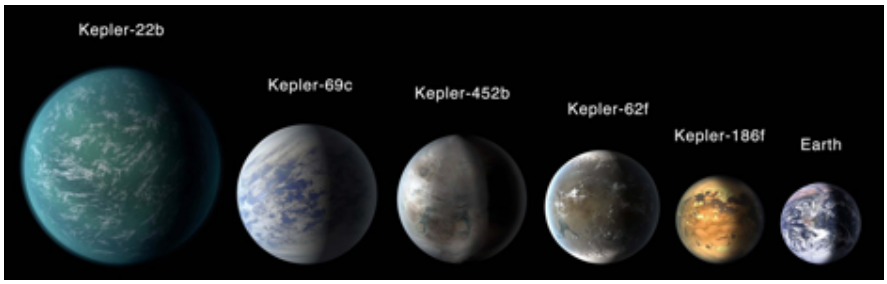
Project Report - Exoplanet Detection Using Ensemble Learning

Supervised by: Myriam TAMI
Authors: Thomas Georges & Manon Lagarde

1. Introduction :

An **exoplanet** is a planet outside our solar system, orbiting a star other than the Sun. Detecting these planets helps scientists explore potential habitable worlds and understand planetary formation.



The goal of this study is to develop an accurate classification model using **ensemble learning techniques** to distinguish between exoplanet candidates and false positives. By leveraging machine learning models such as **Random Forest, Extra Trees, XGBoost, LightGBM, and AdaBoost**, we aim to improve classification performance and determine the most effective model for exoplanet detection.

The dataset used for this analysis comes from **NASA's Kepler mission** and includes observational data on detected planetary signals. The primary objective is to preprocess the data, build predictive models, optimize their performance, and compare their effectiveness.

More Information about the dataset can be found here :
https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html

2. Methodology

2.1 Data Preprocessing

The dataset (`cumulative.csv`) contains features describing exoplanet candidates, including planetary properties, stellar characteristics, and transit signal attributes.

Several preprocessing steps were applied to enhance data quality:

- **Feature Selection:**
  - Non-informative and administrative features such as *kepid*, *kepoi_name*, and *kepler_name* were removed.

    - Features with potential label leakage (e.g., *koi_pdisposition*, *koi_score*) were identified and excluded to prevent unfair model advantages.

- **Handling Missing Values:**
    - Numerical missing values were imputed using the mean strategy.
    - Features with excessive missing values were removed.
- **Feature Engineering :**
    - The target variable *is_candidate* was created as a binary classification label (1: Candidate, 0: False Positive).
- **Correlation Analysis:**
  - Identified most important features for exoplanet detection.
  - Features with **low importance (<0.005)** were removed to simplify the model without affecting performance.
  - Identified redundant features with correlation above 0.75.

