# Model Exploration

Disleve Kanku

# Model Overview

- **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.

# Logistic Regression

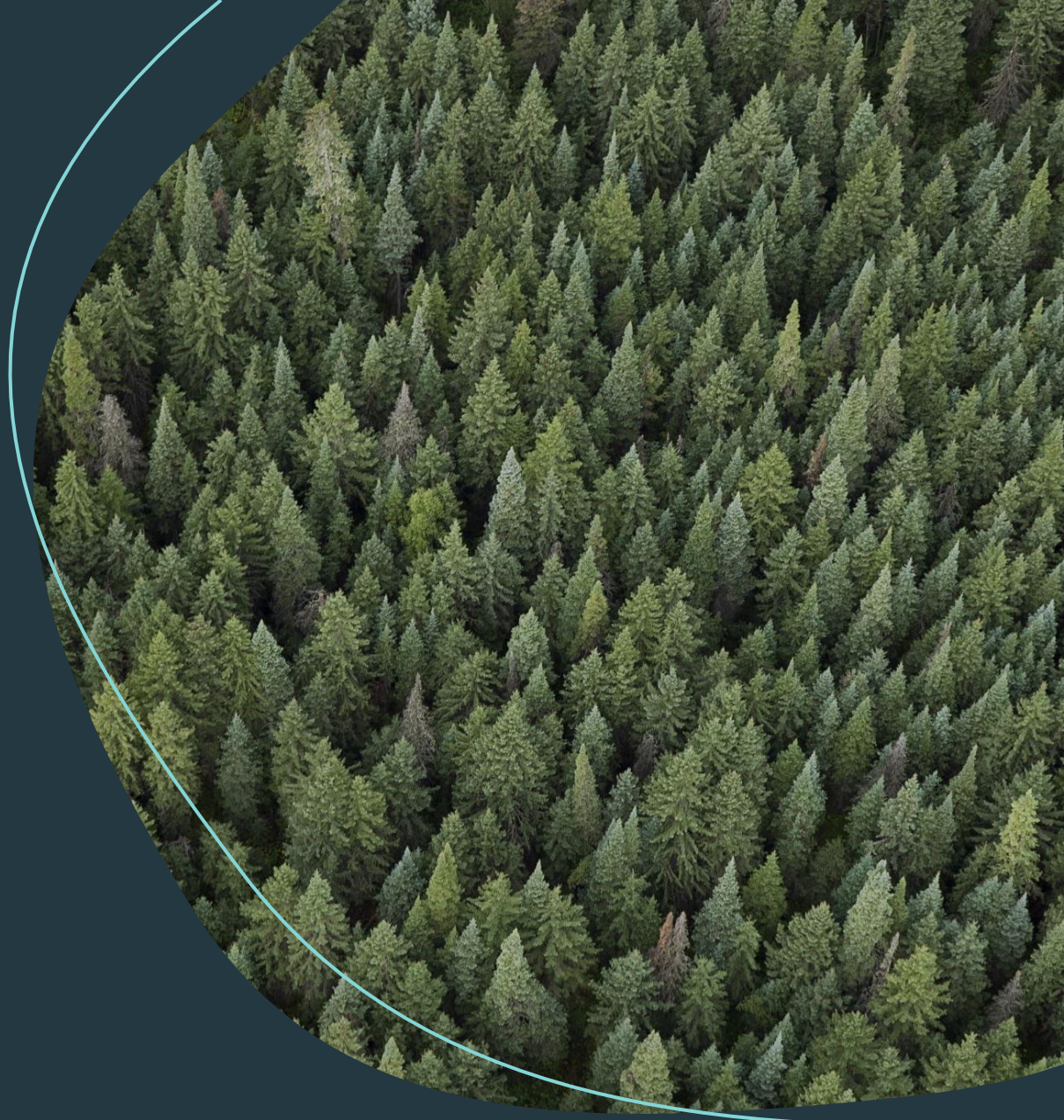
- **Overview of Logistic Regression:**
  - Logistic Regression is a linear model adapted for multi-class classification.
  - It models the probability of each class using a logistic function, making it suitable for estimating class probabilities.
- **Application in My Analysis:**
  - Logistic Regression serves as the baseline model, providing insights into the fundamental relationships between features and classes.
  - Applied to predict the likelihood of each class, aiding in the understanding of initial classification patterns.

# Decision Tree

- **Overview of Decision Tree:**
  - Decision Trees are non-linear models suitable for multi-class classification.
  - They recursively split data based on feature thresholds, creating a tree-like structure of decisions.
- **Application in My Analysis:**
  - Decision Trees capture complex decision boundaries, providing a deeper understanding of the data.
  - Insights gained from the tree structure contribute to feature importance analysis.

