

Predicting Genetic Disorders Using Decision Tree Model with Bayesian Network

By Bridget Manu

Outline

- Introduction
 - Background
 - Research Question
 - Data Collection
 - Model Building
 - Model Evaluation
 - Results and Discussion
 - Conclusion
-

Genetic Disorder

Definition

- Caused by a mutation in the deoxyribonucleic acid or change in number or structure of chromosomes.
 - Mutation can occur as a result of:
 - environmental factors
 - genetic inheritance
 - chromosomal damage
-

Impact on Health

- Early prediction may help in early intervention and treatment
- Provide valuable information for patients and family to make informed decision about their health and lifestyle
- Opportunity for genetic counselling

Decision Tree Model

Decision trees are non-parametric supervised learning model for classification and regression.

Aim: Create a model to predict the value of a target variable based values of the predictor variables

- Pros:

- simple to understand and interpret
- trees can be visualized
- require little preparation of data
- Performs well even if assumptions are somehow violated

- Cons:

- can cause overfitting
 - can be unstable due to small variation in data
 - can create biased trees if certain classes dominate
-

To use Decision Tree Model to
predict genetic disorders
from the given dataset

Goal

Data Features

- The Data consisted of a train dataset and test dataset
 - It contained medical information of children with genetic disorders
 - The patient's age ranges from 0 – 14 years old
 - Raw Train data : 45 variables; 22083 observations
 - Raw Test data: 43 variables; 9463 observations
-

Data was collected from 27 different hospitals across the US

- 3 different genetic disorders

Cleaned Data: 43 variables; 10410 observations

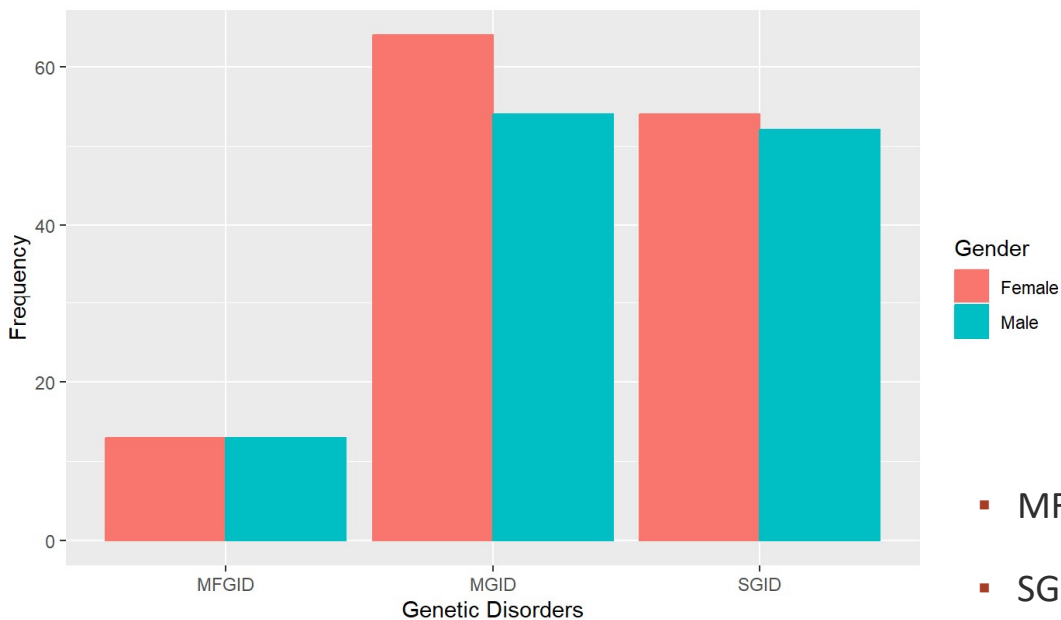
- Converted characters & integer variables into factors
- Replaced missing data with NA

