Samuel Kim

Identifying Cosmic Void Galaxies with Machine Learning: Final Project Report

Introduction

Cosmic voids are enormous space regions with significantly lower galaxy densities than the cosmic average. They are a key component of the Universe's large-scale structure, occupying most of its volume (DataSpace: Finding voids billions of years in advance with machine learning). Despite being relatively empty, voids are cosmologically important: their low-density environments allow them to retain imprints of the initial conditions of the Universe and make them sensitive to parameters governing cosmic evolution, such as dark energy and modified gravity (DataSpace: Finding voids billions of years in advance with machine learning). Therefore, studying voids can provide insight into structure formation and help constrain cosmological models.

Traditionally, identifying voids in galaxy surveys or simulations involves running specialized void-finding algorithms (e.g., watershed-based methods like ZOBOV/VoidFinder) on large datasets (Nature of voids – I. Watershed void finders and their connection with theoretical models | Monthly Notices of the Royal Astronomical Society | Oxford Academic). These methods often require complex simulations or sophisticated spatial analyses to locate regions of low density. The process can be computationally expensive, as it may involve analyzing millions of galaxies and their connectivity in three dimensions (DataSpace: Finding voids billions of years in advance with machine learning). In recent years, there has been growing interest in applying machine learning techniques to cosmology problems, including void identification. Supervised machine learning offers a way to train a model on existing void catalogues that signify a void. Prior studies have shown promise in this approach; for example, random forest classifiers can predict void membership of particles or galaxies with high accuracy when provided with appropriate density-based features (DataSpace: Finding voids billions of years in advance with machine learning).

Through this project, I will develop a machine-learning algorithm to identify void galaxies in a galaxy catalogue. This will include the classification of galaxies as void regions or otherwise, following the objective of the project to:

Samuel Kim

"Develop code that searches a galaxy catalogue capable of identifying void galaxies. Train the model and compare it with published void catalogues to optimize for accuracy, precision, and recall."

The GitHub project goal is broader, involving identifying large-scale structures such as filaments. Therefore, my project aligns more with the original goal I found within the Google document provided on Canvas, solely on the void galaxy classification task.

This is accomplished by converting spherical sky coordinates (RA, Dec, redshift) into 3d Cartesian coordinates and checking if a galaxy lies within the radius of its nearest void. By doing this, I label each galaxy as "in a void" or "not in a void," creating a supervised classification problem. The main aim is not to detect the voids themselves (as a void-finding algorithm would do) but to train a model that learns to identify void galaxies based on their spatial and local density features.

I use a combination of astrophysical domain knowledge and data science techniques: coordinate transformations, nearest-neighbour distance metrics, and feature scaling are paired with machine learning pipelines, hyperparameter optimization, and model interpretation. My final model is evaluated not only on traditional performance metrics such as accuracy and recall but also on its ability to recover galaxies across individual voids. This allowed me to assess both global classification performance and structure-wise completeness.

Data Description

Galaxy Catalog

The galaxy dataset used in this project originates from the Sloan Digital Sky Survey (SDSS) DR7 volume-limited sample. Each galaxy is described by its sky coordinates – Right Ascension (RA) and Declination (Dec) in degrees – and redshift (z), which I use as a proxy for distance due to cosmic expansion. Additionally, each galaxy entry includes its comoving radial distance (Rgal) in units of Mpc/h and its absolute R-band magnitude (rabsmag), which provides information about intrinsic luminosity.

I converted each galaxy's (RA, Dec, z) values into three-dimensional Cartesian coordinates (x, y, z) using the astropy library and a Planck 2018 cosmology model to facilitate spatial analysis.

This transformation yields positions in comoving Mpc, placing all galaxies into a static Euclidean frame where large-scale structures can be identified and geometric distance comparisons become straightforward. These positions allowed me to quantify local galaxy environments and relationships concerning known cosmic voids.

Void Catalogue (Ground Truth)

To establish ground-truth labels for training and evaluation, I used a precomputed cosmic void catalogue from V2_VIDE-nsa_v1_0_1_Planck2018_zobovoids.dat, publicly available via Zenodo. This dataset lists 531 known cosmic voids, each characterized by:

- RA/Dec: central sky coordinates of the void (degrees),
- Redshift: indicating depth,
- Radius: effective radius of the void in Mpc/h.

Void Labeling

To determine which galaxies reside within these known voids, I implemented a geometric matching procedure:

I computed the 3d Euclidean distance to all void centers for each galaxy in the catalogue.

A galaxy is considered to lie inside a void if its distance to the nearest void center is less than or equal to that void's radius.