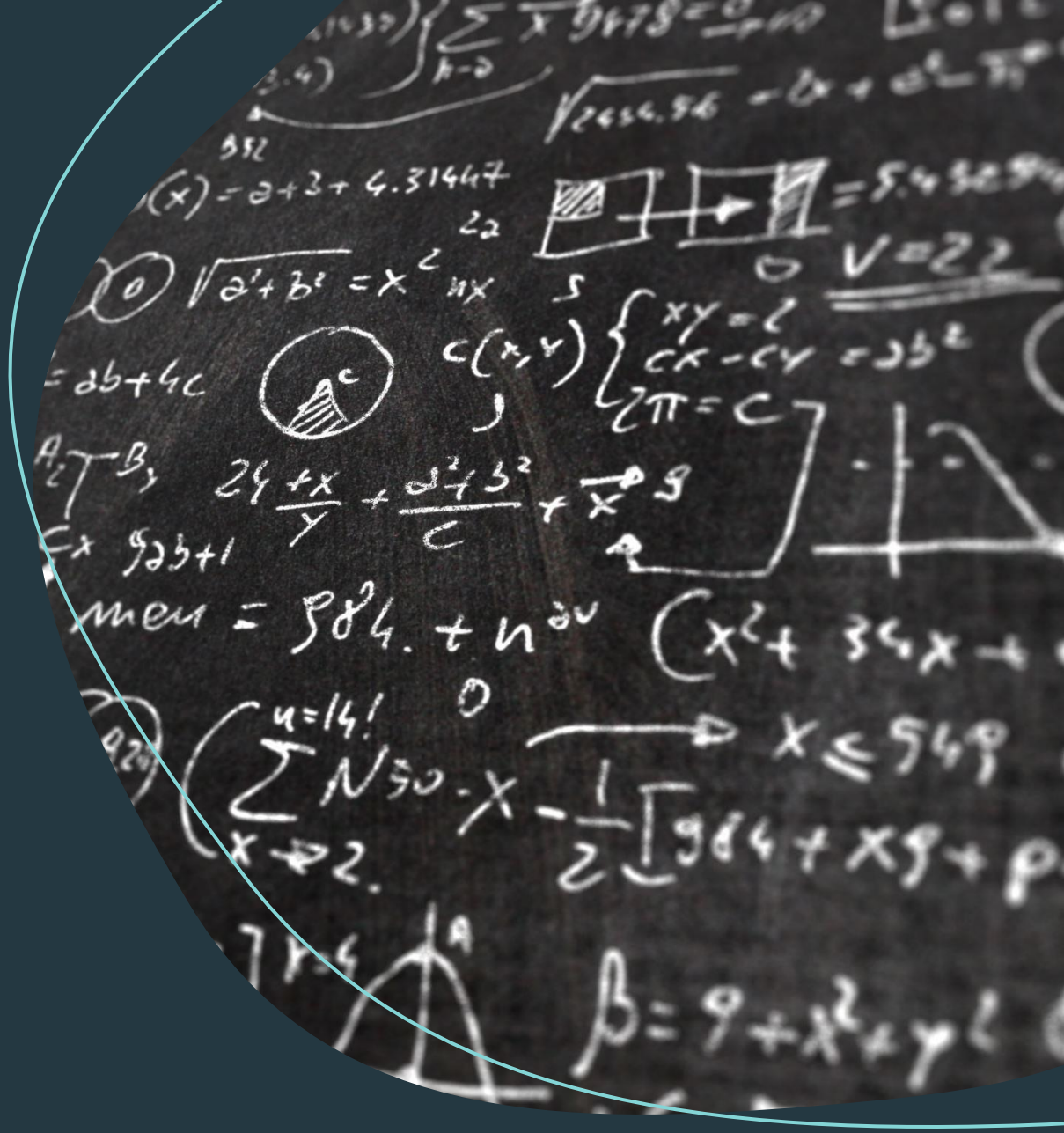
# Random Forest

- **Overview of Random Forest:**
  - Random Forest is an ensemble model built on multiple Decision Trees.
  - It aggregates predictions from individual trees to improve overall predictive accuracy and generalization.
- **Application in My Analysis:**
  - Random Forest enhances robustness by mitigating overfitting and capturing diverse patterns in the data.
  - As an ensemble, it offers improved performance compared to individual Decision Trees.



