Project report

Executive Summary

The Genetic Disorder Prediction project has reached its conclusion, presenting significant findings and results. Leveraging machine learning models and a carefully curated dataset, the project aimed to predict the likelihood of individuals having a genetic disorder. This report summarizes key findings, model performances, and insights gained throughout the project.

Key Findings

1. Data Exploration and Preprocessing

Data Quality:

The genomes-and-genetics-hackerearth-ml dataset showed varying degrees of data quality.

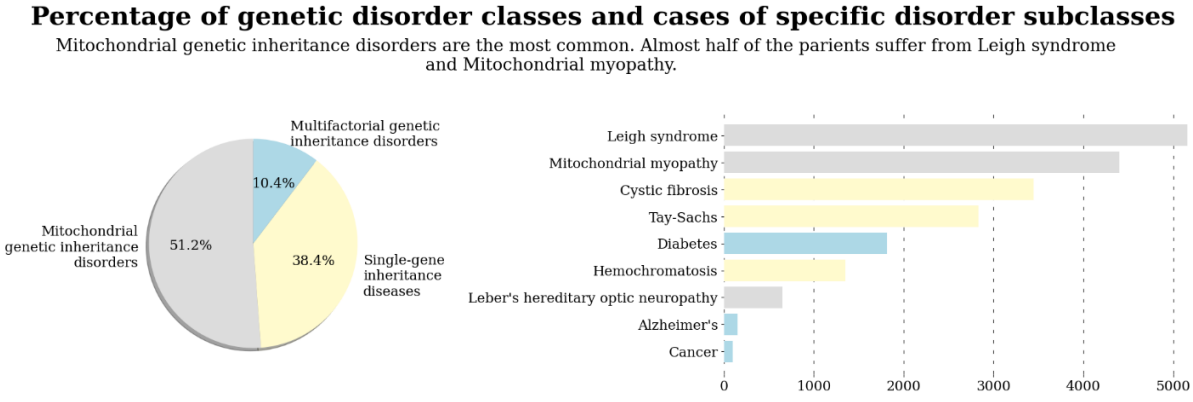
Preprocessing efforts focused on handling missing values, encoding categorical variables, and scaling numerical values

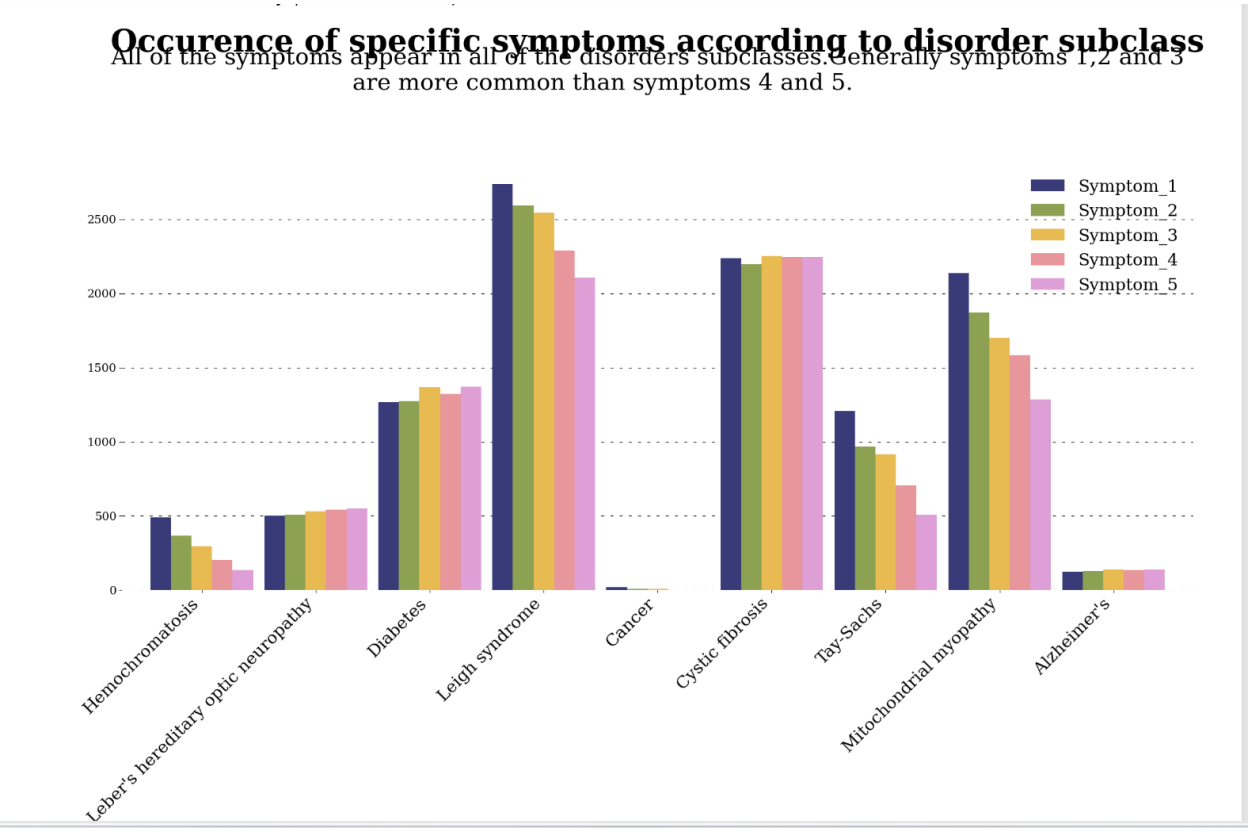
2. Exploratory Data Analysis (EDA)

Data Distribution:

EDA revealed diverse distributions within patient demographics, genetic information, and medical data.