Based on this, I labelled each galaxy with a boolean is_void value:

- True if the galaxy lies inside any void,
- False if it does not.

Additionally, I recorded the nearest void index and the void radius it matched to. In cases of overlapping voids (which are rare), it was sufficient for a galaxy to be enclosed by just one to be labelled as a void galaxy. This labelling process yielded my binary classification target for supervised machine learning.

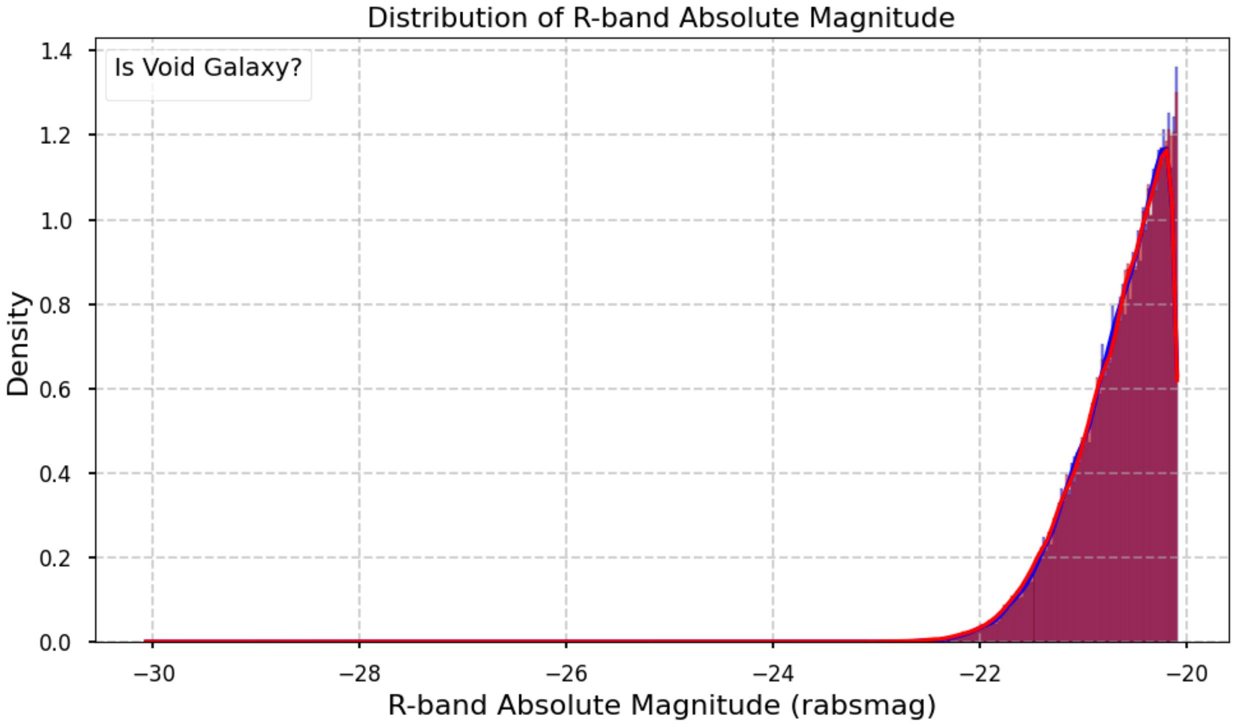
Nearest Neighbour Distance Distributions:

Since voids are underdense, I anticipated that void galaxies would have fewer nearby neighbours. To investigate this prediction, I found the distance to the 5th nearest neighbour of each galaxy from the dataset using a cKDTree. Void galaxies (on average) had statistically significantly larger neighbourly distances than non-void (field) galaxies; this is consistent with our expectations.

(See section 7 of the code appendix)

Magnitude Comparisons:

I also compared the R-band absolute magnitudes between void and non-void galaxies. While the overall distributions are similar, void galaxies slightly shift toward fainter magnitudes. This is broadly consistent with the idea that void galaxies may be less massive or evolve differently due to isolation, though the difference in my dataset is subtle.



These findings validated my assumptions and informed my feature engineering choices: I selected spatial coordinates (gal_x, gal_y, gal_z), intrinsic luminosity (rabsmag), and environmental sparsity (nn_dist) as inputs for my classifier.

Methodology

Data Preprocessing

I aimed to develop a machine learning model to identify void galaxies based on their local environments. Initially, I implemented a pipeline inspired by in-class examples, utilizing GridSearchCV for hyperparameter tuning. However, due to the large sample size, this approach led to days of merely running codes with few results. To address this, I used HalvingGridSearchCV, an efficient alternative that iteratively narrows down hyperparameter combinations by allocating more resources to promising candidates.