Feature Selection

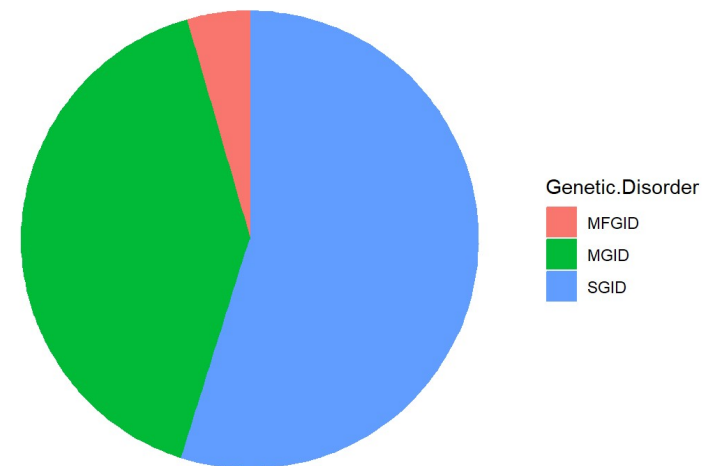
- Bayesian Network – Conditional Probability Table (CPT)
 - Using CPT & Bayesian network, 11 variables were selected
 - Updated cleaned data: 11 variables; 10410 observations
-

Data Visualization

Frequency of each Genetic disorder among the Genders



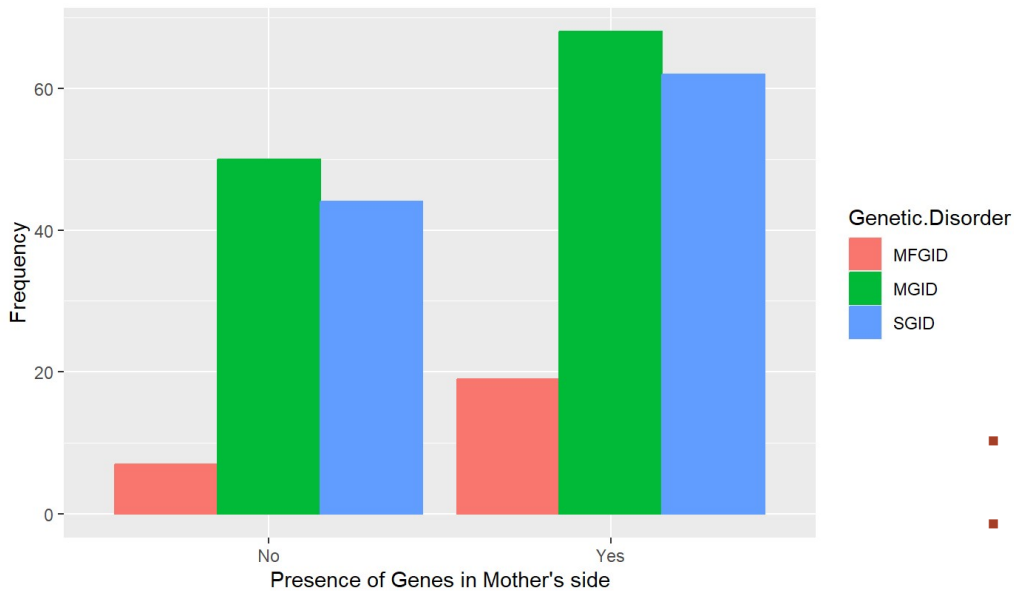
Pie Chart of the Proportion of the Genetic Disorders



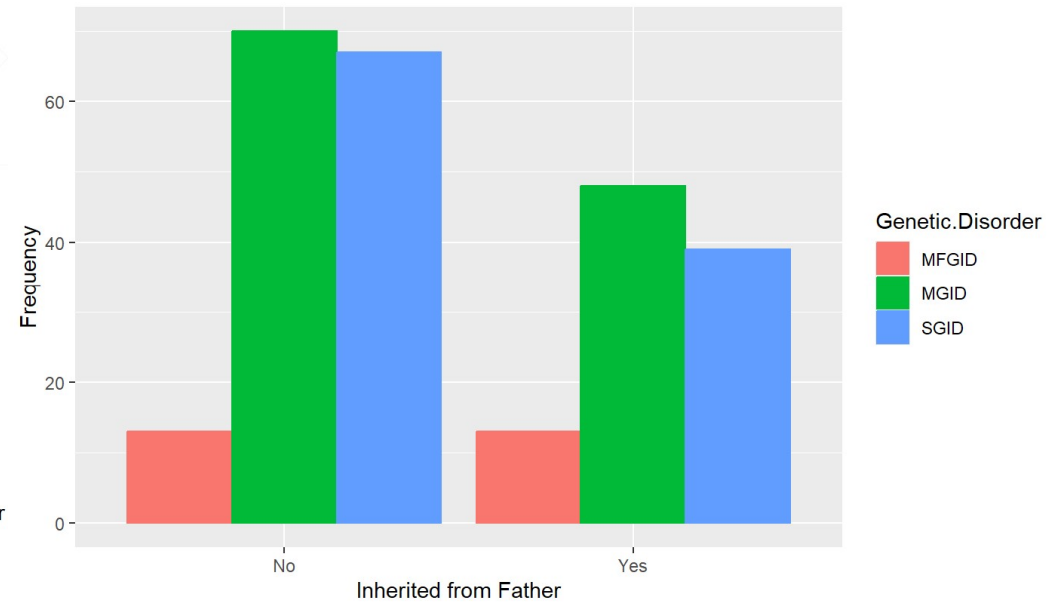
- MFGID - Multifactorial genetic inheritance disorders
- SGID - Single-gene inheritance diseases
- MGID - Mitochondrial genetic inheritance disorders

