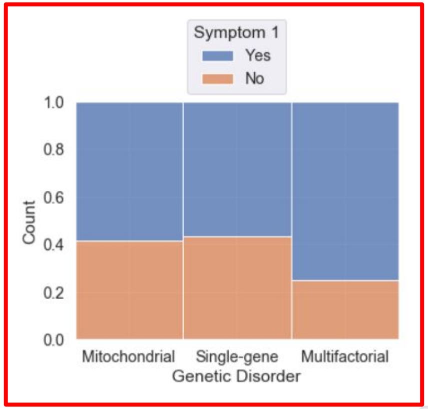
Mother's conditions

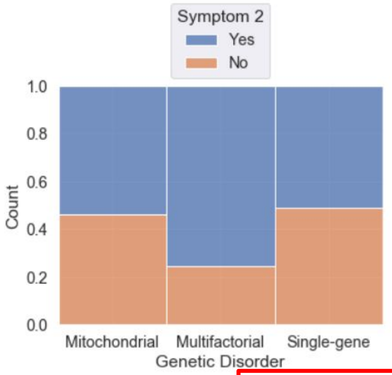
Assisted conception
IVF/ART
H/O anomalies in previous pregnancies
No. of previous abortions
...

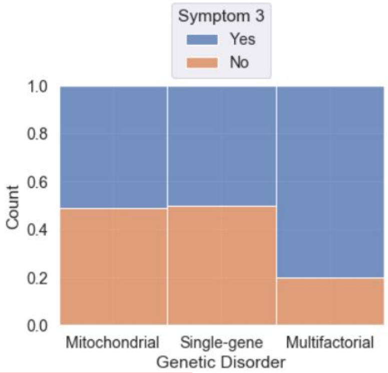
Most variables do not show influence on genetic disorders

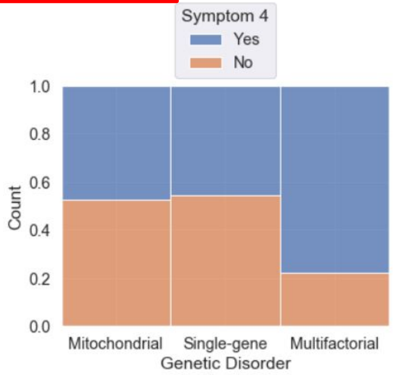


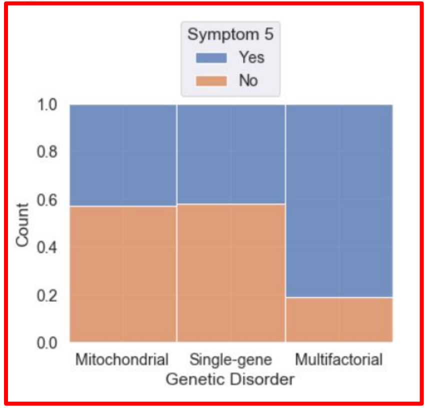
Symptoms


Symptom 1
Yes
No
1.0
0.8
0.6
0.4
0.2
0.0
Count
Mitochondrial
Single-gene
Multifactorial
Genetic Disorder

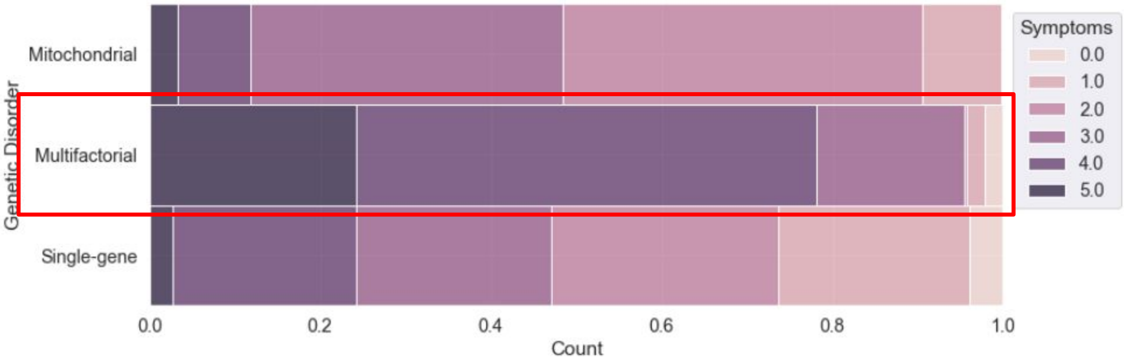

Symptom 2
Yes
No
1.0
0.8
0.6
0.4
0.2
0.0
Count
Mitochondrial
Multifactorial
Single-gene
Genetic Disorder