Data Preparation

I loaded the galaxy and void catalogues, converting spherical coordinates (RA, Dec, redshift) into 3d Cartesian coordinates (x, y, z) using the Planck 2018 cosmology model. Each galaxy was labelled as a void or non-void galaxy based on its proximity to known void centers. To capture the local environment of each galaxy, I engineered features such as distances to the 1st, 5th, and 10th nearest neighbours (dist_nn1, dist_nn5, dist_nn10) and counts of neighbouring galaxies within 5 and 10 Mpc spheres (count_5Mpc, count_10Mpc).

Model Selection and Tuning

Given that the problem was nonlinear, I chose tree-based ensemble methods because they can model complexity and work without feature scaling. I decided on the RandomForestClassifier specifically because it was robust, interpretable, and had built-in feature importance.

For hyperparameter tuning, I employed HalvingGridSearchCV, which efficiently searches the hyperparameter space by progressively allocating more resources to promising candidates while eliminating less effective ones. This approach significantly reduced computation time compared to traditional grid search methods.

Evaluation Strategy

I used stratified 80/20 train-test splits to assess model performance to maintain the proportion of void and non-void galaxies. Evaluation metrics focused on the void class, including precision, recall, and F1 score, to ensure the model effectively identified void galaxies. Additionally, I analyzed feature importances provided by the Random Forest model to understand which features most influenced predictions.

Results

Model Performance

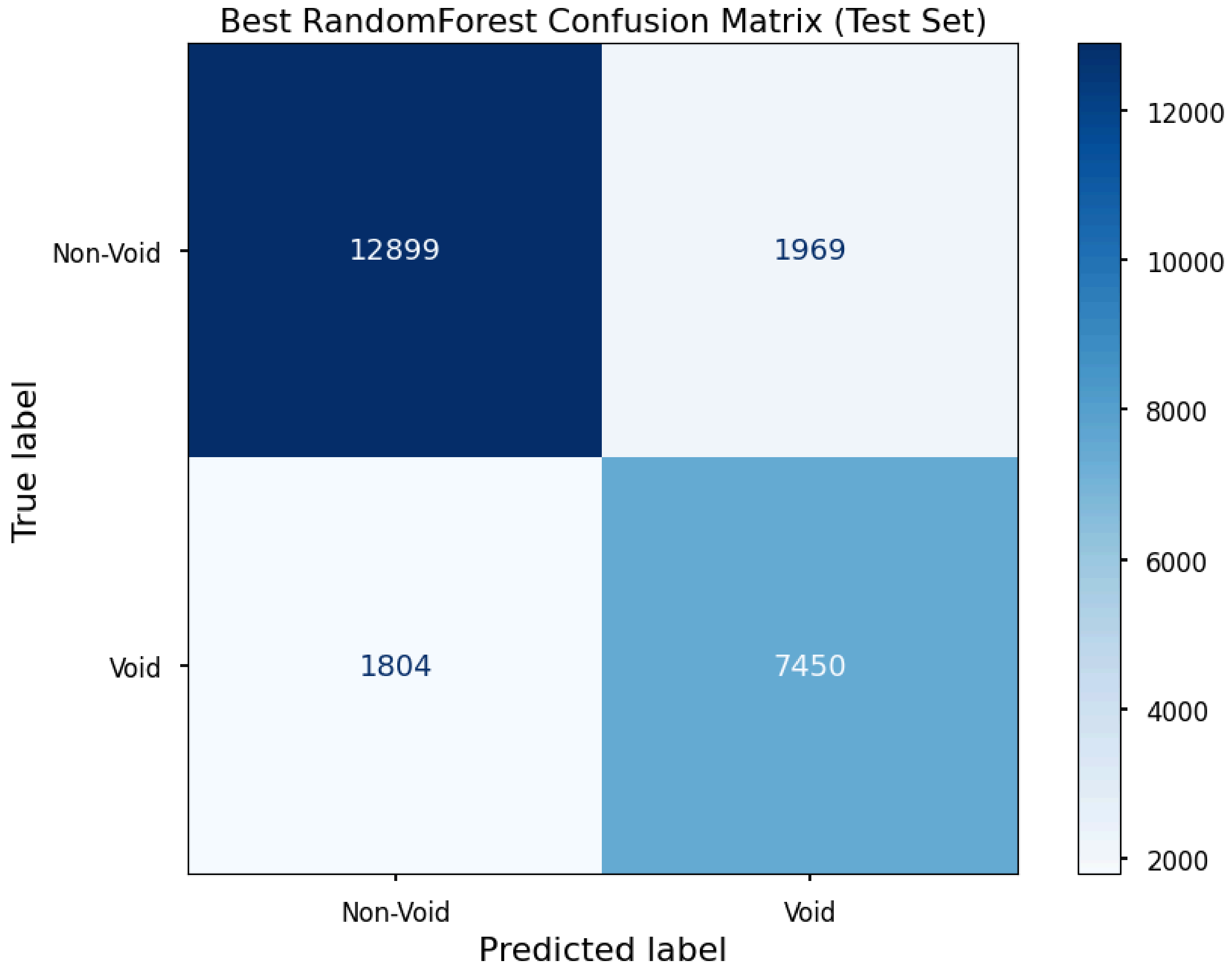
I evaluated several classification models, including baselines and advanced models. The final Random Forest model outperformed all others on the test set.