Data Visualization

Bar Graph of Frequency of each Disorders vs. Presence



Bar Graph of Frequency of each Disorders vs. Presence



- MFGID - Multifactorial genetic inheritance disorders
- SGID - Single-gene inheritance diseases
- MGID - Mitochondrial genetic inheritance disorders

- To develop the model:
 - 80% of dataset to train the model
 - Train Data: 11 variable; 8376 observations.
 - Subset of relevant variables for predicting the target variable
 - Genes.in.mother.s.side; Inherited.from.father; Maternal.gene ; Paternal.gene; Symptom.1; Symptom.2; Symptom.3; Symptom.4; Symptom.5
 - Tuned hyperparameters to avoid overfitting;

Model Building:

Decision Tree

- Learns simple decision rules based on the data features
- Each node on the tree is a data feature that plays a role in the outcome
- The leaf nodes are the outcomes -> targeted variable

- To test the model,
 - Test data: 9 variables; 5318 observations
 - Test Data: 11 variable; 2034 observations (20%)
- Confusion Matrix (Error Matrix) comprises of:
 - Accuracy of the model
 - Confidence Interval of the accuracy
 - No Information Rate: Baseline accuracy
 - P-value [Acc > NIR]: evaluates significance of model compared to NIR
 - Kappa: agreement between predictions and actual values. 1 = perfect; 0 = most likely a chance
 - MT P-value: Biased or not; small p-value indicates strong evidence of model bias

Result:

Confusion Matrix (Error Matrix)

Provides thorough analysis of true positive, true negative, false negative and false positive predictions.

Confusion Matrix and Statistics

| Reference | | | |
|------------|-------|------|------|
| Prediction | MFGID | MGID | SGID |
| MFGID | 71 | 47 | 48 |
| MGID | 101 | 846 | 527 |
| SGID | 40 | 135 | 219 |

Overall Statistics

Accuracy : 0.5585
95% CI : (0.5366, 0.5802)
No Information Rate : 0.5054
P-Value [Acc > NIR] : 9.053e-07

Kappa : 0.1967

McNemar's Test P-Value : < 2.2e-16

Statistics by Class:

| | Class: MFGID | Class: MGID | Class: SGID |
|----------------------|--------------|-------------|-------------|
| Sensitivity | 0.33491 | 0.8230 | 0.2758 |
| Specificity | 0.94786 | 0.3757 | 0.8589 |
| Pos Pred Value | 0.42771 | 0.5739 | 0.5558 |
| Neg Pred Value | 0.92452 | 0.6750 | 0.6494 |
| Prevalence | 0.10423 | 0.5054 | 0.3904 |
| Detection Rate | 0.03491 | 0.4159 | 0.1077 |
| Detection Prevalence | 0.08161 | 0.7247 | 0.1937 |
| Balanced Accuracy | 0.64138 | 0.5994 | 0.5673 |

