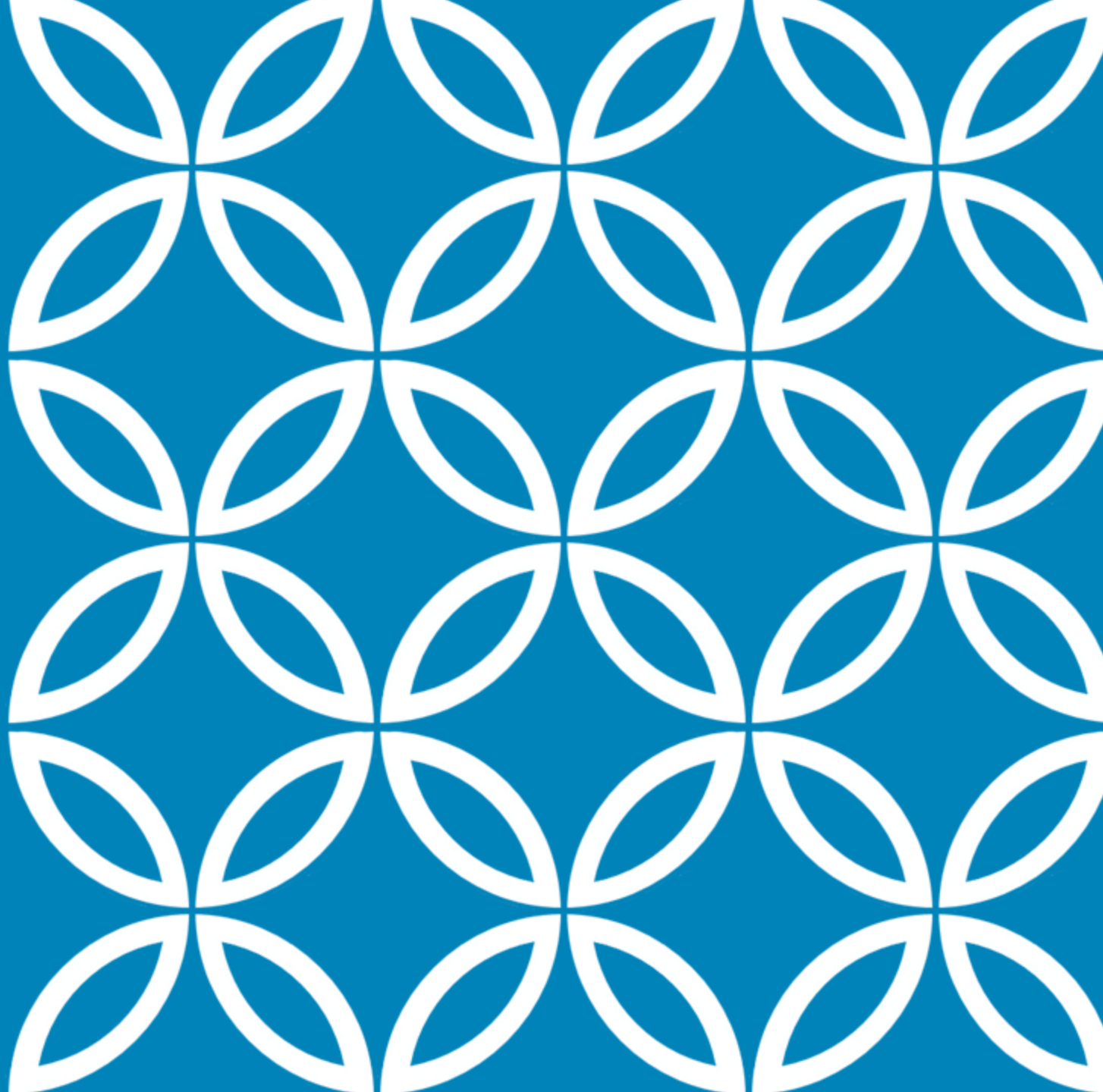
DATA MINING FINAL PROJECT: TEAM 2

Rupali Kakadia, Krupasankari Ragunathan,
Sanjith Krishna Venkatesh Kumar



OUR INITIAL APPROACH

1. *Logistic Regression*:
   - Easy to interpret and a strong baseline for classification problems.
   - Performs well when relationships are linear and data is well-preprocessed.
   - ROC against Validation set: 0.80
2. *Decision Trees*:
   - Simple, interpretable, and effective for capturing hierarchical relationships
   - Often used as a foundation for methods like Gradient Boosted Trees.
   - ROC against Validation set: 0.87