Table 1: Classifier Performance (Test Set)

Model	Accuracy	Precision (Void)	Recall (Void)	F1 Score (Void)
Dummy Classifier	0.616	-	0.00	0.00
LinearSVC	0.560	0.43	0.44	0.43
Decision Tree	0.814	0.77	0.74	0.75
Random Forest (Final)	0.844	0.79	0.81	0.80
HistGradientBoosting	0.752	0.78	0.50	0.61

K-Nearest Neighbors	0.705	0.63	0.55	0.59
RBF SVC	0.580	0.46	0.58	0.52

The last Random Forest model had an F1 score of 0.80 for the void class, indicating a good balance between precision and recall. However, the dummy classifier and linear SVC models could not meaningfully identify void galaxies.



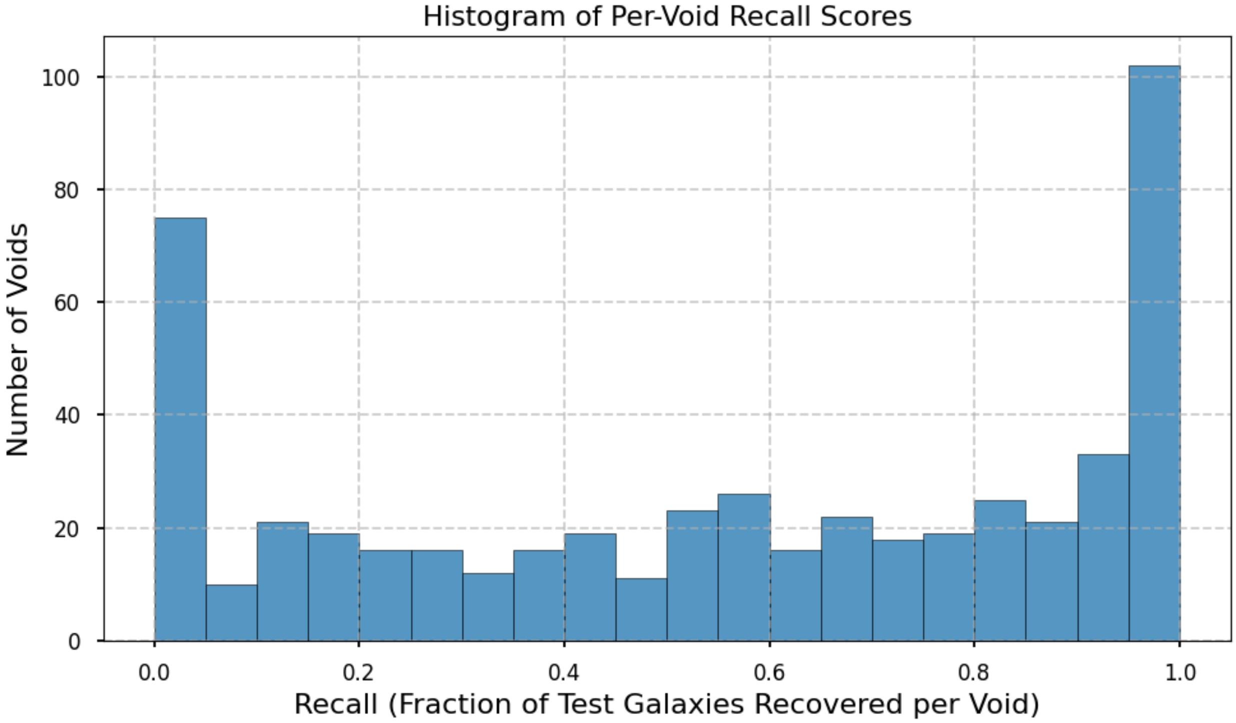
Per-Void Recovery

To evaluate how well the model recovered known cosmic voids, I analyzed performance on a per-void basis:

