# Capstone Presentation

Presented by Serena Hy

# Introduction

This presentation outlines the capstone project focused on analyzing genetic disorders using advanced data science techniques.

**Objectives**: The goal is to identify patterns and correlations between genetic markers and health outcomes, aiding in the understanding and prediction of genetic disorders.

**Approach**: The project utilizes a comprehensive dataset and applies various data science methodologies, from data wrangling to machine learning modeling, to derive actionable insights.

# Project vision and mission

This project explores the critical role of genetics in health and disease, highlighting its significance in the burgeoning field of personalized medicine. Personalized medicine, tailoring healthcare strategies to individual genetic profiles, is becoming increasingly vital in providing effective treatments.

01. **Research Focus**: The project investigates the correlation between genetic markers and various health outcomes to understand the implications of genetic variations.

02. S**ignificance**: Understanding these correlations is crucial for early diagnosis and targeted treatment strategies for genetic disorders.

03. **Challenge**: Identifying reliable patterns and predictions in a complex dataset requires sophisticated data analysis techniques.

# Project Process

## 01

**Data Wrangling**: Prepare the "Of Genomes and Genetics" dataset for analysis, which includes loading the data, handling missing values, and understanding its structure.

## 02

**Exploratory Data Analysis (EDA):** Analyze the dataset to identify patterns, anomalies, or relationships between features, focusing on genetic variations and potential health outcomes.

## 03

**Pre-processing & Training Data Development:** Process the data into a format suitable for modeling, which includes feature selection, normalization or standardization, and splitting the dataset into training and testing sets.

## 04

**Modeling**: Apply statistical and machine learning models to identify and validate the correlation between genetic markers and health outcomes.

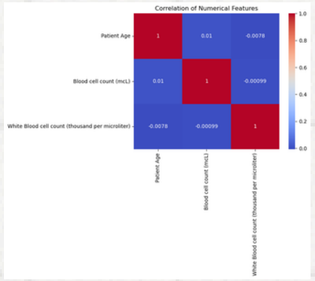
# Data Overview

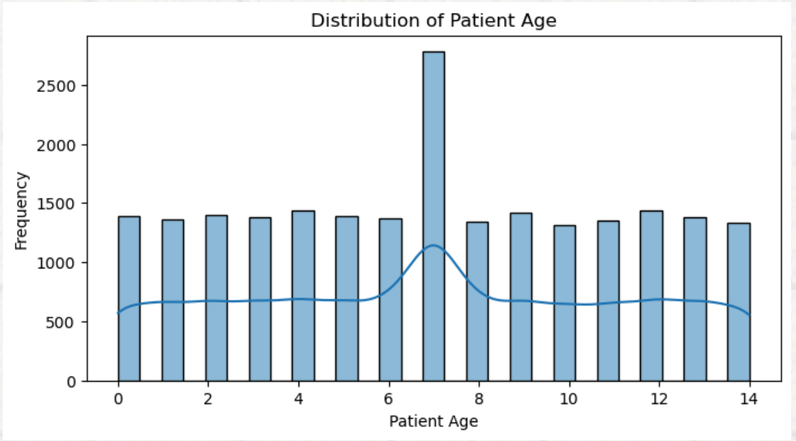
**Data Sources**: The dataset originates from extensive genetic studies and includes both clinical and genetic information of patients.

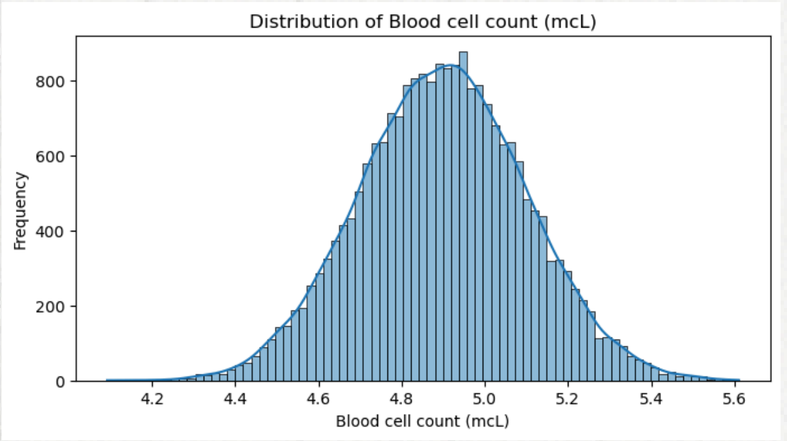
**Key Features**: Includes patient age, genetic markers, health outcomes, and other relevant medical data.

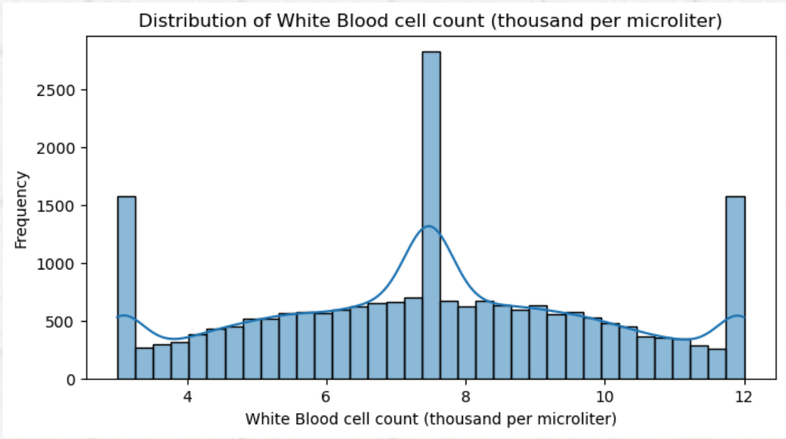
**Data Volume**: The dataset encompasses over 20,000 patient records, making it comprehensive for robust analysis.

