AI Boot Camp Project 2

Earthquake Impact

Team Members:

(Pedro Zurita)

Christoph Guenther

Ashwini Kumar

Background

- **Impact of Earthquakes:** can lead to significant destruction, economic loss, and fatalities due to their sudden and unpredictable nature.

- **Challenges in Prediction:** accurately forecasting location, and magnitude of earthquakes remains a major challenge in the field of seismology.

Project Objectives

1. **Predict Impact of Earthquakes in the US**
2. **Evaluate Various ML Classifiers**
3. **Build Model Using Best Classifier**

Metric to Evaluate Classifiers: balanced accuracy score on test dataset

Process

Overview of data collection, cleanup and exploration process

1. **Data Retrieval**:
    a. USGS Earthquake data
    b. ISRIC Soil Density data
2. **Data Pre-processing:**
    a. Remove NaN’s
    b. Define Target (Modified Mercalli Intensity (MMI) scale)
    c. Address data leakage
    d. Scale data

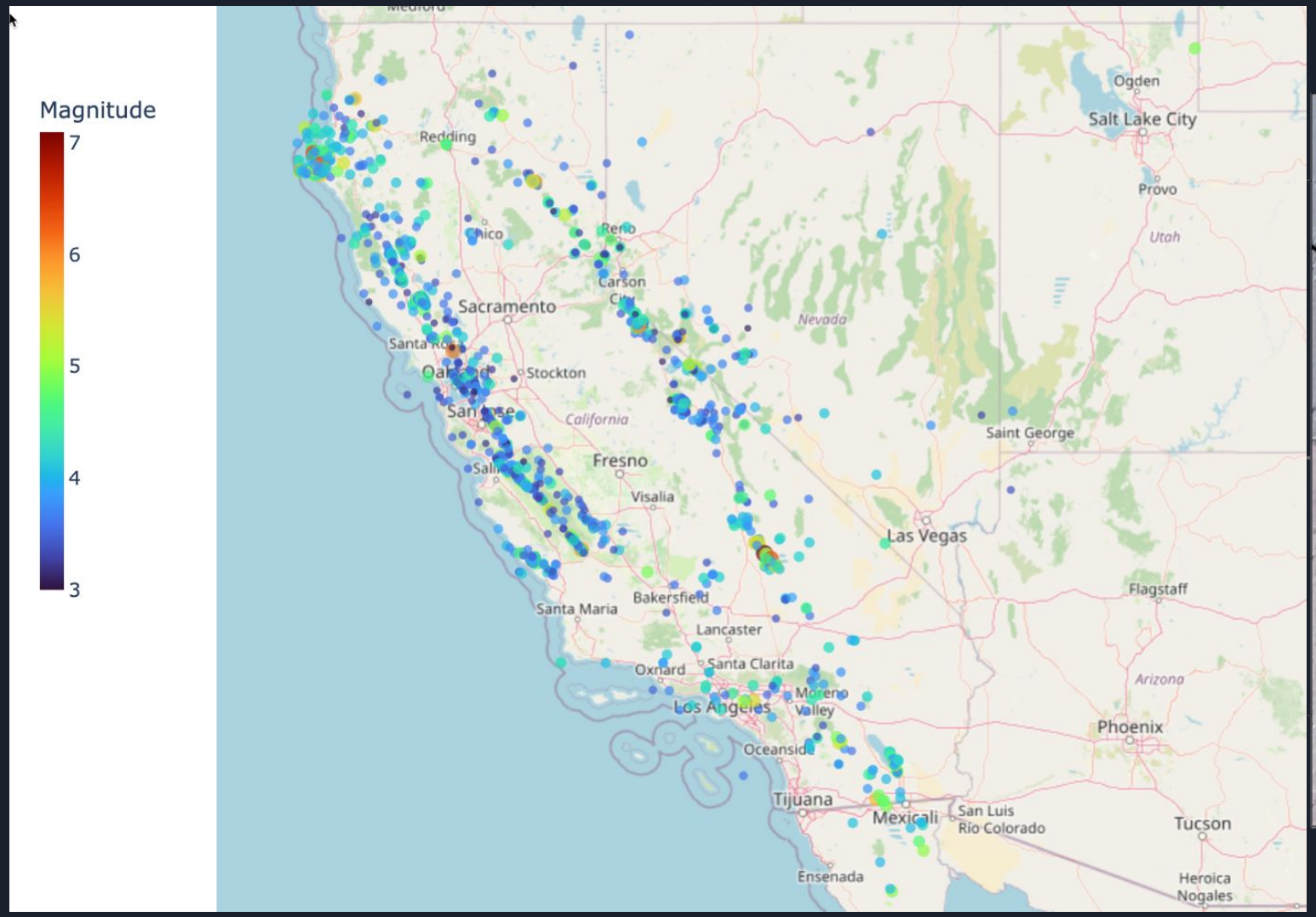
Model Optimization

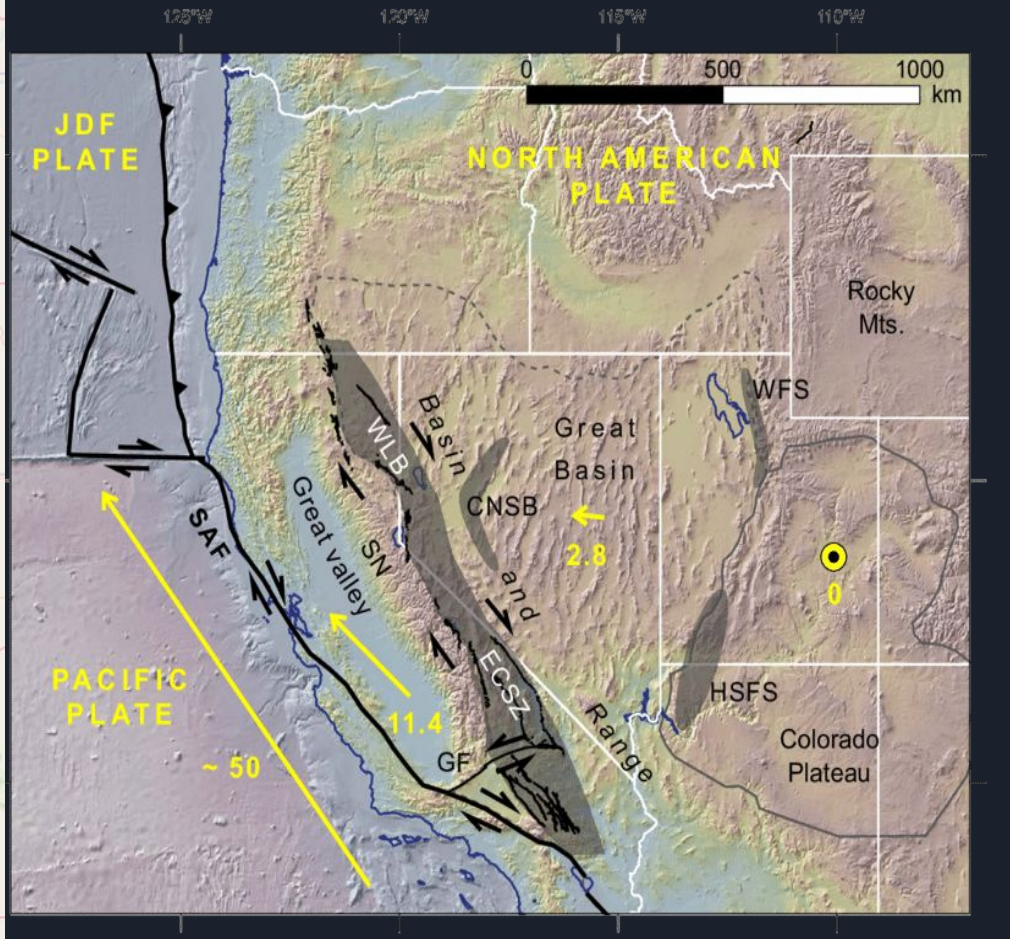
1. **Split Data into Train & Test Data**
2. **Evaluate Classifiers**:
    a. Random Forest Classifier (RFC)
    b. Multinomial Logistic Regression
    c. Support Vector Machine (SVM)
    d. K Nearest Neighbors (KNN)
    e. Decision Tree (DT)
3. **Address Over-Fitting**
    a. P-Values
    b. PCA
    c. Hyperparameter Tuning
4. **Address Random State Variable Dependency**
5. **Build Model with Best Performing Classifier**