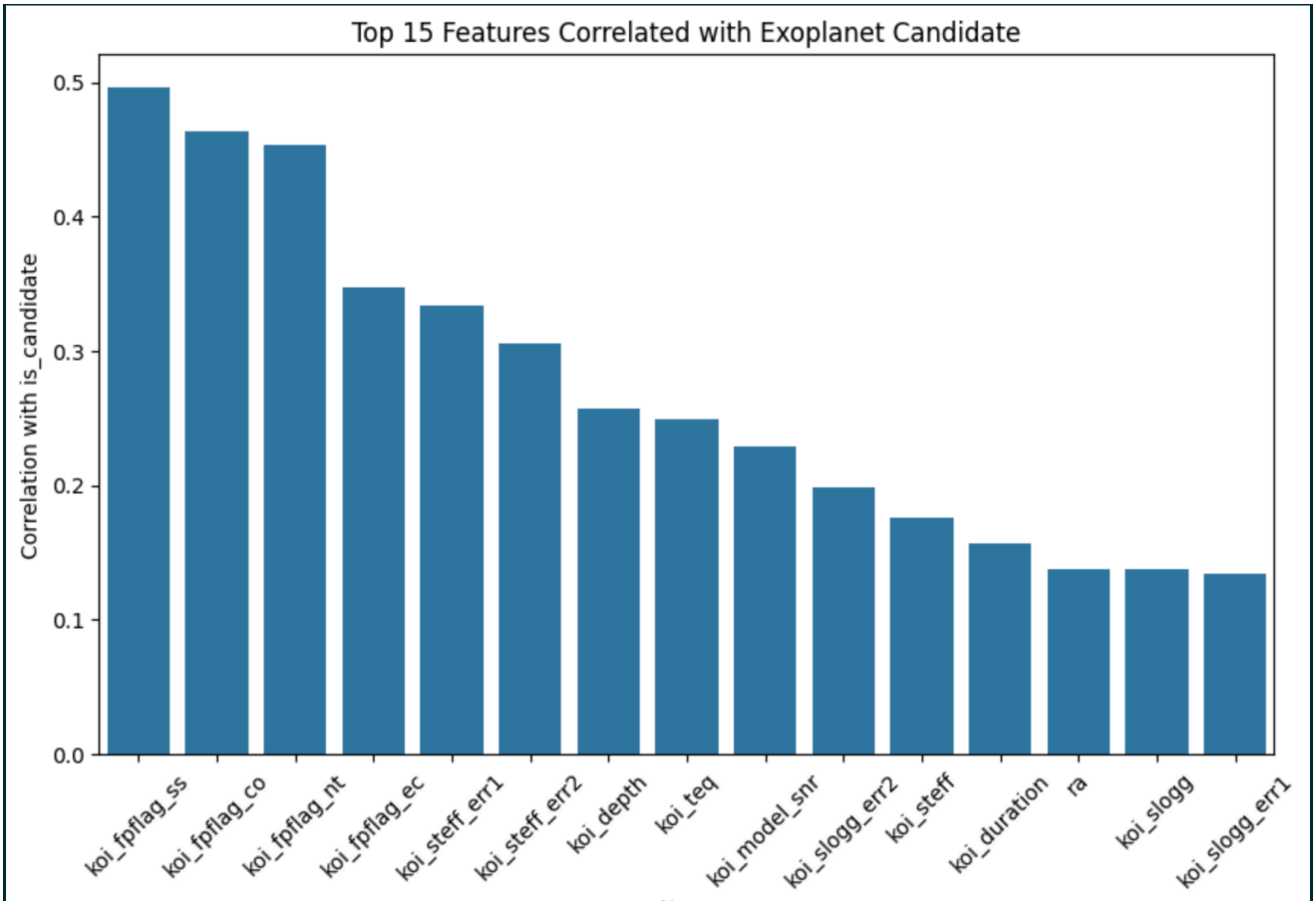


Figure : Top 15 features correlated with Exoplanet Candidacy (Target Variable)