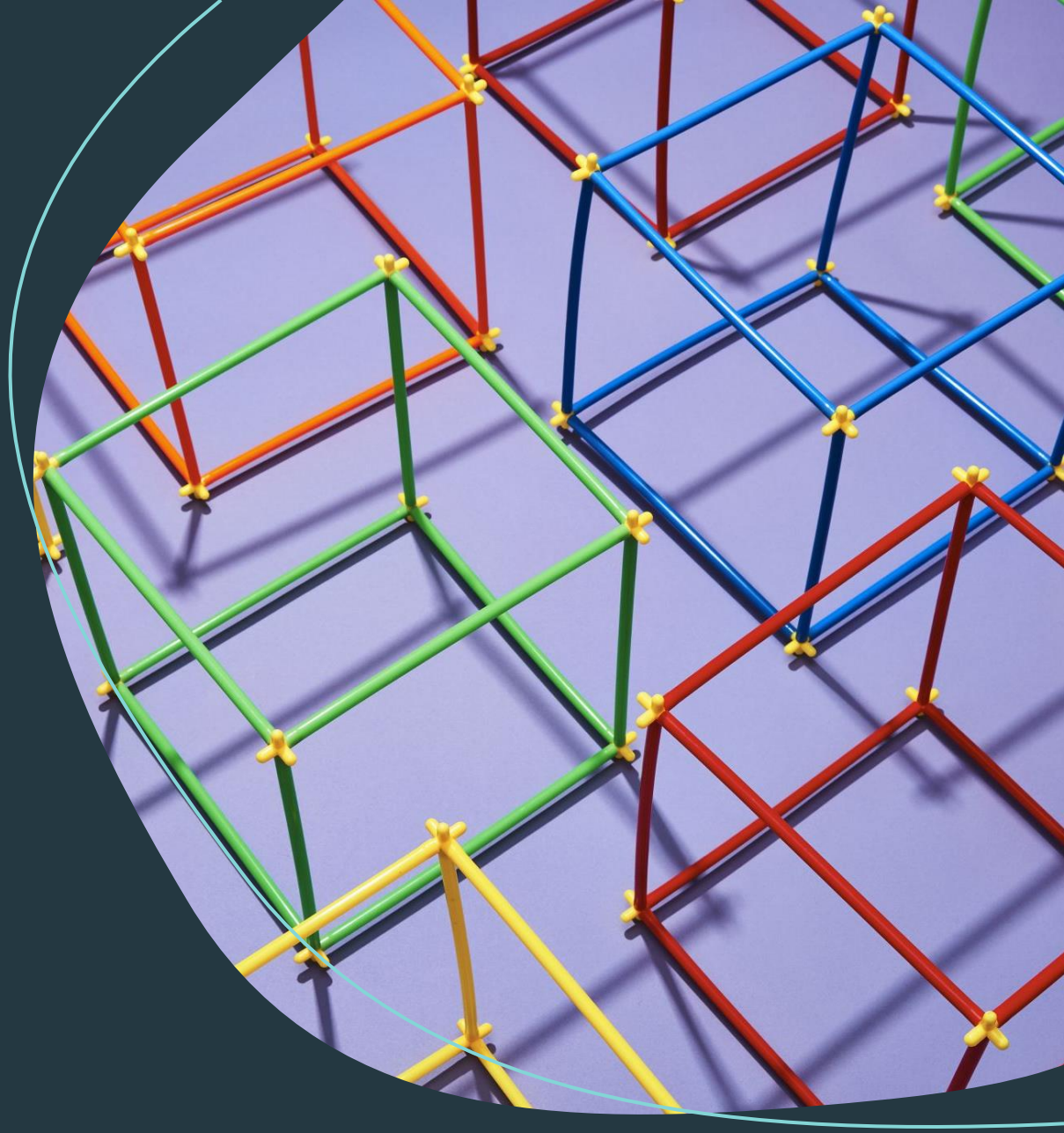
# Gaussian Naïve Bayes

- **Overview of Gaussian Naive Bayes:**
  - Gaussian Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem.
  - It assumes that features are conditionally independent, given the class.
- **Application in My Analysis:**
  - Well-suited for datasets with continuous features and a Gaussian distribution assumption.
  - Efficient and effective for multi-class classification tasks.



# XGBoost

- **Overview of XGBoost:**
  - XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm that uses decision trees as base models.
  - It employs a gradient boosting framework, sequentially adding trees to correct errors made by previous ones.
- **Application in My Analysis:**
  - XGBoost excels in capturing complex relationships and interactions within the data.
  - Provides high predictive accuracy and is resistant to overfitting.



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 |

# Model Performance

- The table shows model performance for all our models
- Random Forest is the best performing model

# Feature Importance

- Table showing feature importance for our used models

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 |

# Model Comparison

- **Logistic Regression:**
  - Achieved an accuracy of 52%.
  - Well-suited for linear relationships but may not capture complex patterns.
- **Decision Tree:**
  - Accuracy reached 48%.
  - Good at capturing non-linear relationships; prone to overfitting.
- **Random Forest:**
  - Highest accuracy at 55%.
  - Robust against overfitting, thanks to ensemble learning.
- **Gaussian Naive Bayes:**
  - Achieved 52% accuracy.
  - Performs well with assumptions of independent features.
- **XGBoost:**
  - Outperformed individual decision trees with 54% accuracy.
  - Effective in capturing complex relationships; resistant to overfitting.

# Strengths and weaknesses

- **Logistic Regression:**
  - *Strengths:* Simple, interpretable.
  - *Weaknesses:* Limited complexity, may underperform for 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 independence assumptions.
- **XGBoost:**
  - *Strengths:* High accuracy, resistant to overfitting.
  - *Weaknesses:* Complexity may affect interpretability.

# Challenges

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

# Conclusion

1. **Model Overview:**
    1. Introduced and implemented Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and XGBoost.
2. **Model Details:**
    1. Explored each model's overview, application in analysis, and highlighted feature importance.
3. **Model Comparison:**
    1. Compared performance metrics, strengths, and weaknesses of each model.
4. **Challenges:**
    1. Discussed challenges faced during implementation and strategies employed to overcome them.