Challenges Encountered

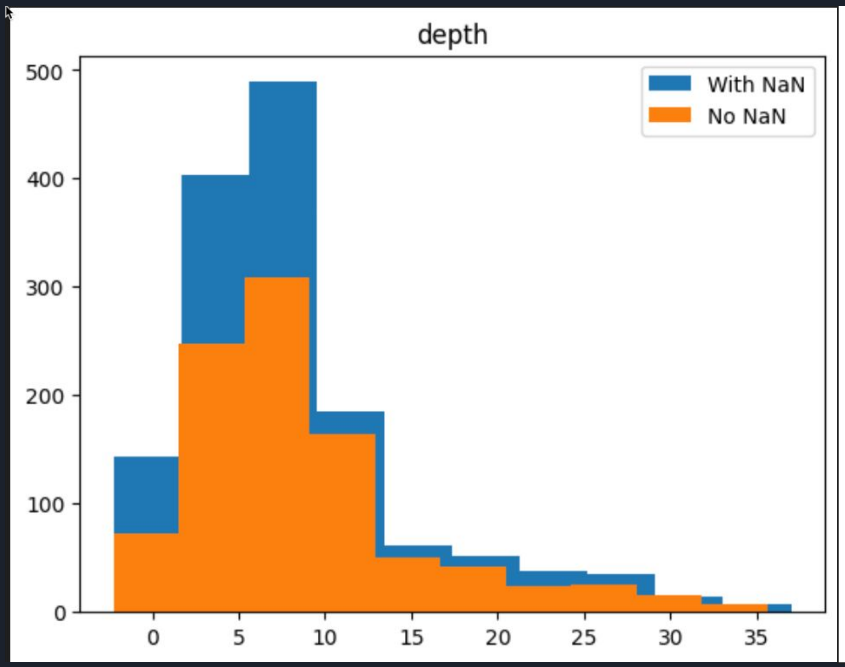
- Unable to Reach Accuracy Score Target
- Lack of Geological and Soil Science Domain Knowledge
- Combining Soil and Earthquake Data
- Over-Fitting of Best Classifier
- Dependency of Classifier Performance on Random State Variable