Identified potential correlations and patterns that informed subsequent modeling decisions.





3. Model Development

Model Selection:

Logistic regression, decision trees, random forests, gaussian naïve bayes and xgboost models were chosen for their suitability to the problem.

Each model underwent training on a subset of the dataset.

- **Logistic Regression:** Serves as a foundational multi-class classification model, establishing an initial performance baseline.
- **Decision Tree:** Captures intricate decision boundaries and provides insights into feature importance for multi-class scenarios.
- **Random Forest:** Ensemble model designed to enhance robustness and predictive accuracy in a multi-class context.
- **Gaussian Naive Bayes:** Assumes feature independence and efficiently handles high-dimensional data, making it suitable for multi-class classification.
- **XGBoost:** Widely recognized for high performance and scalability, specifically tailored for multi-class classification challenges.

4. Model Evaluation

Performance Metrics:

Model evaluation utilized metrics such as accuracy, precision, recall, and F1 score.

Cross-validation and validation datasets were crucial in assessing model robustness.

```
Performance Metrics:
```

	Model	Mean Score	Std Score
0	Logistic Regression	0.518536	0.003803
1	Decision Tree	0.471512	0.010593
2	Random Forest	0.539305	0.010550
3	Gaussian Naive Bayes	0.519085	0.003184
4	XGBoost	0.522925	0.013151

5. Hyperparameter Tuning

Optimization Strategies:

Fine-tuned model hyperparameters to enhance predictive performance.

Balanced the models to avoid overfitting and underfitting.

Model Performances

Logistic Regression

- Accuracy: 52%

Random Forest

- Accuracy: improved from 55% to 57%.

XGBoost

- Accuracy: improved from 54% to 57%

Discussion of strengths and weaknesses.

- **Logistic Regression:**
    - *Strengths:* Simple, interpretable.
    - *Weaknesses:* Limited complexity may underperform intricate relationships.
- **Decision Tree:**
    - *Strengths:* Captures non-linear patterns.
    - *Weaknesses:* Prone to overfitting, sensitive to small changes in data.
- **Random Forest:**
    - *Strengths:* Robust, high accuracy.
    - *Weaknesses:* Less interpretable due to ensemble nature.
- **Gaussian Naive Bayes:**
    - *Strengths:* Efficient with assumptions met.
    - *Weaknesses:* Relies on independent assumptions.
- **XGBoost:**
    - *Strengths:* High accuracy, resistant to overfitting.
    - *Weaknesses:* Complexity may affect interpretability.

5. Challenges Faced

Feature Engineering:

Initial features lacked predictive power; required careful selection and transformation.

- **How Challenges Were Overcome:**

Feature Engineering:

1. Conducted in-depth exploratory data analysis (EDA) to identify relevant features.

Employed domain knowledge to create new features, enhancing model performance.

6. Feature Importance Analysis:

```
Feature Importance:
```

	Feature	Logistic Regression	Decision Tree	Random Forest	XGBoost
0	Patient_age	-0.004569	0.084051	0.083908	0.031886
1	Genes_mother_side	0.016161	0.131834	0.116086	0.033299
2	Inherited_father	-0.008715	0.049090	0.055427	0.030930
3	Maternal_gene	-0.007977	0.132041	0.107730	0.034186
4	Paternal_gene	0.080980	0.034273	0.028209	0.047533
5	Blood_cell_count	0.098078	0.027966	0.027695	0.048638
6	Status	0.047473	0.033008	0.028484	0.041198
7	Respiratory_rate	-0.039249	0.028736	0.033058	0.036442
8	Heart_rate	0.016755	0.030504	0.032456	0.033461
9	Follow_up	-0.009937	0.026665	0.032166	0.035319
10	Gender	0.004135	0.026697	0.032077	0.032692
11	Folic_acid	-0.013139	0.029858	0.032193	0.034488
12	Assisted_conception	-0.034200	0.048809	0.048335	0.032022
13	History_previous_pregnancies	0.013141	0.028584	0.032303	0.032154
14	Previous_abortions	-0.002013	0.029379	0.031855	0.031085
15	Birth_defects	-0.053161	0.027749	0.032292	0.032749
16	White_blood_cell_count	-0.026633	0.033787	0.032857	0.030975
17	Blood_test	-0.015844	0.057496	0.057279	0.034367
18	Symptom_1	0.087994	0.035012	0.028042	0.062284
19	Symptom_2	0.128644	0.021007	0.027309	0.073596
20	Symptom_3	0.055133	0.034274	0.034925	0.070396
21	Symptom_4	0.152798	0.017467	0.031289	0.089065
22	Symptom_5	-0.159141	0.031713	0.034027	0.071236

Conclusion

The Genetic Disorder Prediction project has successfully provided a foundation for predicting genetic disorders using machine learning. While facing challenges, the project navigated data complexities and implemented models that show promise in their predictive capabilities. The insights gained and recommendations presented pave the way for future enhancements and research in genetic disorder prediction. The documentation stands as a valuable resource for understanding the project's journey and outcomes.