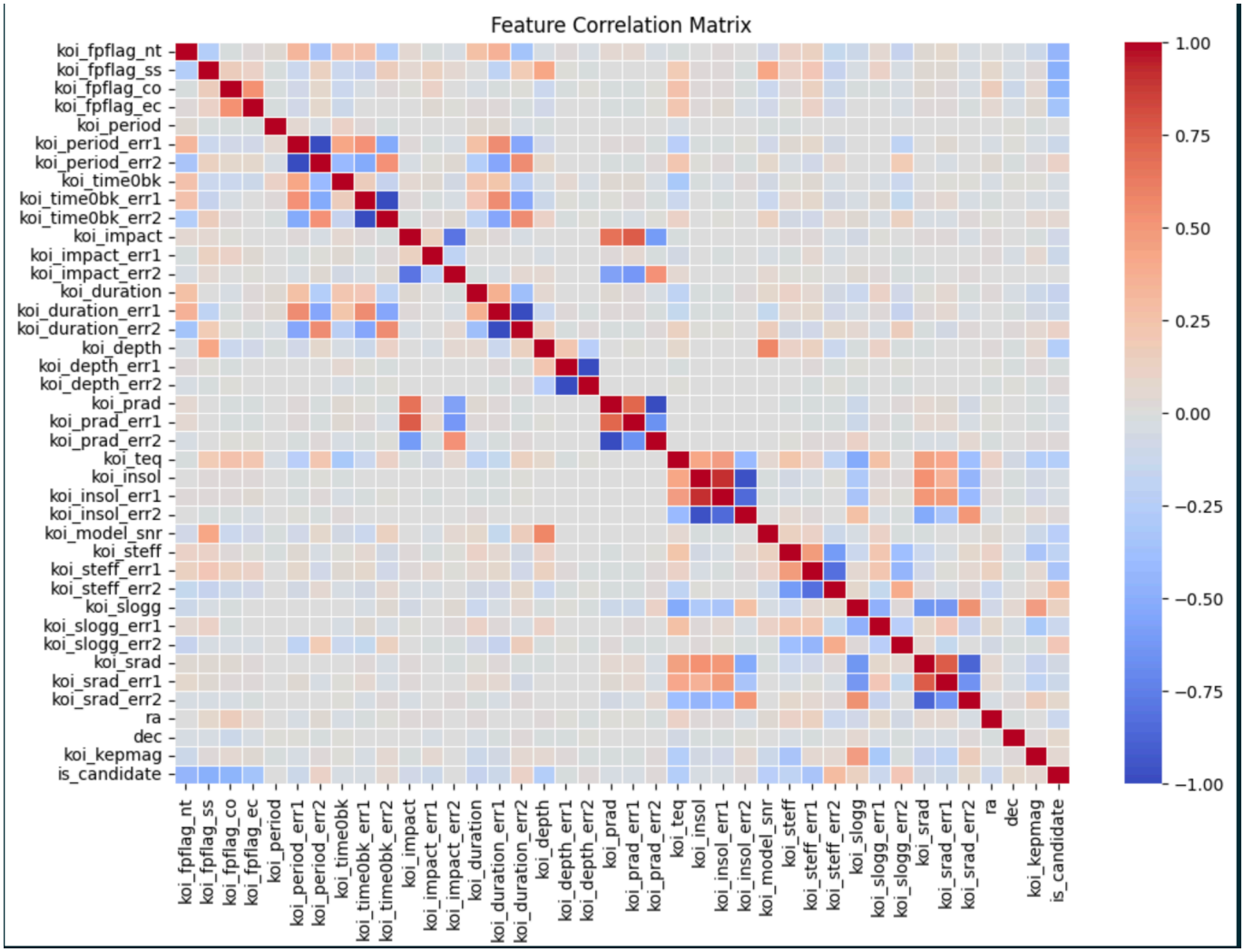


Figure : Correlation Matrix between the Features

2.2 Model Training & Optimization

We use the *koi_pdisposition* variable as the target label:

- `1` = Candidate exoplanet
- `0` = False Positive

We implemented five ensemble learning models to compare performance :

2. **Random Forest (RF)** - A bagging-based ensemble model using decision trees.
3. **Extra Trees (ET)** - Similar to Random Forest but introduces additional randomness.
4. **XGBoost (XGB)** - A gradient boosting algorithm known for efficiency and high performance.
5. **LightGBM (LGBM)** - A boosting algorithm optimized for speed and large datasets.
6. **AdaBoost (ADA)** - A boosting method that sequentially corrects errors from previous models.

2.2.1 Hyperparameter Tuning

To optimize model performance, we applied **hyperparameter tuning** using:

- **GridSearchCV** for smaller models (e.g., Random Forest), where exhaustive search is feasible.
- **RandomizedSearchCV** for XGBoost, LightGBM, and AdaBoost to efficiently explore large parameter spaces.

Key parameters optimized include:

- `n_estimators`, `max_depth`, `learning_rate`, `subsample`, `min_samples_split`.

2.2.2 Different Train-Test Splits Testing

Additionally, we tested different train-test splits (70% training, 30% testing) to assess whether model performance depends on the 80%-20% default split.

3. Results

3.1 Random Forest Classifier Analysis :

	Model	Training Accuracy	Testing Accuracy
0	Random Forest – Model#1 (RF1)	1.0	0.987454
1	Cross-Validation RF1	_	0.974171
2	Random Forest – Model#2 (RF2)	0.98693	0.986409
3	Cross-Validation RF2	_	0.974590
4	Random Forest – Optimized Model (OM)	1.0	0.988500
5	Cross-Validation OM	_	0.984185

The table compares different Random Forest models in terms of training accuracy, testing accuracy, and cross-validation (CV) performance:

RF1 (Baseline Model):

- Training accuracy is **1.0**, indicating potential overfitting.
- Testing accuracy (**0.9874**) shows good generalization but suggests overfitting.
- Cross-validation accuracy (**0.9741**) is slightly lower.

RF2 (Regularized Model):

- We used Regularization and tried reducing overfitting by tuning `max_depth` and increasing `min_samples_leaf`
- Training accuracy drops slightly to **0.9869**, with testing accuracy of **0.9864**.
- Cross-validation accuracy (**0.9745**) improves, indicating better generalization.

Optimized Random Forest (OM):

- This is our Optimized Random Forest Model, obtained by tuning our hyperparameters with Gridsearch.

- Achieves **100% training accuracy**, confirming model complexity.
- Testing accuracy improves to **0.9885**, making it the best-performing model.
- Cross-validation accuracy (**0.9841**) is also higher than previous models, showing improved generalization.

Conclusion : The optimized model (OM) performs the best overall, achieving the highest testing and cross-validation accuracy.