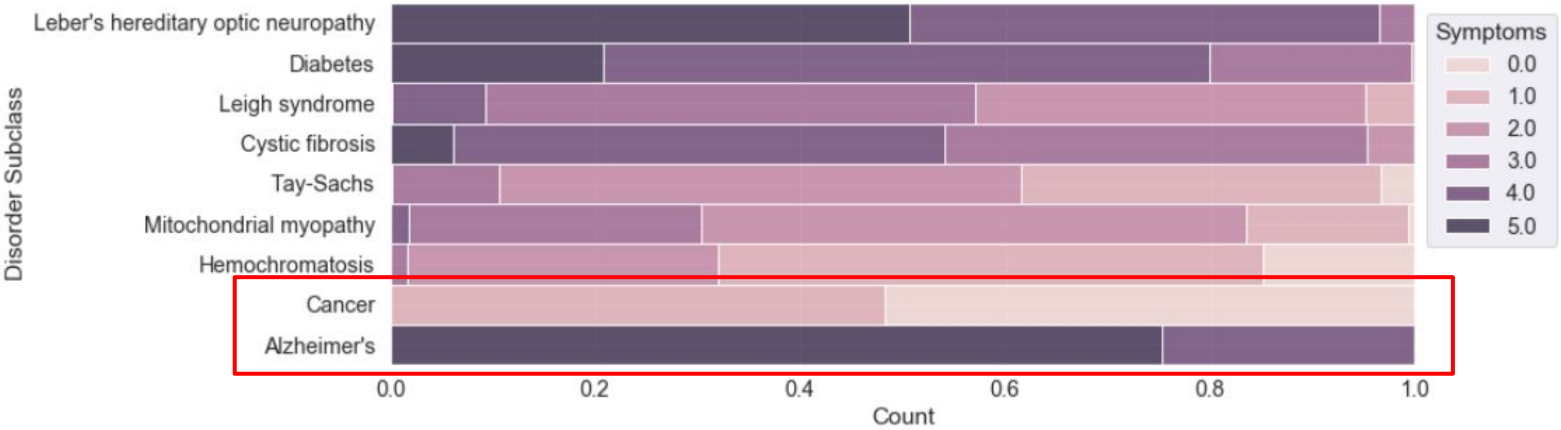

Symptom 3
Yes
No
1.0
0.8
0.6
0.4
0.2
0.0
Count
Mitochondrial
Single-gene
Multifactorial
Genetic Disorder


Symptom 4
Yes
No
1.0
0.8
0.6
0.4
0.2
0.0
Count
Mitochondrial
Single-gene
Multifactorial
Genetic Disorder

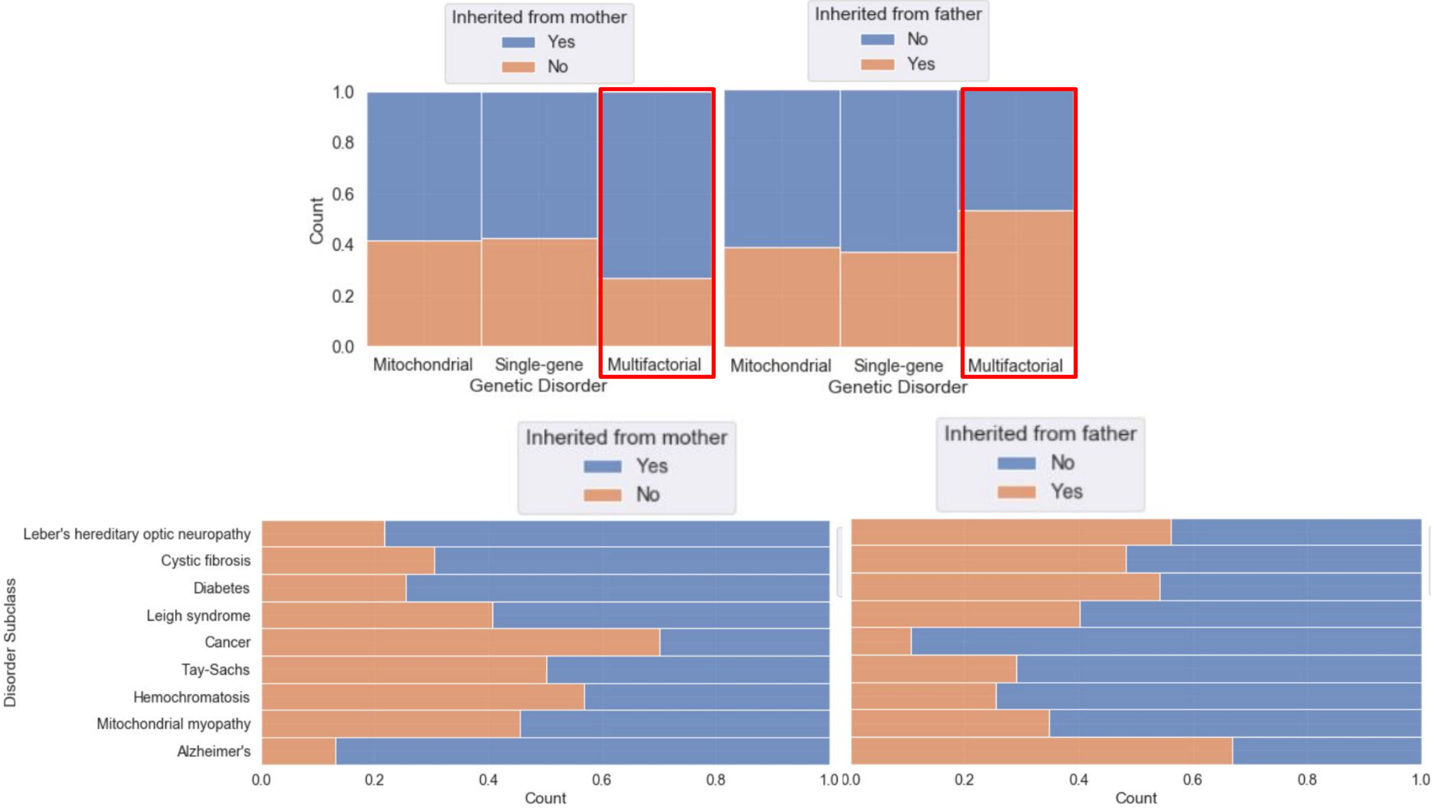

Symptom 5
Yes
No
1.0
0.8
0.6
0.4
0.2
0.0
Count
Mitochondrial
Single-gene
Multifactorial
Genetic Disorder