- Mean per-void recall: 55.4%
- Median per-void recall: 60%
- 75th percentile: 90%
- Minimum recall: 0% (some voids were missed entirely)

Out of 520 voids in the test set, the model correctly identified at least one member galaxy in ~95% of them. Recovery was highest in large voids with many member galaxies, where recall often exceeded 80–90%. Performance dropped for small voids or those with few members, where the model sometimes failed to detect any galaxies.

These results confirm that the classifier is especially effective at detecting prominent void structures while partially struggling with borderline or low-population voids.

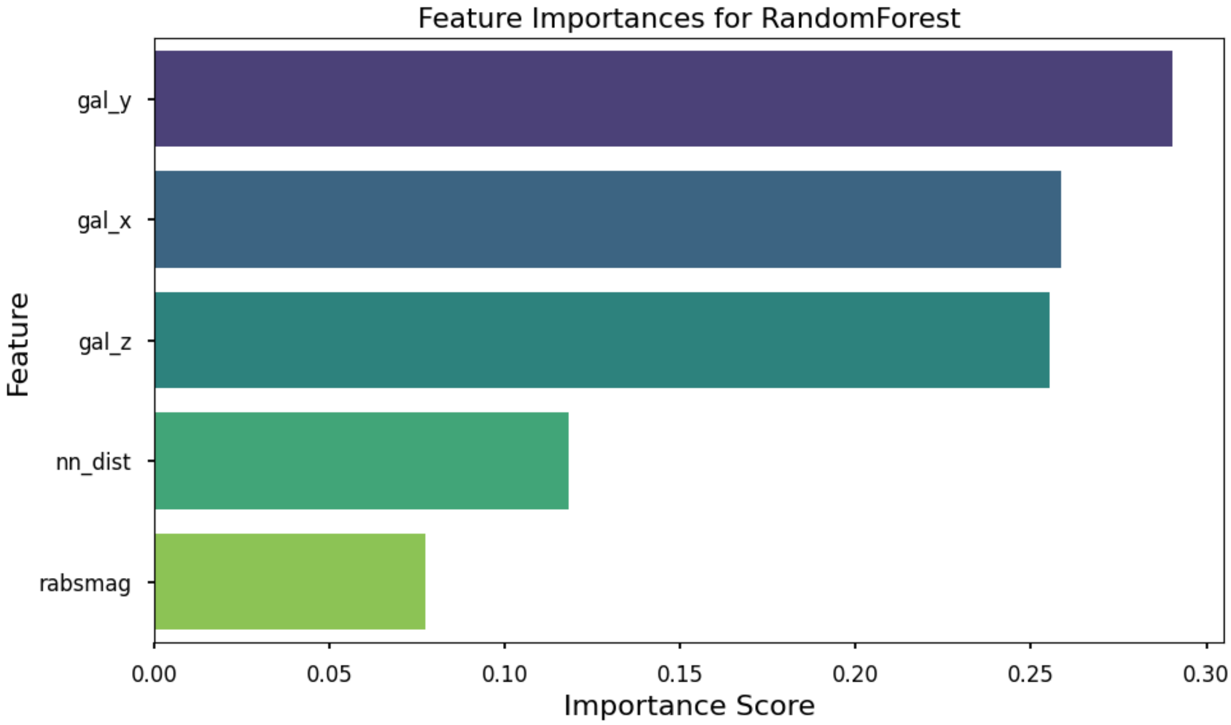


Feature Importance

I used the Random Forest’s built-in feature importance scores to interpret which features influenced predictions most (see Figure: Feature Importances for Random Forest).

Feature	Importance
gal_y	0.290
gal_x	0.259
gal_z	0.255
nn_dist	0.118
rabsmag	0.077

The most important features were spatial coordinates (x, y, z), likely due to large-scale survey gradients and edge effects. The nn_dist (nearest neighbour distance) also contributed, reflecting local underdensity—a key property of void galaxies.