3.1.1. Feature Importance Analysis :

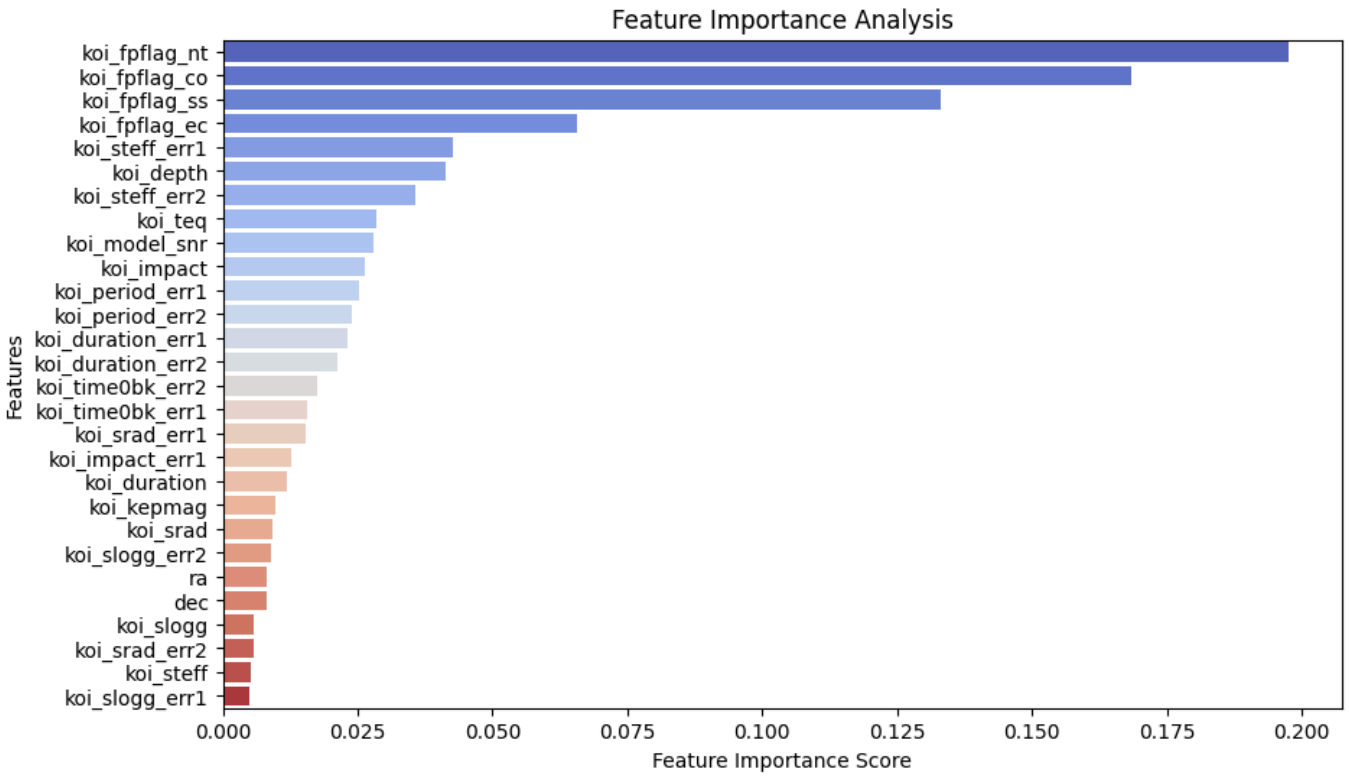


Figure : Feature Importance Analysis

Using the Random Forest model, feature importance was analyzed, and features with low importance (<0.005) were removed to simplify the model without significantly affecting performance.

3.2 Model Performance Comparison

3.2.1 Models Accuracy Comparison :

A comparison of training accuracy, testing accuracy, and cross-validation accuracy is summarized below:

Model	Training Accuracy	Testing Accuracy	Cross-Validation Accuracy
Random Forest	1.0	0.9874	0.9741
Extra Trees	1.0	0.9869	0.9745

Model	Training Accuracy	Testing Accuracy	Cross-Validation Accuracy
XGBoost	0.998	0.9879	0.9841
LightGBM	0.998	0.9879	0.9842
AdaBoost	0.991	0.9823	0.9735