Symptoms

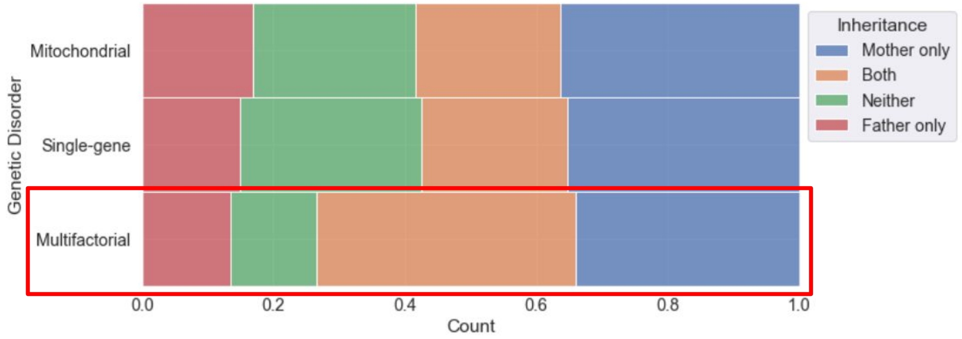

Genetic Disorder
Mitochondrial
Multifactorial
Single-gene
0.0
0.2
0.4
0.6
0.8
1.0
Count
Symptoms
0.0
1.0
2.0
3.0
4.0
5.0

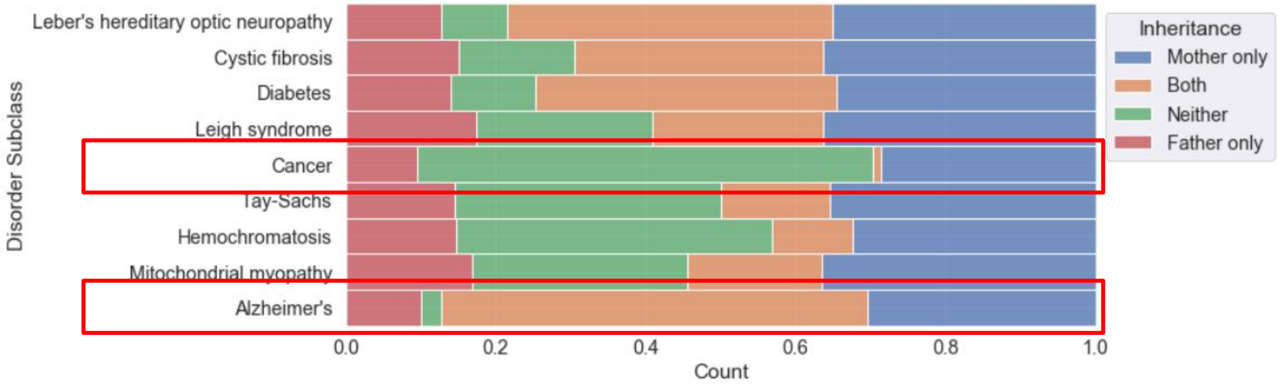

Disorder Subclass
Leber's hereditary optic neuropathy
Diabetes
Leigh syndrome
Cystic fibrosis
Tay-Sachs
Mitochondrial myopathy
Hemochromatosis
Cancer
Alzheimer's
0.0
0.2
0.4
0.6
0.8
1.0
Count
Symptoms
0.0
1.0
2.0
3.0
4.0
5.0

Inheritance



Inheritance



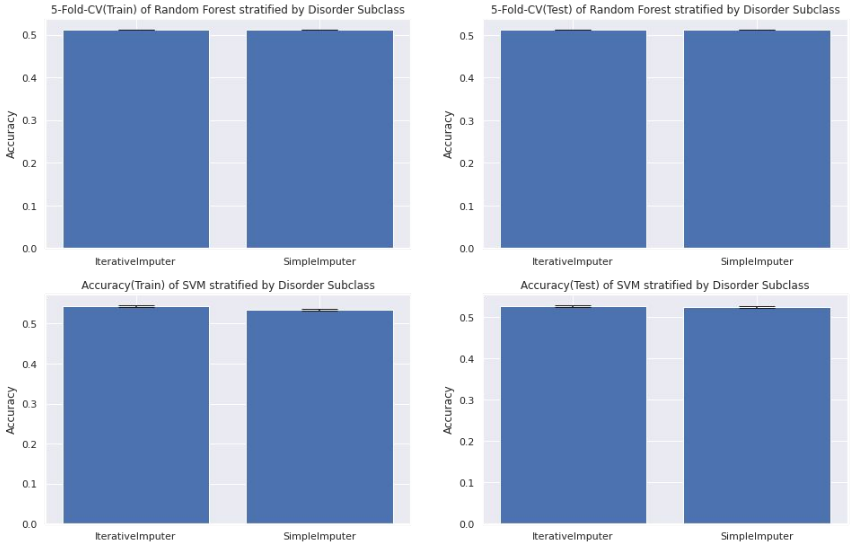


02

Core Analysis

Comparison: Missing Value Imputation

- SimpleImputer – mode of each column
- IterativeImputer – model each predictor as a function of other predictors