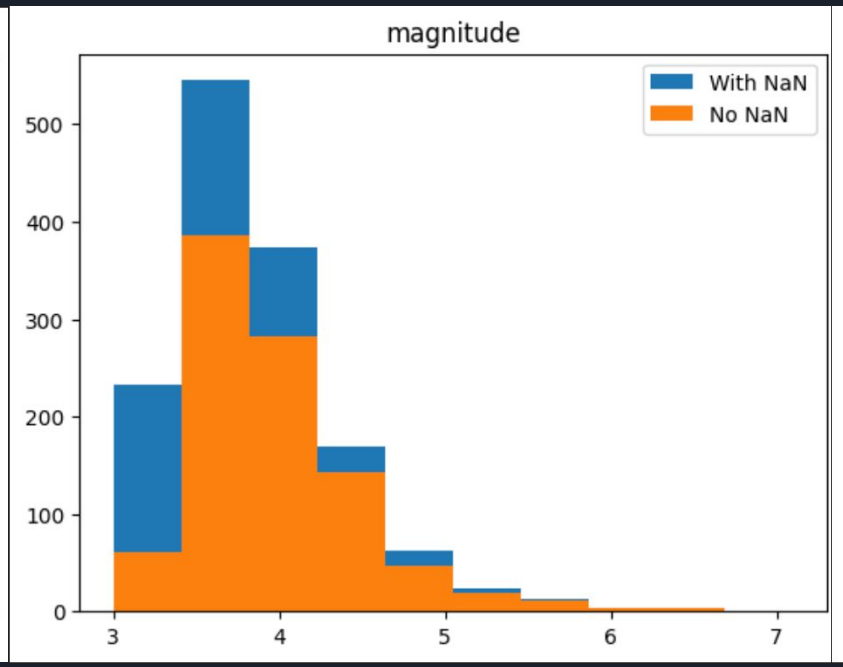
Earthquake Occurrences

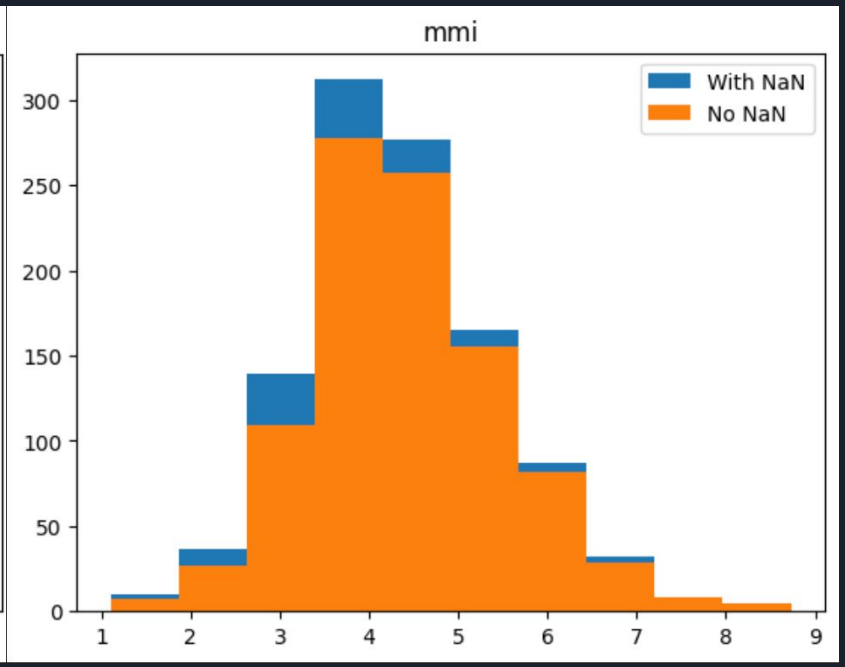




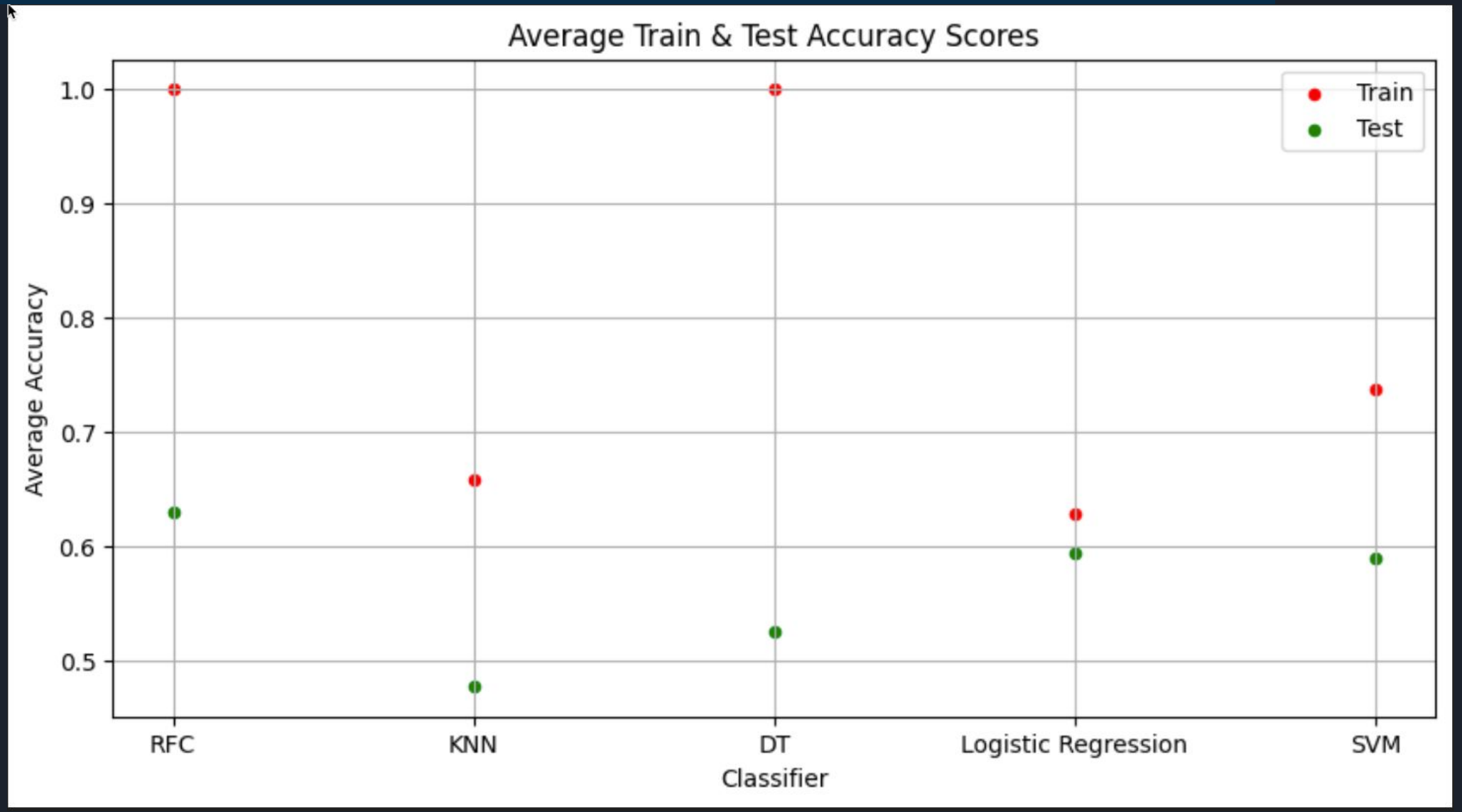
Effect of Removing NaN Values



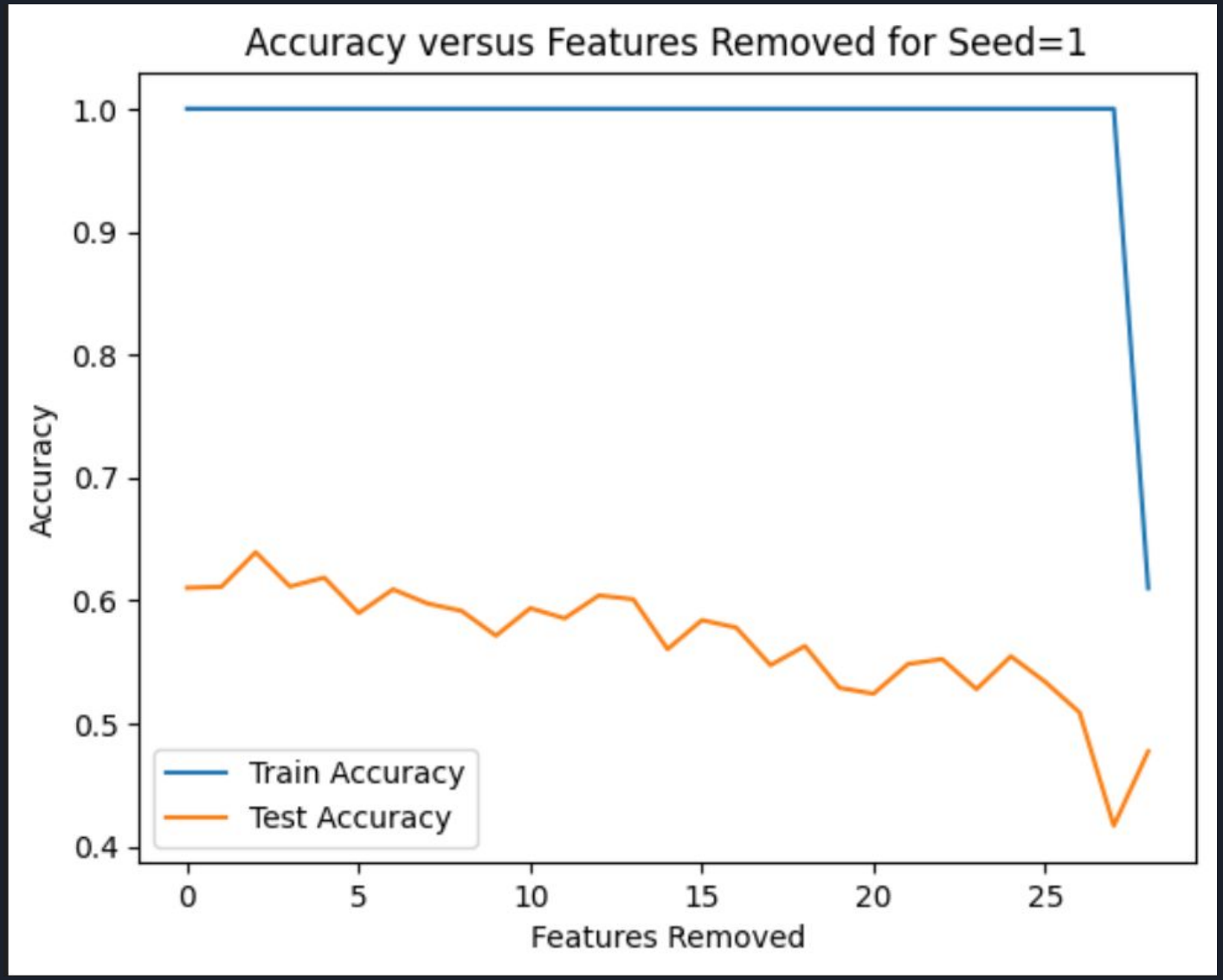




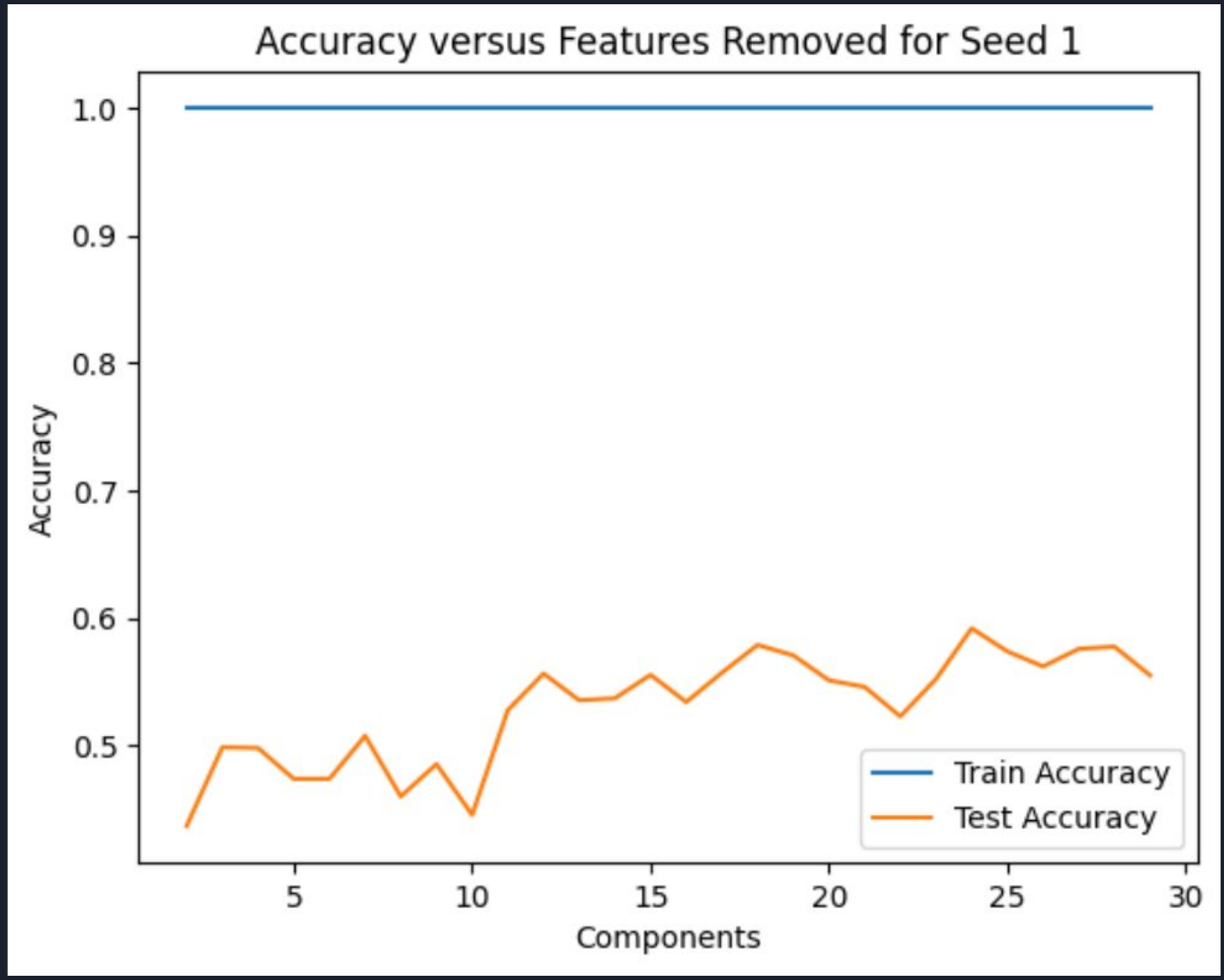
Evaluating Classifiers



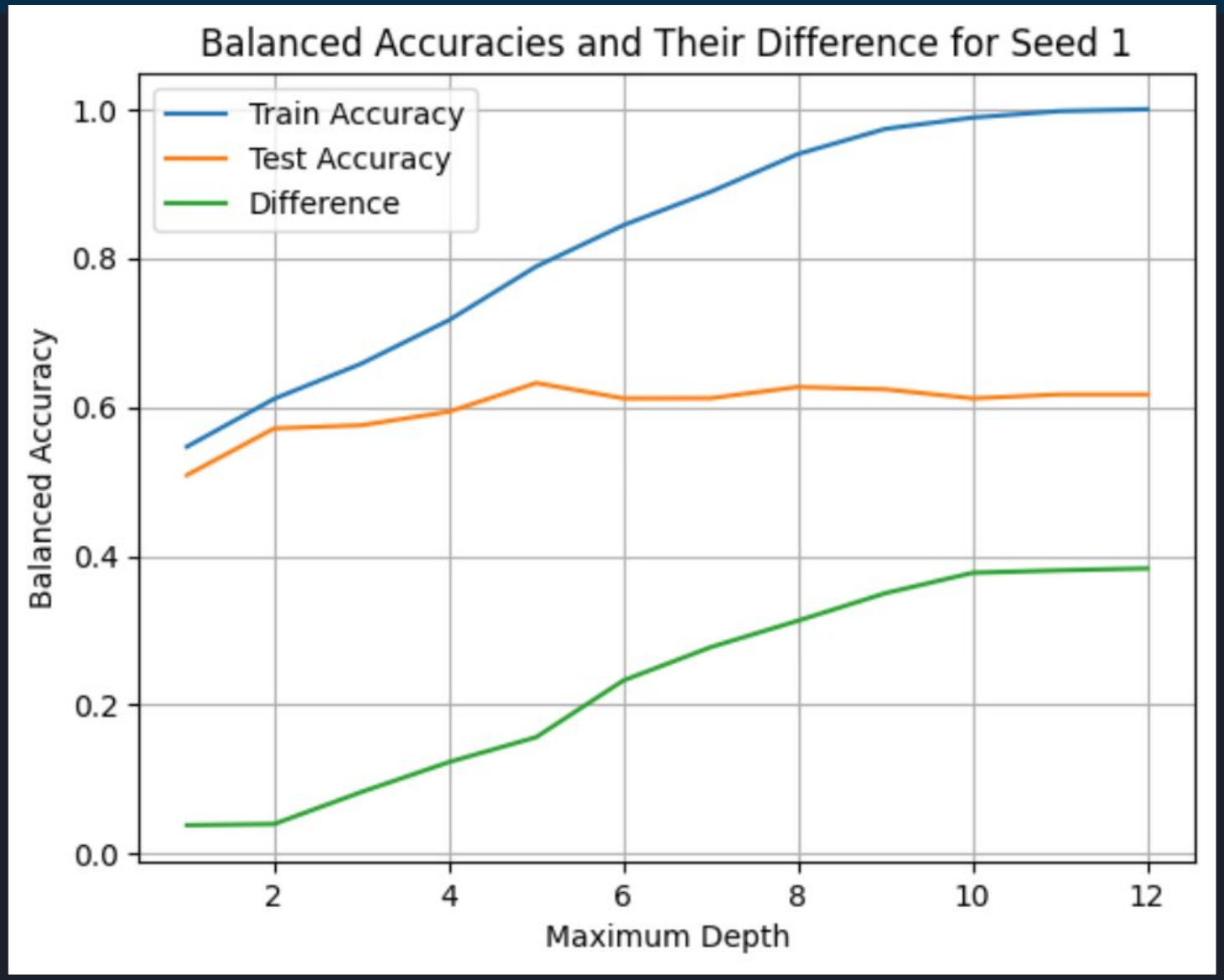