Figure : Model's Accuracy Comparison

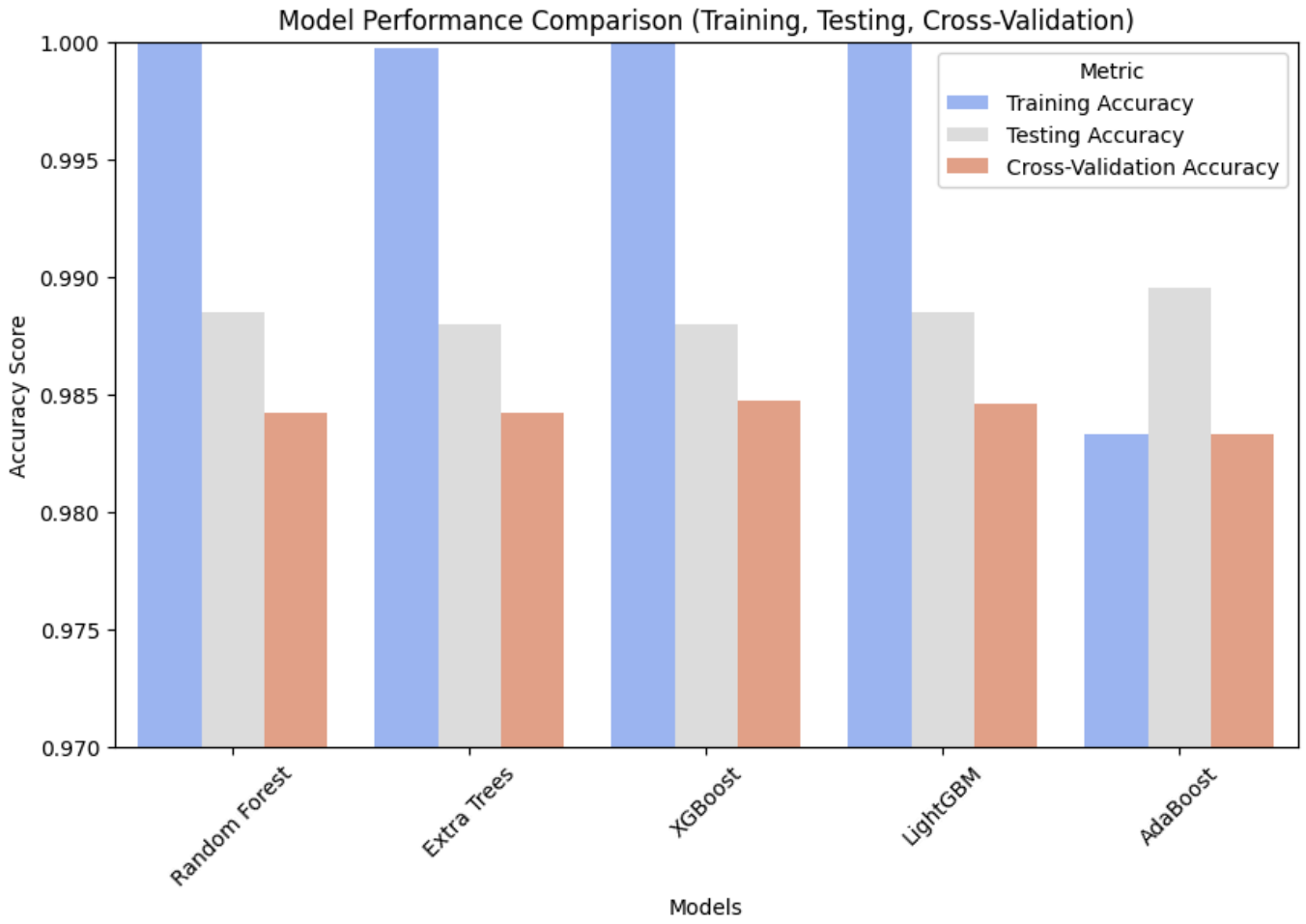


Figure : Visualization of Model's Accuracy Comparison

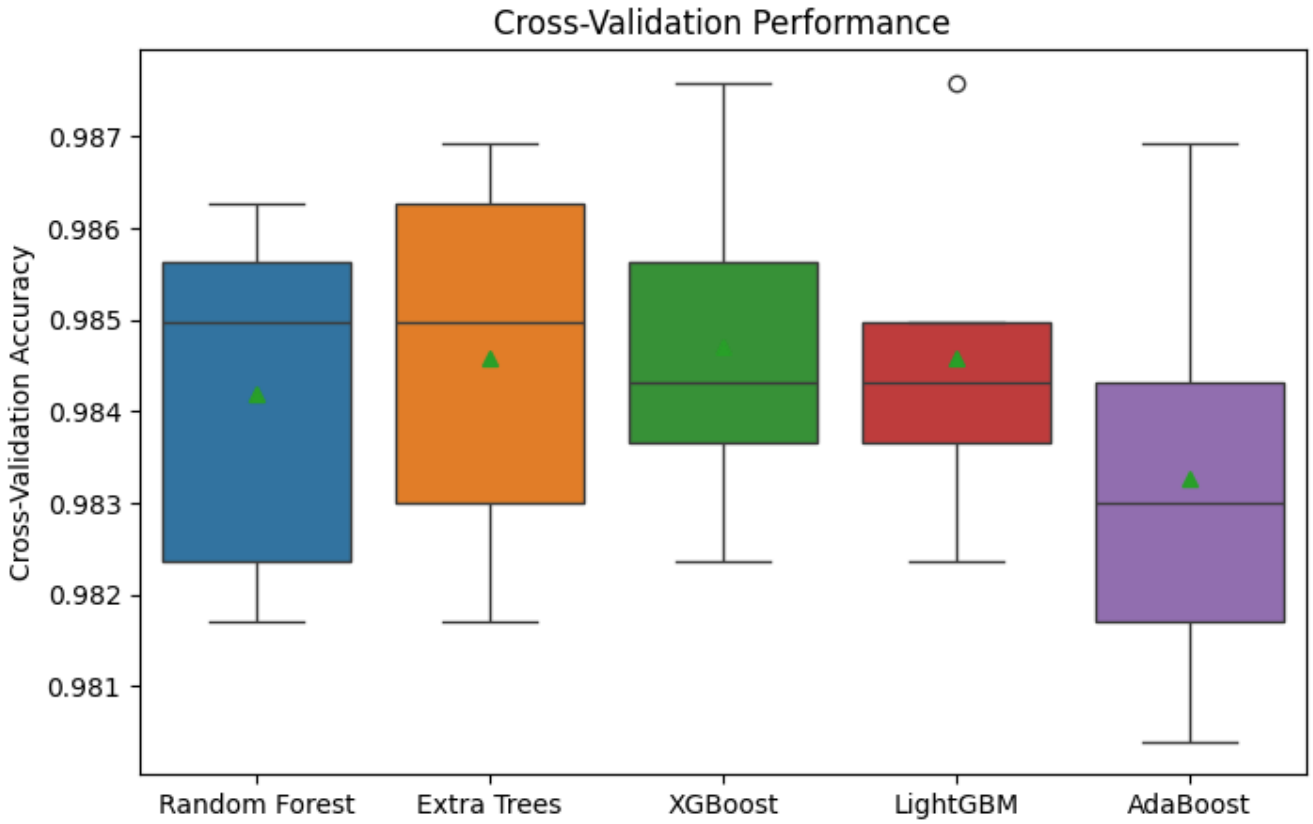


Figure : Model's Cross-Validation Accuracy Comparison

Key Observations:

- **Random Forest and Extra Trees show perfect training accuracy**, indicating potential **overfitting**.