5-Fold-CV(Train) of Random Forest stratified by Disorder Subclass

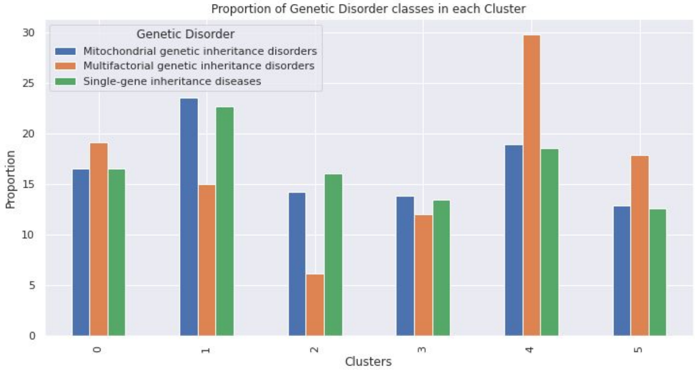
5-Fold-CV(Test) of Random Forest stratified by Disorder Subclass

Accuracy(Train) of SVM stratified by Disorder Subclass

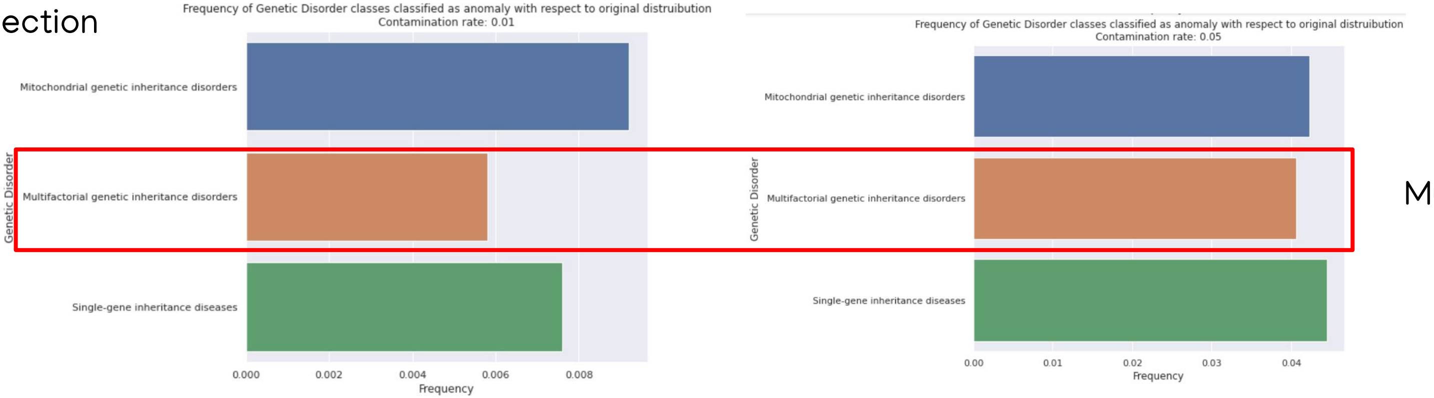
Accuracy(Test) of SVM stratified by Disorder Subclass