Confusion Matrix

Sensitivity: proportion of actual (+) that are correctly identified

Specificity: : proportion of actual (-) that are correctly identified

Pos Pred Value: proportion of (+) predictions that are correct

Neg Pred Value: proportion of (-) predictions that are correct

Prevalence: proportion that belongs to (+ class

Detection Rate: # of true (+) predictions by model/ total # of instances

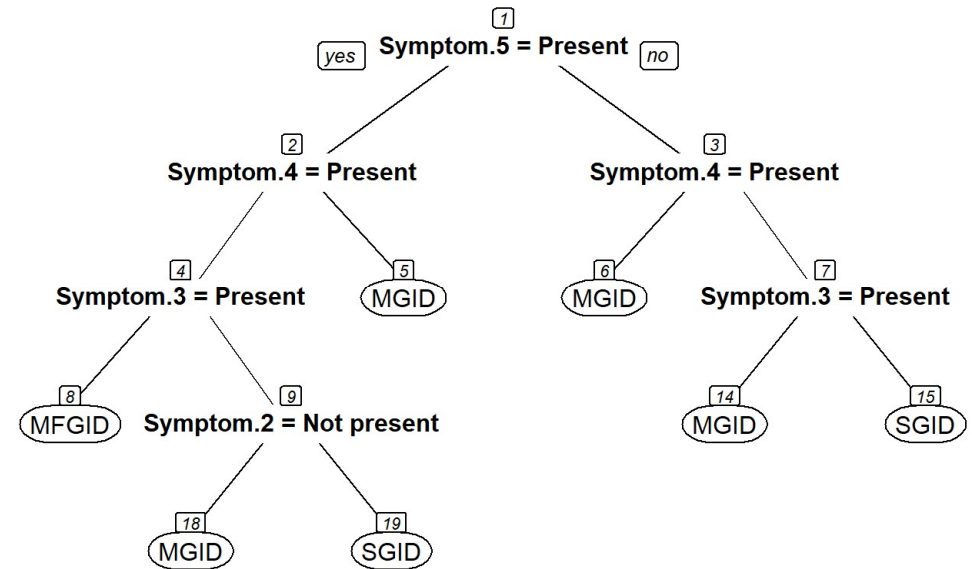
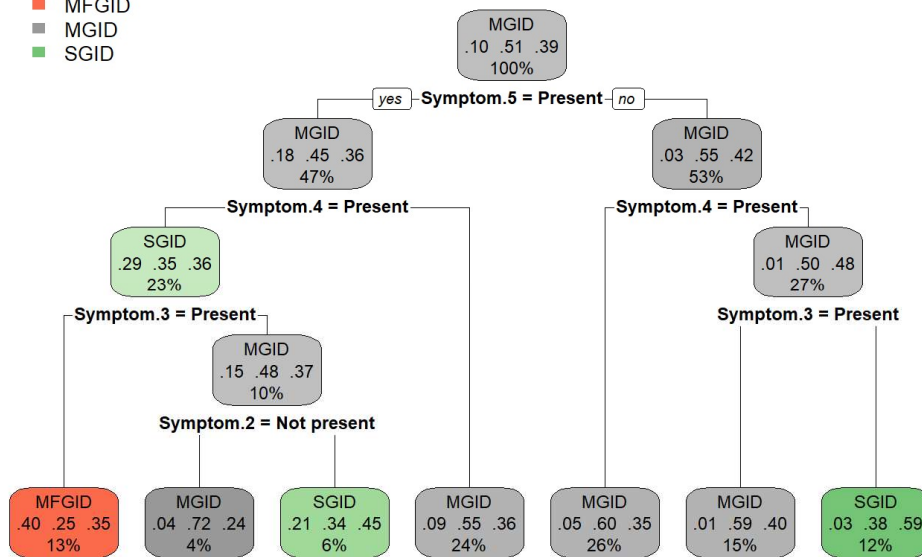
Detection Prevalence: # of (+) predictions by model/ total # of instances

Balanced Accuracy: average of sensitivity and specificity

Decision Tree of Test Dataset

* Optimized decision tree

■ MFGID
■ MGID
■ SGID



- Accuracy can be improved by adding more highly correlated variables to create the decision tree.

1

- Accuracy can be improved by transforming existing features to increase the performance of the model

2

- Accuracy can be improved by handling imbalanced classes.

3

Conclusion

Citations

- Kumar, A. (2021). Predict the genetic disorders dataset-of genomes. [www.kaggle.com.
https://www.kaggle.com/datasets/aibuzz/predict-the-genetic-disorders-datasetof-genomes?rvi=1](https://www.kaggle.com/datasets/aibuzz/predict-the-genetic-disorders-datasetof-genomes?rvi=1)
- [Decision Tree Models in Python — Build, Visualize, Evaluate | by Mustafa Adel Amer | Towards Data Science](#)