- **XGBoost and LightGBM achieved the highest cross-validation accuracy (~98.4%)**, suggesting better generalization.
- **AdaBoost performed slightly worse** than the other boosting models but still remained competitive.

3.2.2. Confusion Matrix Analysis :

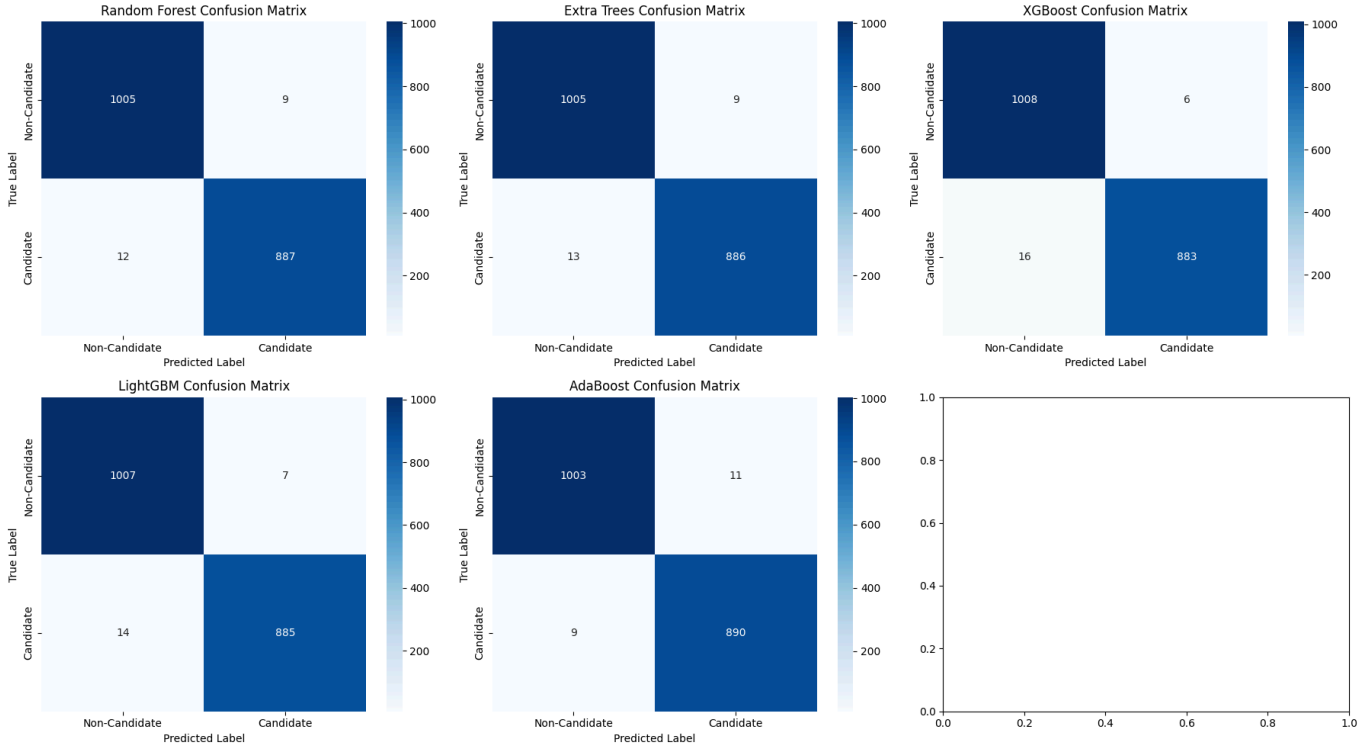


Figure : Model's Confusion Matrix

The confusion matrices reveal the classification performance of each model in terms of false positives (FP) and false negatives (FN).