Unsupervised Learning

KModes Clustering

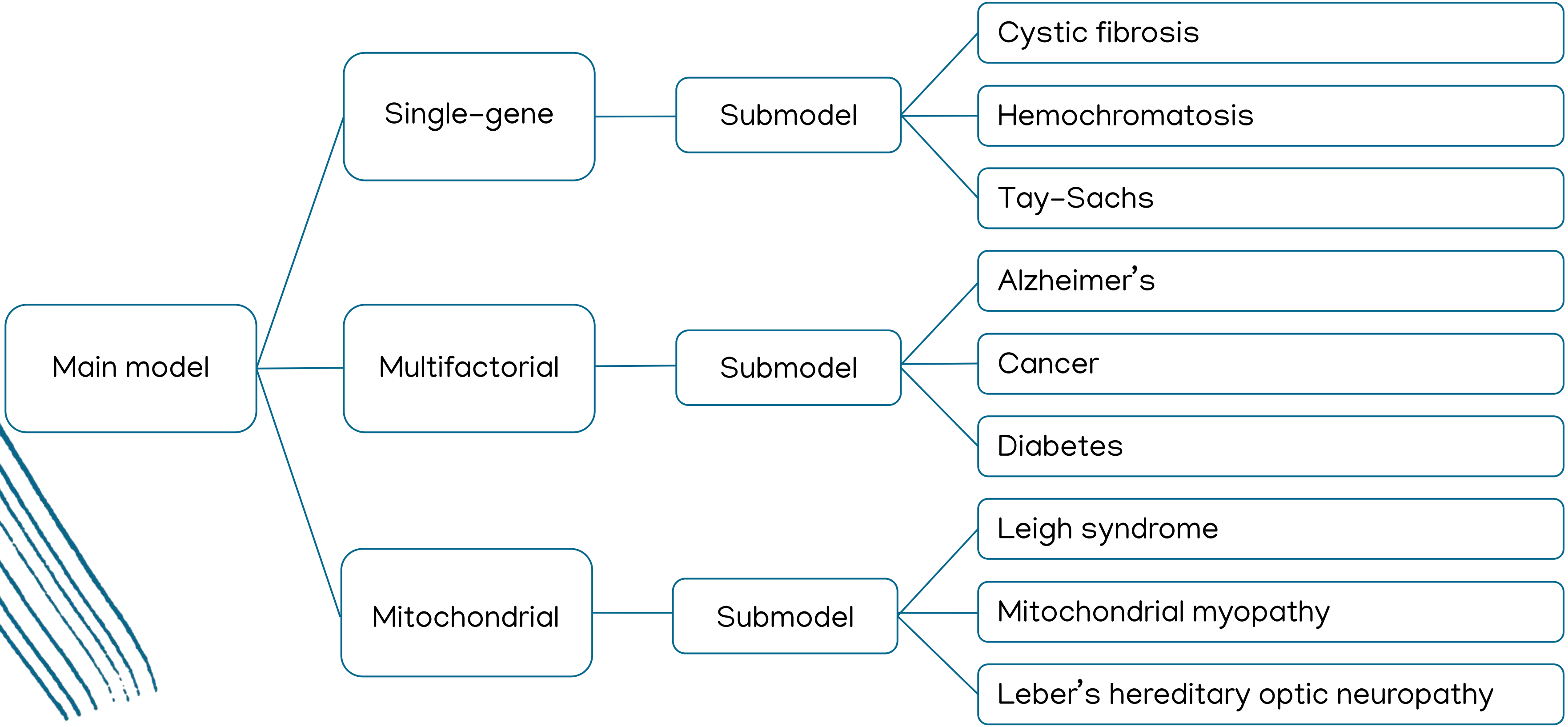


kNN Anomaly Detection



Multifactorial

Supervised Learning



Main model

Single-gene

Submodel

Cystic fibrosis

Hemochromatosis

Tay–Sachs

Multifactorial

Submodel

Alzheimer's

Cancer

Diabetes

Mitochondrial

Submodel

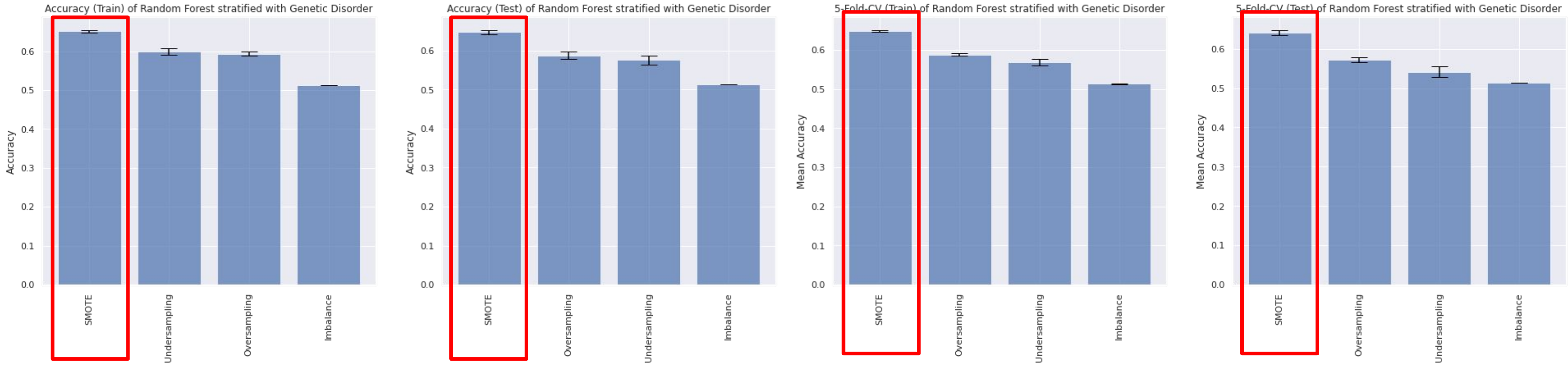
Leigh syndrome

Mitochondrial myopathy

Leber's hereditary optic neuropathy

Comparison: Class Imbalance Treatment

Accuracy



Train

Test

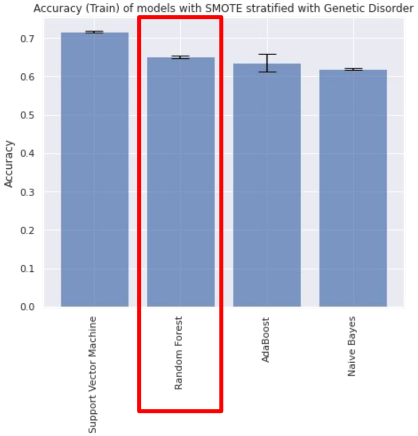
Train
5–fold CV

Test
5–fold CV

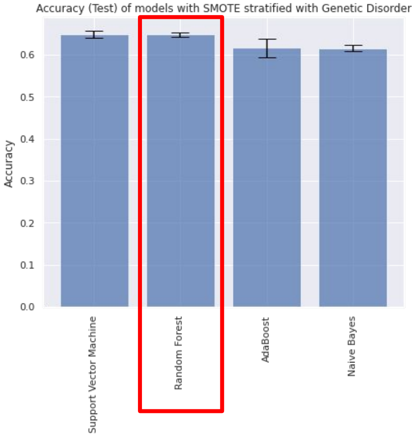
- SMOTE (red box) > Oversampling > Undersampling > Imbalance
- SMOTE is used for final models