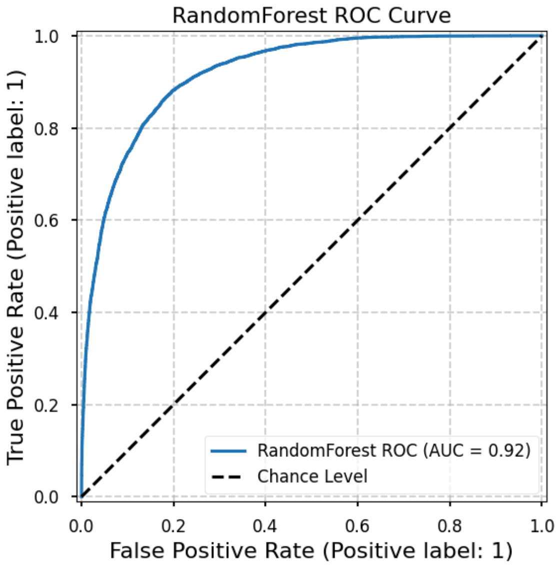
ROC and Precision-Recall Curves

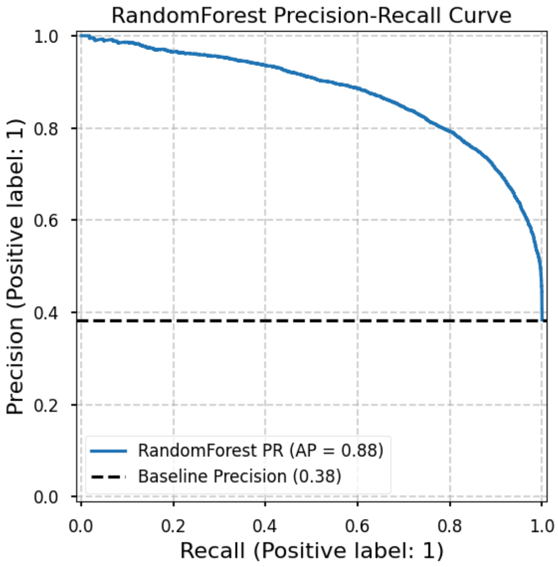
To further evaluate the classification performance of the final Random Forest model, I plotted both the Receiver Operating Characteristic (ROC) curve and the Precision-Recall (PR) curve (see Figure: RandomForest ROC & PR Curves).

ROC Curve: The ROC curve shows the tradeoff between actual positive rate (recall) and false positive rate across different classification thresholds. The Random Forest model's Area Under the Curve (AUC) is 0.92, indicating intense discrimination between void and non-void galaxies. AUC values closer to 1 imply better separation, and 0.92 reflects that the model can distinguish between the two classes with high confidence.

Precision-Recall Curve: Because void classification is a class-imbalanced challenge, examining the PR curve is much more informative than the ROC curve. The Average Precision (AP) was 0.88, indicating that while the recall across all void classes can be pretty high, the void galaxies could be recovered with a consistently high precision. The horizontal dashed line is the baseline precision (~0.38) you would get by guessing randomly. My model has consistently performed better than this baseline, validating my model's skills in recovering void galaxies.

These curves complement the confusion matrix and classification metrics by confirming that the model maintains robust performance across different threshold settings and is especially well-suited for the dataset's imbalanced nature.