Addressing RFC Over-Fitting: P-Values



Addressing RFC Over-Fitting: PCA



Addressing RFC Over-Fitting: RFC Hyperparameter Tuning



Final Model Performance

Preprocessing data ...

Dropping rows with 'NaN' values:
Dropped 32.98% of rows.
There are 957 rows remaining.
There are 37 columns remaining.

Dropping columns:
Number of rows remaining: 957.
Number of columns remaining: 31.

Creating X and y:
Number of rows in X: 957.
Number of columns in X: 30.
Number of elements in y: 957.

==========

Best Classifier: RandomForestClassifier(max_depth=6):
Balanced Train Accuracy Score: 0.842.
Balanced Test Accuracy Score: 0.610.

Classification Report:

	precision	recall	f1-score	support
0	0.66	0.69	0.67	91
1	0.53	0.62	0.57	87
2	0.76	0.52	0.62	62
accuracy			0.62	240
macro avg	0.65	0.61	0.62	240
weighted avg	0.64	0.62	0.62	240

Future Considerations

- **PSHA (Probabilistic Seismic Hazard Assessment)**
  - Risk analysis over a given period and return periods (e.g “once in 100 years”, etc**)**
- **DSHA (Deterministic Seismic Hazard Analysis)**
  - Scenario analysis & Confidence estimates
- **Correlate Soil Density with MMI Measurement**
  - Where was the MMI recorded, and what soil data was collected at that location
- **Better Understanding the Geology of Earthquakes**

Questions?