Comparison: Classification Model Performance

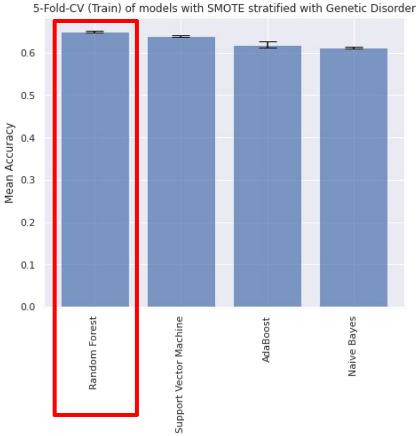
Accuracy



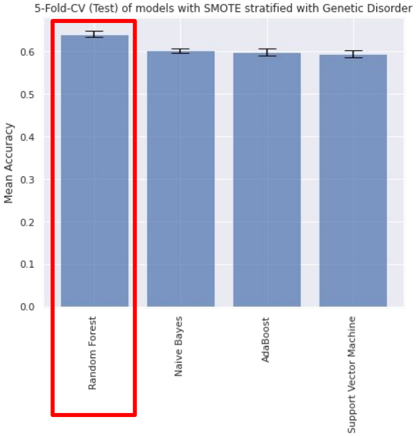
Train



Test



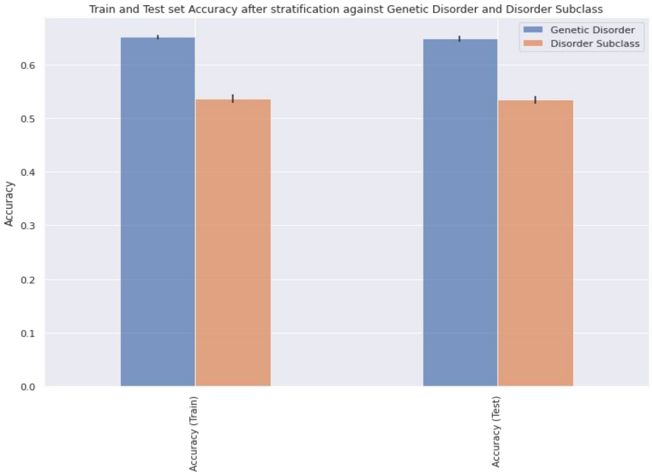
Train
5-fold CV

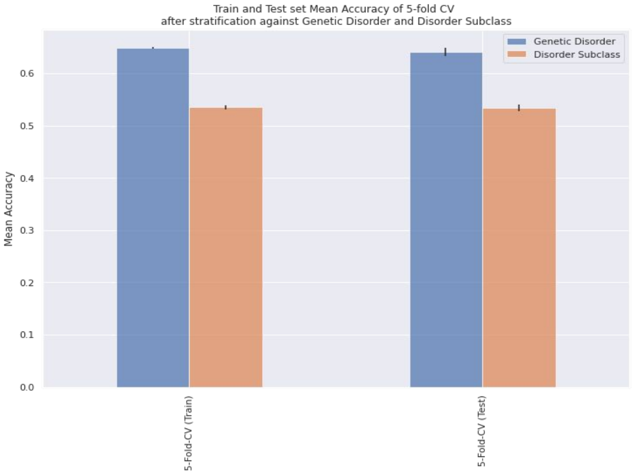


Test
5-fold CV

- Support Vector Machine shows signs of overfit
- Random Forest (red box) shows more consistent performance

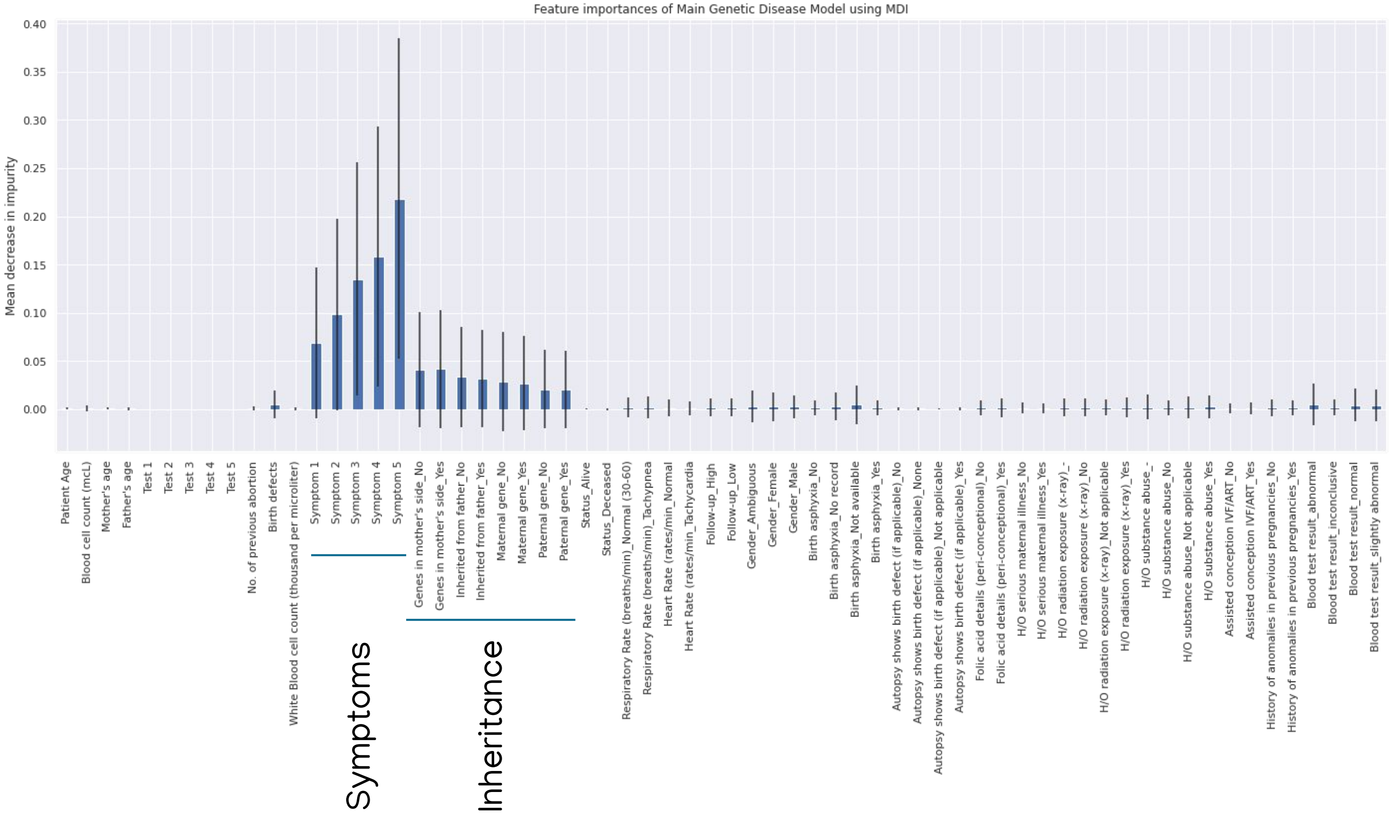
Comparison: Stratification of Classes





- Better performance for Genetic Disorder model when stratified against Genetic Disorder
- Final model stratified against Disorder Subclass to keep same train and test sets for all models