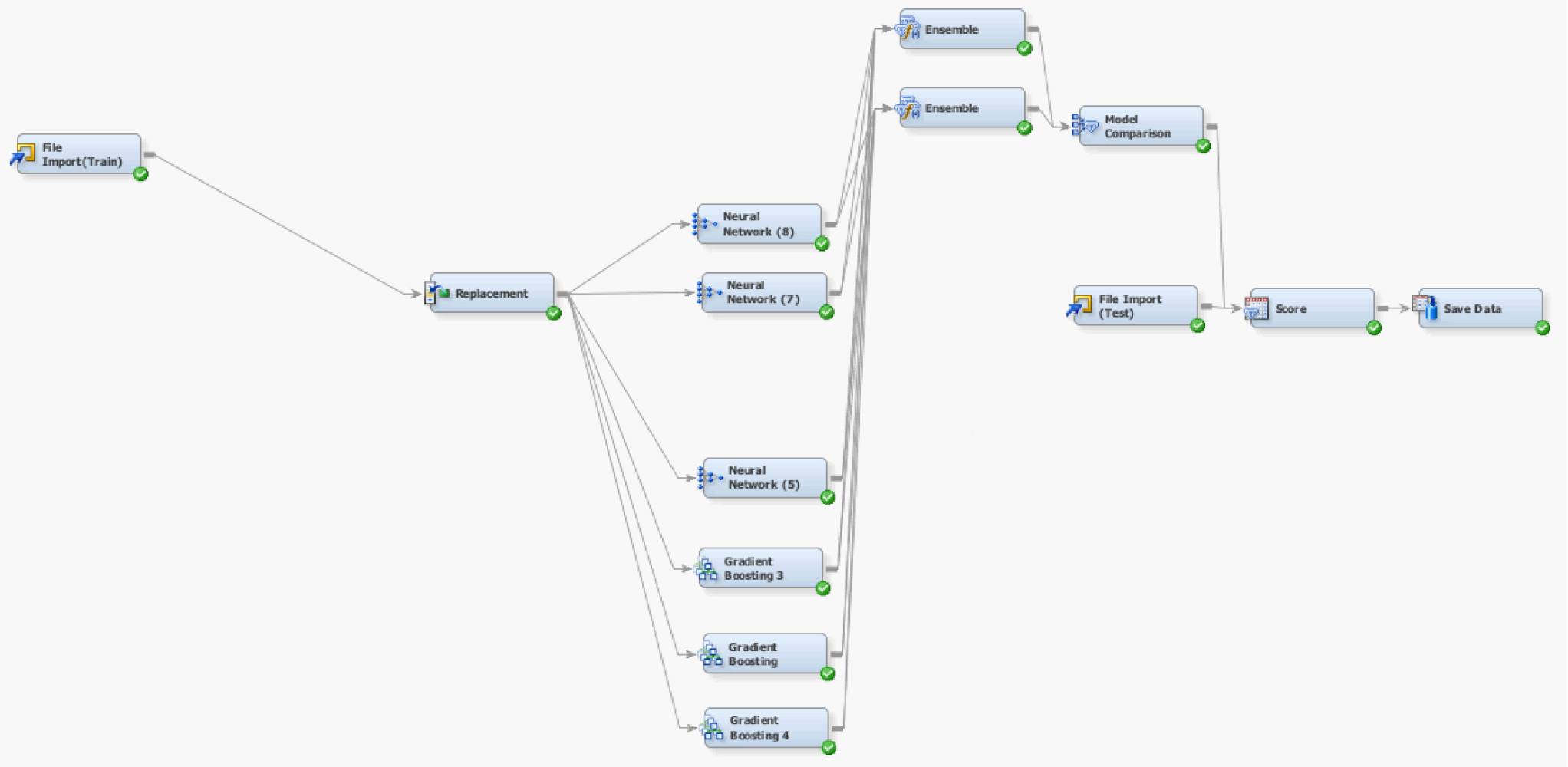
FURTHER INTO THE PROJECT…

3. *Gradient Boosted Trees*:
   - High performance for tabular data.
   - Incrementally corrects errors, making it robust to overfitting with proper hyperparameter tuning.
   - ROC against Validation set: 0.92
4. *Neural Networks*:
   - Effective for modeling complex feature interactions.
   - Performs well when large datasets allow for learning intricate patterns.
   - ROC against Validation set: 0.93

COMBINING GBT AND NEURAL NETWORKS: THE ENSEMBLE METHOD

- GBT captures non-linear relationships and performs well with tabular data.
- NN excels in detecting complex patterns and interactions.
- Together, they create a robust model that balances bias and variance.