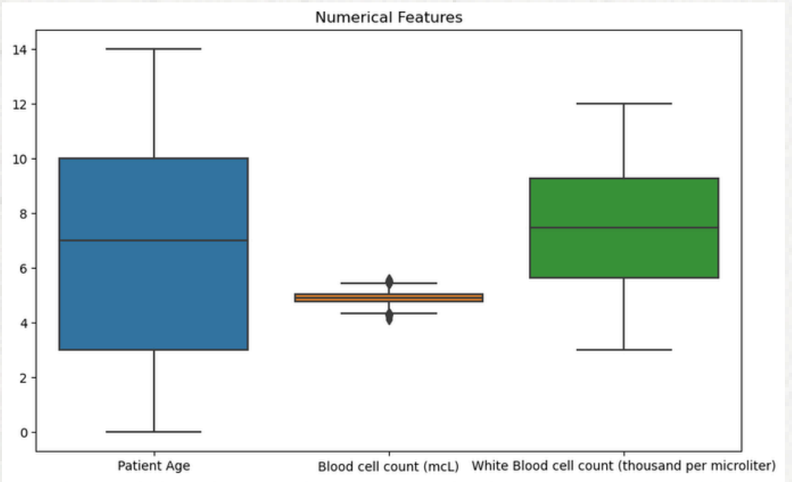
# Exploratory Data



# Mind Map

Exploring Data

## Data Wrangling

**Data Cleaning**: Handling missing values, removing duplicates, and correcting data errors to ensure data quality.
**Data Transformation**: Standardizing formats, encoding categorical data, and normalizing numerical values to enhance model performance.
**Data Splitting**: Segmenting the data into training and testing sets to evaluate model effectiveness accurately.

## Exploratory Data Analysis

**Statistical Summaries**: Review of statistical measures to understand distributions, variability, and central tendencies of the data.
**Visual Explorations**: Use of plots and charts to visualize relationships between features and identify potential correlations or outliers.
I**nitial Findings**: Identification of key features likely influencing genetic disorders, setting the stage for deeper analysis in modeling.

## Pre-processing & Training

**Feature Engineering**: Creation of new variables that enhance the model's ability to predict genetic disorders, such as ratios or aggregated features.
**Data Normalization**: Standardization and normalization of data to ensure that model inputs are on a comparable scale, improving learning efficiency.
T**raining and Test Split:** Division of the data into training and validation sets to ensure robust model evaluation and prevent overfitting.

## Model Development

**Model Selection**: Utilization of RandomForest and XGBoost classifiers to model genetic disorders based on the identified features.
**Parameter Tuning**: Application of grid search techniques to optimize model parameters for best performance.
**Model Training**: Training models on preprocessed data, evaluating using cross-validation to ensure generalizability.

## Model Evaluation

**Evaluation Metrics**: Use of accuracy, precision, recall, and F1-score to evaluate the performance of the models.
**Comparison of Models**: Comparative analysis of RandomForest and XGBoost models based on performance metrics and cross-validation scores.
**Model Diagnostics**: Analysis of confusion matrices and error rates to identify areas for model improvement.

## Best Model Selection

**Selection Criteria**: Selection based on highest accuracy and F1-score, stability across different datasets, and computational efficiency.
**Final Model:** XGBoost was selected due to its superior performance on cross-validation scores and its ability to handle diverse data features effectively.
**Model Justification**: XGBoost provides a good balance of speed and predictive power, making it suitable for large-scale genetic data analysis.

## Models Tested

Various models were evaluated including RandomForest, XGBoost, and logistic regression. Each model was tested for its ability to accurately predict genetic disorders based on the dataset.

## Selection Criteria

Selection based on highest accuracy and F1-score, stability across different datasets, and computational efficiency.

## Final Model

XGBoost was selected due to its superior performance on cross-validation scores and its ability to handle diverse data features effectively.

## Model Justification

XGBoost provides a good balance of speed and predictive power, making it suitable for large-scale genetic data analysis.

# 55.4%

## accuracy

with the XGBoost model showing the best overall performance in terms of handling the imbalanced dataset. The model metrics comparison higlighted the need for further tuning and possibly gathering more data to improve the model's predictve capabilities.

# Final reflections and future steps

**Project Overview**: This project applied advanced data science techniques to analyze genetic markers and predict health outcomes.

**Key Findings**: The analysis revealed significant correlations between genetic markers and specific genetic disorders, providing insights for potential clinical applications.

**Future Work**: Further research could explore more complex models and larger datasets to refine predictions and extend the applicability to other genetic conditions.

# Thank you very much!