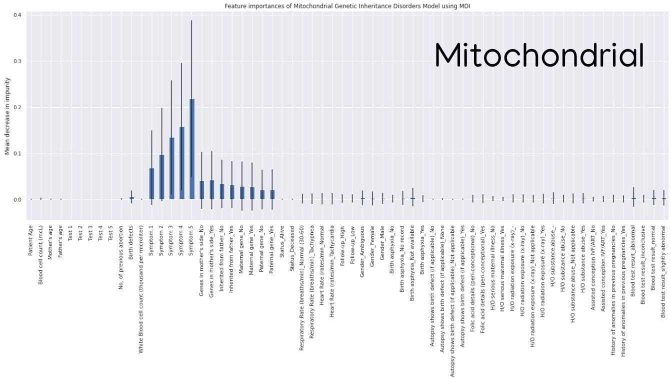
Variable Importance



Symptoms

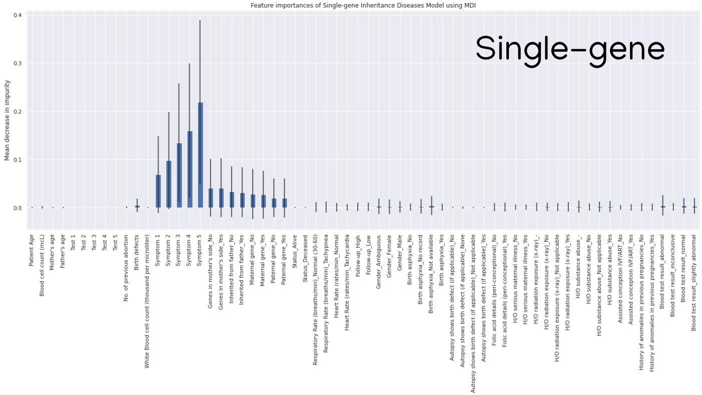
Inheritance

Variable Importance



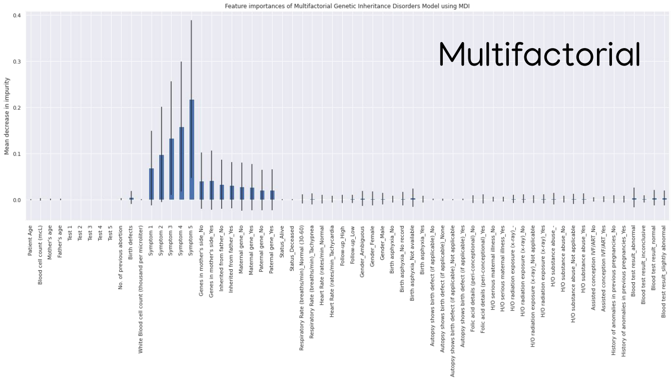
Feature importances of Mitochondrial Genetic Inheritance Disorders Model using MDI

Mitochondrial



Feature importances of Single-gene Inheritance Diseases Model using MDI

Single-gene

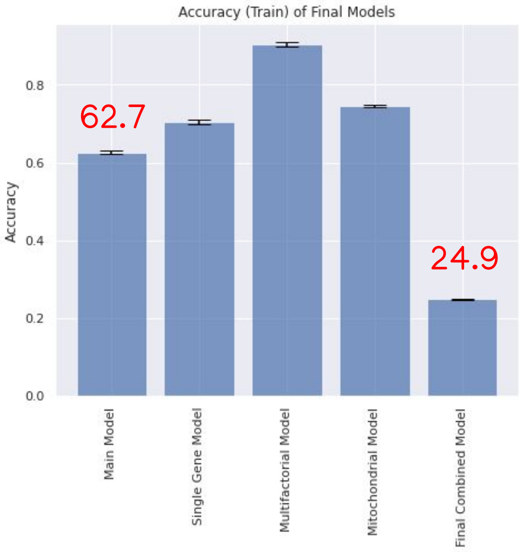


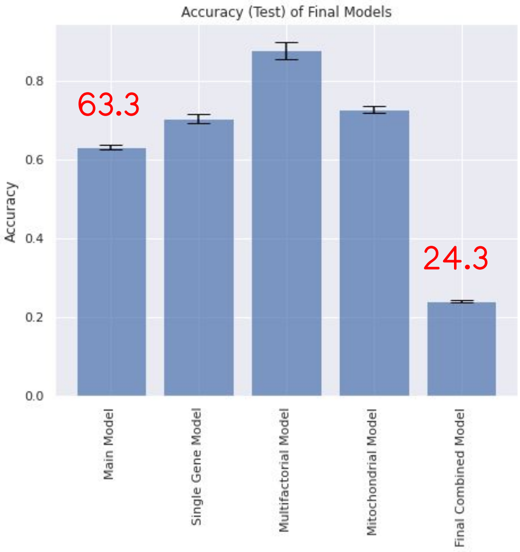
Feature importances of Multifactorial Genetic Inheritance Disorders Model using MDI