- **XGBoost and LightGBM** exhibit the lowest false positives (FP), meaning they are better at avoiding incorrect identification of non-candidates as exoplanets.
- **AdaBoost** has the **lowest false negatives (FN)**, correctly identifying more true exoplanets than other models.
- **Random Forest and Extra Trees** show slightly higher misclassification rates compared to boosting models, indicating they may not generalize as well.

Overall, boosting-based models (XGBoost, LightGBM, and AdaBoost) demonstrate better balance between false positives and false negatives, making them preferable for exoplanet classification.

3.2.3 Precision-Recall & ROC Curve Analysis :

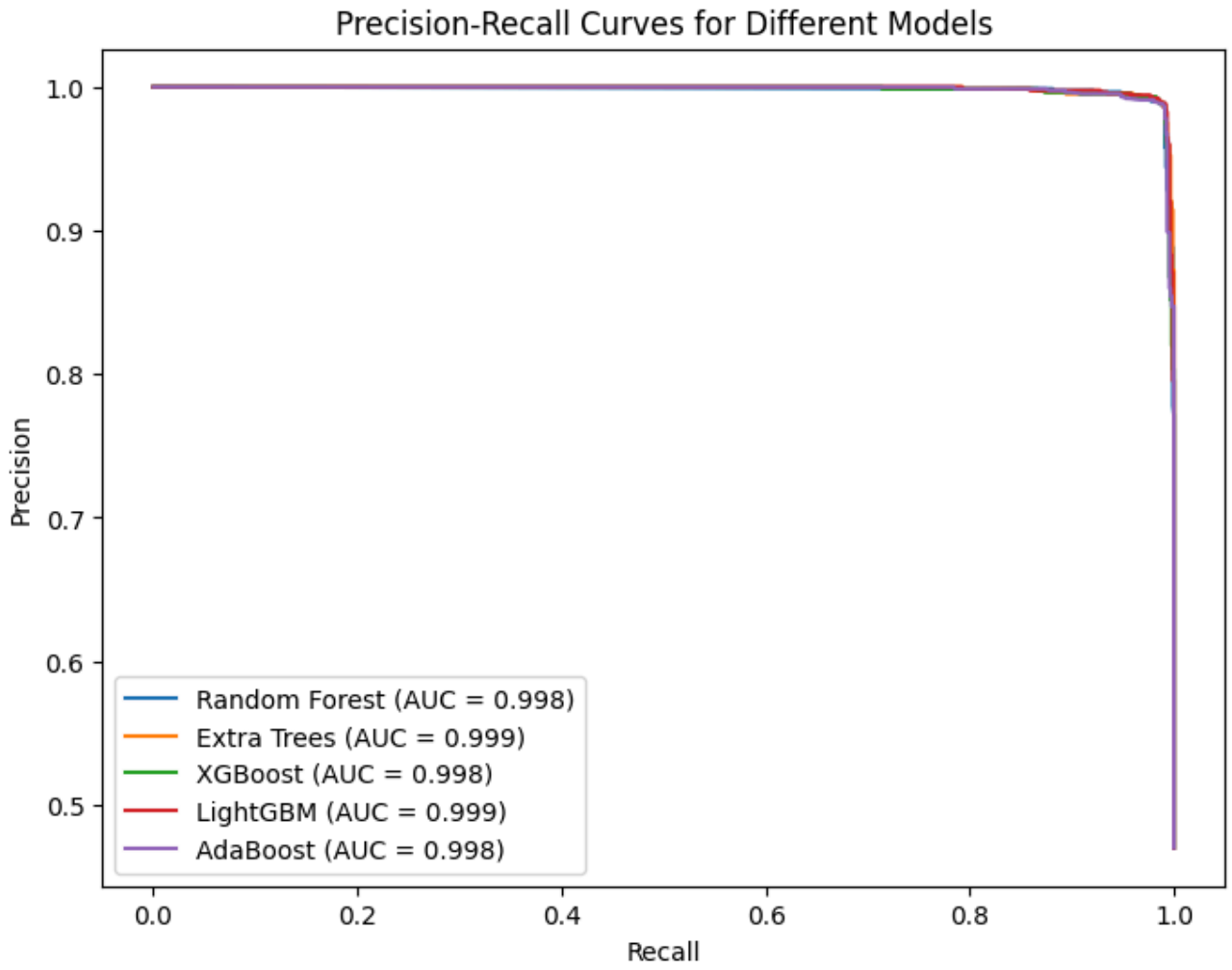


Figure : Model's Precision-Recall Curve

Precision-Recall (PR) curves help evaluate model performance, particularly in imbalanced datasets.