BEST ALGORITHM: ENSEMBLE
(PRIVATE LEADERBOARD: 95.806, PUBLIC LEADERBOARD: 95.162)



WHY ENSEMBLE MODELS?

1. *Why Ensemble Models Work*:
   - Combines strengths of diverse algorithms.
   - Reduces the weaknesses of individual models (e.g., GBT is prone to overfitting, NN needs large data).
   - Increases generalizability and reduces variance.

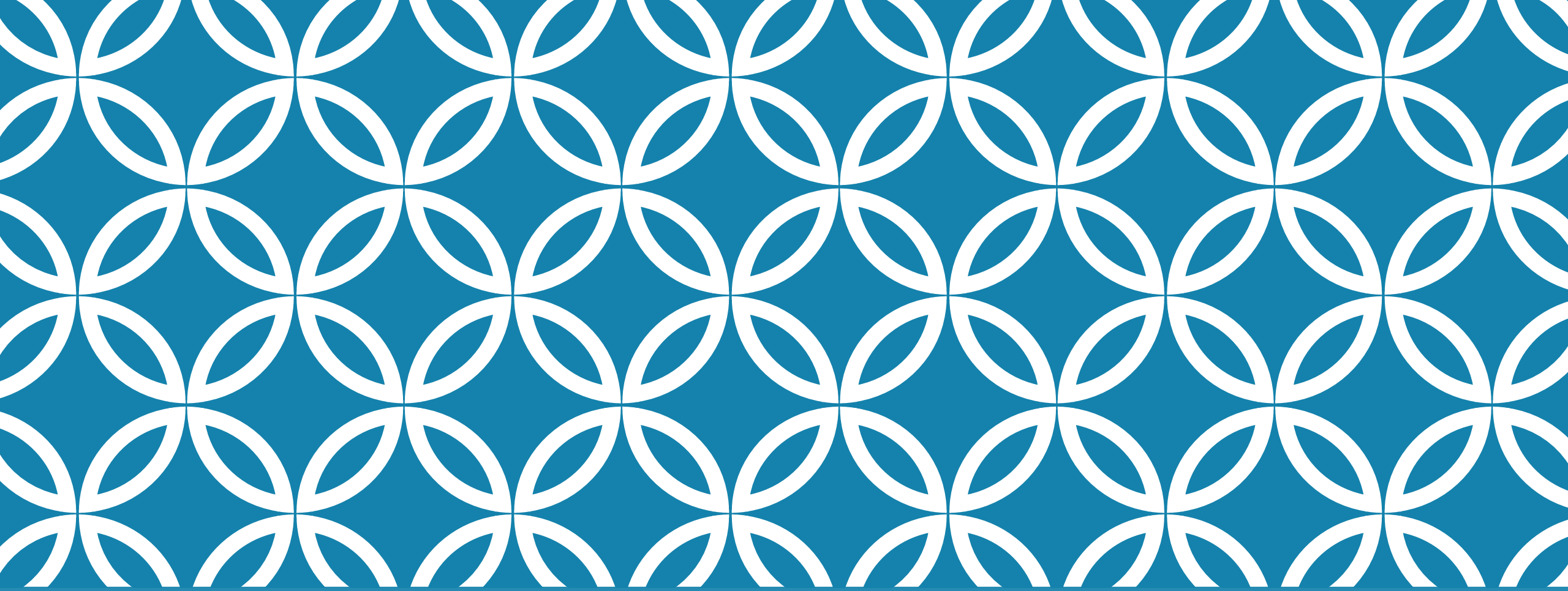
LEARNINGS: PREPROCESSING AND ITS IMPACT

1. *Preprocessing Learnings*:
   - **Transformation**: Useful for Logistic Regression and Neural Networks to normalize data for stability.
   - **Outlier Removal**: Effective for Gradient Boosted Trees but risky for Neural Networks as it may remove critical patterns.

Example: Filtering outliers improved GBT performance but hurt NN as it reduced the data diversity.

LEARNINGS: OVERFITTING AND ENSEMBLE INSIGHTS

1. *Overfitting*:
   - Observed in Neural Networks when hyperparameters were too complex.
   - Mitigated using careful tuning.
2. *Common Learning from Ensembles*:
   - AUC scores improved consistently as ensembles reduced both model bias and variance.



THANK YOU