Multifactorial

Final Model Accuracy

- IterativeImputer to fill missing values
- SMOTE imbalance treatment
- Disorder Subclass stratification
- 9 features: 5 symptoms + 4 inheritance factors
- Random Forest classification model

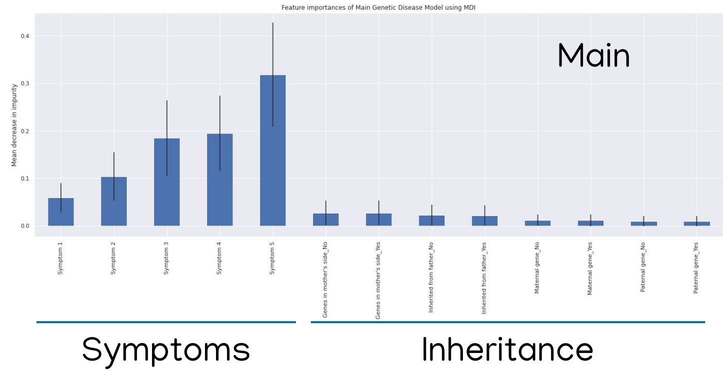


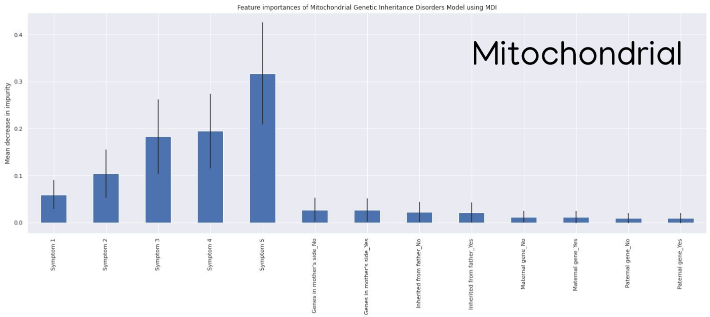


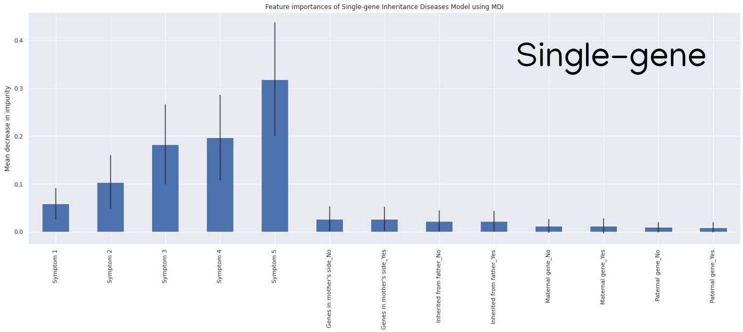
03

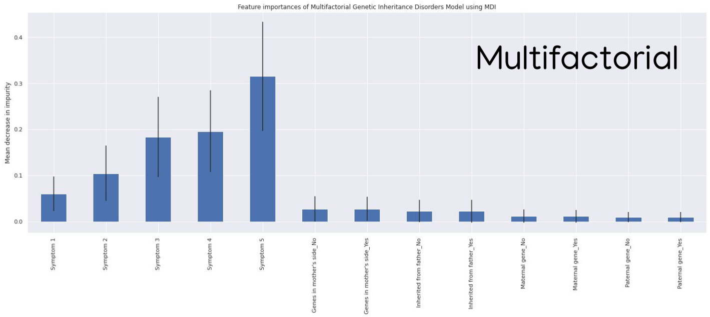
Conclusion

Variable Importance in Reduced Complexity Models


Feature importances of Main Genetic Disease Model using MDI

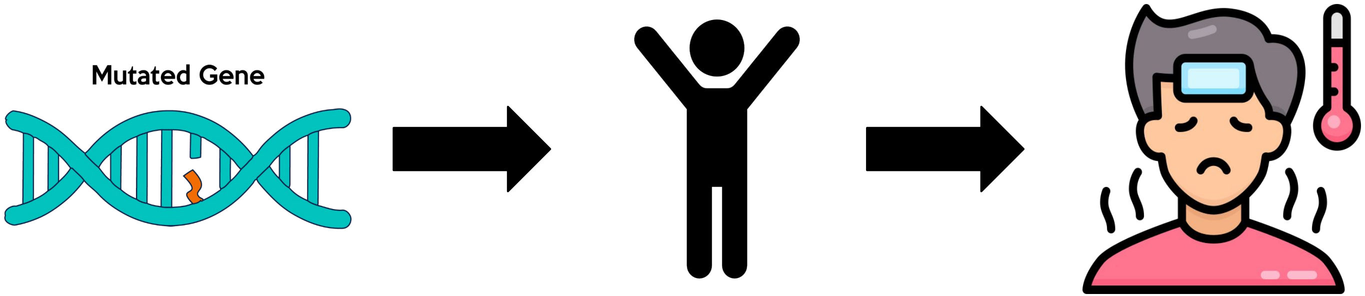

Feature importances of Mitochondrial Genetic Inheritance Disorders Model using MDI


Feature importances of Single-gene Inheritance Diseases Model using MDI


Feature importances of Multifactorial Genetic Inheritance Disorders Model using MDI

Conclusions

- Inheritance factors and Symptoms are important in determining Genetic Disorder and Disorder Subclass values
- Important variables same as what was found from EDA
- Exact symptoms and genes involved were not defined, difficult to make scientific connections



Thank you