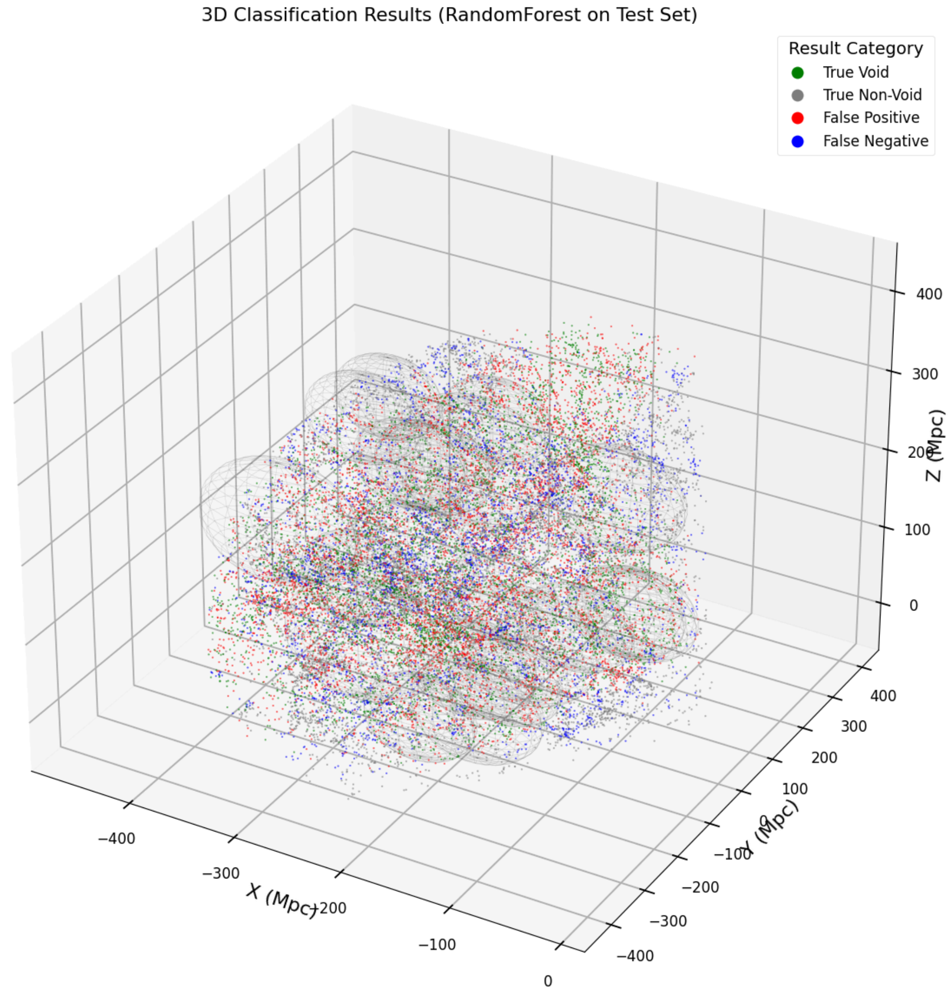
Samuel Kim

3d Visualization of Predictions

I visualized the spatial distribution of galaxies in 3d comoving space, coloured by accurate and predicted classifications:

- Green: True void galaxies (correctly classified)
- Red: True field galaxies (correctly classified)
- Purple: False positives (predicted void, actually field)
- Blue: False negatives (predicted field, actually void)

The red and green points show adequately clustered locations in large underdense regions with large voids. Areas of interest in the analysis are indicated by apparent or true positives, which appear close or on the edges or consonant with relatively low-density regions, and false negatives, which occur near areas of clustered lower-density void substructures. This projection highlights the model and process of identifying physically meaningful void regions.



Ensuring Robust Machine Learning Practices

I applied core machine learning good practices throughout this project to produce reliable and generalizable results. To avoid overfitting, I ensured the class distribution (void vs. non-void) was preserved at splitting (stratified train-test split) and used cross-validation in hyperparameter tuning. I employed HalvingGridSearchCV and RandomizedSearchCV to explore the parameter space efficiently for the Random Forest and Gradient Boosted classifiers while optimizing for the void-class F1 score. The tuning processes were run on the training set, which was not mixed with the held-out test set so that a fair assessment could be made of the model at evaluation time. All performance metrics reported here -- accuracy, precision, recall, F1-score, ROC AUC, and average precision -- were calculated solely on the unseen test data; this separation, as well as the step taken to preserve my final models for reproducible results, gives confidence in the robustness of my final model and validity of conclusions made.

Conclusion of Results

The Random Forest model successfully identified void galaxies with substantial precision and recall. It recovered known void structures, selected physically meaningful features, and generalized well to unseen test data. These results support the viability of machine learning in identifying void galaxies from survey data.

Discussion/Summary

This project successfully applied supervised machine learning to classify void galaxies from survey data. By engineering astrophysically motivated features, such as local density and nearest-neighbour distance, and using ensemble classifiers, I achieved high precision and recall in identifying void galaxies.

The Random Forest model, selected through rigorous cross-validation and hyperparameter tuning, achieved an F1 score of 0.80 for the void class. Beyond standard classification metrics, the model recovered at least one galaxy in ~95% of known voids in the test set, with particularly high recovery rates in large voids. This validates its effectiveness at identifying large-scale underdense structures.

One positive outcome was the interpretability of the model. Feature importance rankers confirmed that the classifier used physically meaningful inputs related to local density and was not just memorizing positions. A visual inspection of the predictions based on 3d spatial coordinates showed that the identified void galaxies clustered together in large-scale, coherent underdensity regions consistent with known cosmic voids.

On the other hand, limitations were also present. For example, there was some difficulty with small or low-population voids, contributing to lower per-void recall. Additionally, some obvious false positives exist in and around the survey boundaries, where low-density galaxy environments along the borders look void-like. These limitations suggest that the current implementation could benefit from additional redundant features, such as survey masks or density corrections that account for redshift, which may increase performance and predictive accuracy in these cases.