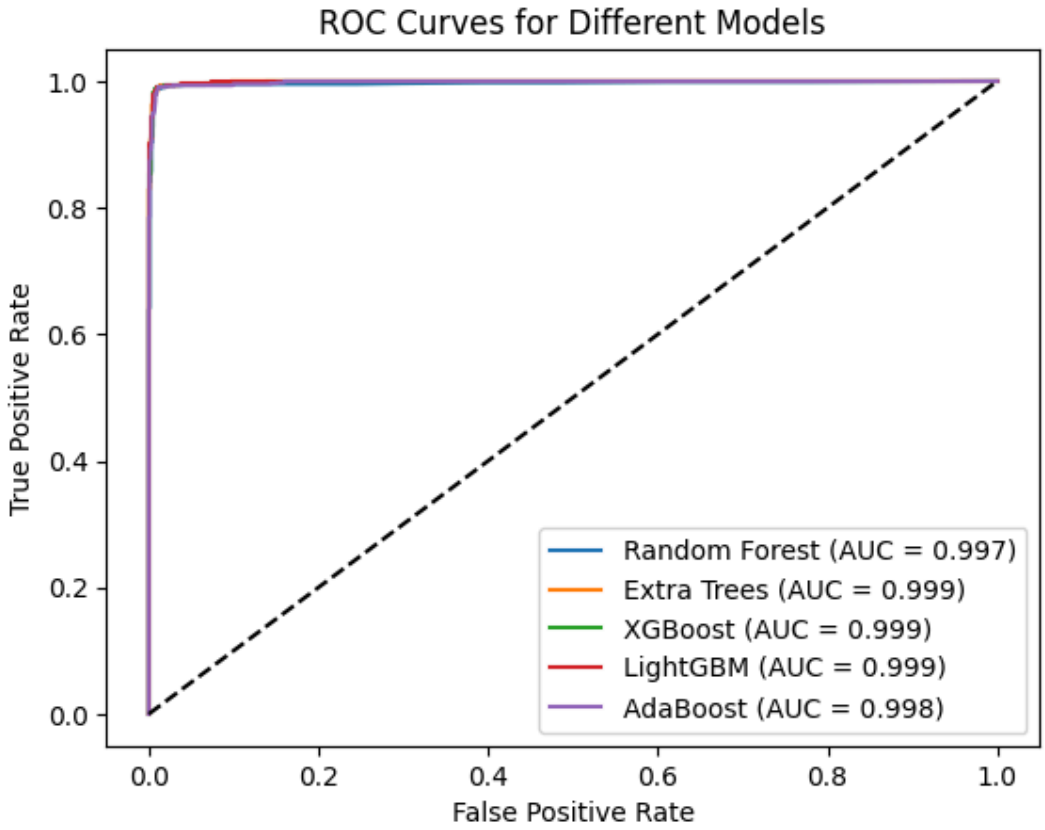


Figure : Model's ROC Curve

The ROC curves measure the models' ability to distinguish between exoplanets and false positives across different classification thresholds.

- **Extra Trees, XGBoost, and LightGBM** achieved **high AUC scores**, showing strong classification capabilities.
- **Random Forest also performed well**, but slightly lower than boosting-based models.
- **AdaBoost, while strong overall, had a slightly lower AUC compared to the others.**

Overall, all models achieve **high AUC scores (>0.98)**, indicating strong discriminatory power.

Since exoplanet detection requires **high precision (minimizing false positives) and high recall (minimizing false negatives)**, XGBoost and LightGBM remain the best choices, with AdaBoost also performing well.

4. Discussion & Recommendations

4.1 Addressing Overfitting

- Random Forest and Extra Trees models showed **100% training accuracy**, suggesting that they memorized the training data.
- This can be mitigated by tuning parameters like `max_depth` and increasing `min_samples_leaf` as we did in second model testing (RF2).

4.2 Boosting Models Performed Best

- **XGBoost and LightGBM outperformed other models**, making them the best choices for exoplanet detection.
- Further optimization of **learning rates and boosting strategies** could enhance performance.

4.3 Computational Efficiency

- **RandomizedSearchCV was preferred over GridSearchCV** for large parameter spaces, significantly reducing computation time.
- Future optimizations could include reducing feature space and limiting the number of hyperparameter combinations.

4.4 Final Model Recommendation

- **LightGBM or XGBoost** is the most suitable model due to its balance of **accuracy, computational efficiency, and generalization**.

5. Conclusion

This study successfully applied **ensemble learning methods** to classify exoplanet candidates with high accuracy (~98%). The results showed that **boosting models (XGBoost, LightGBM) were the most effective**, while **Random Forest and Extra Trees exhibited potential overfitting**. The combination of **feature selection, hyperparameter tuning, and boosting algorithms** significantly classification performance, especially in the cross-validation testing.

Future improvements could focus on **further reducing overfitting, exploring stacking model, and testing additional datasets** to validate our results.