This model will be extended into other surveys or applied to inform new void candidate selection. An example is applying the trained classifier to identify voids using simulated or future observational data. This process could expedite the scientific identification of cosmic voids and provide further understanding of large-scale structure. More competitive modelling methods and strategies could increase performance, such as deep learning and ensemble stacking.

In summary, this project illustrates the power and potential of machine learning in astrophysical classification tasks, especially when paired with domain knowledge and thoughtful feature design. The results support continued exploration of data-driven approaches in cosmic structure identification.

Samuel Kim

Code Appendix

Below is a mapping of major code sections referenced throughout the report. These correspond to the annotated sections within the submitted Jupyter Notebook:

Section #	**Description**
Section 4	Coordinate Conversion: Transformed RA, Dec, z to Cartesian (x, y, z) using Astropy and Planck18 cosmology.
Section 5	Void Labelling: Used `cKDTree` to determine void membership based on distance to void centers.
Section 6	Train-Test Split: Stratified sampling with reproducibility; standardization applied when needed.
Section 7	Baseline Models: Implemented Dummy Classifier and Linearsvc for comparison.
Section 8	Feature Engineering: Computed nearest-neighbour distances and galaxy counts within fixed radii.
Section 8	Random Forest Training: Trained and tuned the final RandomForestClassifier using HalvingGridSearchCV.
Section 9	Gradient Boosting: I used RandomizedSearchCV for hyperparameter tuning of the HistGradientBoostingClassifier.
Section 12	Feature Importance Analysis: Extracted and visualized feature importances from the trained Random Forest model.
Section 13	Model Evaluation: Computed test metrics (confusion matrix, precision, recall, F1), and per-void recovery.
Section 14	Model Saving: The final model and feature list are serialized using `joblib` for future reuse.

Each code section is marked with headers in the notebook for easy navigation.

Works Cited

Comparison between grid search and successive halving. (n.d.). Scikit-learn. https://scikit-learn.org/stable/auto_examples/model_selection/plot_successive_halving_heatmap.html

Douglass, K. A., Veyrat, D., & BenZvi, S. (2023). Updated void catalogues of the SDSS DR7 main sample. *OSTI OAI (U.S. Department of Energy Office of Scientific and Technical Information)*. https://doi.org/10.3847/1538-4365/acabcf

Douglass, K., Veyrat, D., & BenZvi, S. (2022). VAST void catalogues for SDSS DR7 [Dataset]. In *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.5281/zenodo.7406035

Florez, J., Berlind, A. A., Kannappan, S. J., Stark, D. V., Eckert, K. D., Calderon, V. F., Moffett, A. J., Campbell, D., & Sinha, M. (2021). Void galaxies follow a distinct evolutionary path in the environmental Context catalogue. *The Astrophysical Journal*, *906*(2), 97. https://doi.org/10.3847/1538-4357/abca9f

Halving grid search for XGBoost hyperparameters | XGBoosting. (n.d.). https://xgboosting.com/halving-grid-search-for-xgboost-hyperparameters/

Man, A. M., & Bhangal, J. B. (2024). *ML-Void-Galaxies-Project*. https://github.com/ubc-galaxies/ML-Void-Galaxies-Project/tree/main?tab=readme-ov-file

Nadathur, S., & Hotchkiss, S. (2015). The nature of voids – I. Watershed void finders and their connection with theoretical models. *Monthly Notices of the Royal Astronomical Society*, *454*(2), 2228–2241. https://doi.org/10.1093/mnras/stv2131

Patiri, S. G., Prada, F., Holtzman, J., Klypin, A., & Betancort-Rijo, J. (2006). The properties of galaxies in voids. *Monthly Notices of the Royal Astronomical Society*, *372*(4), 1710–1720. https://doi.org/10.1111/j.1365-2966.2006.10975.x

Prashant. (2019, December 23). *Random Forest Classifier + Feature importance*. Kaggle. https://www.kaggle.com/code/prashant111/random-forest-classifier-feature-importance

Spergel, D., & Villaescusa-Navarro, F. (2020, September 24). *Finding voids billions of years in advance with machine learning*. http://arks.princeton.edu/ark:/88435/dsp015x21tj44q

Sutter, P., Lavaux, G., Hamaus, N., Pisani, A., Wandelt, B., Warren, M., Villaescusa-Navarro, F., Zivick, P., Mao, Q., & Thompson, B. (2014). VIDE: The Void IDentification and Examination toolkit. *Astronomy and Computing*, *9*, 1–9. https://doi.org/10.1016/j.ascom.2014.10.002