



RANDOM FOREST

GROUP 5: KRIPANJALI DHUNGANA,
JACK LYNN, NORMAN MORRIS, &
SAM WAINRIGHT

INTRODUCTION

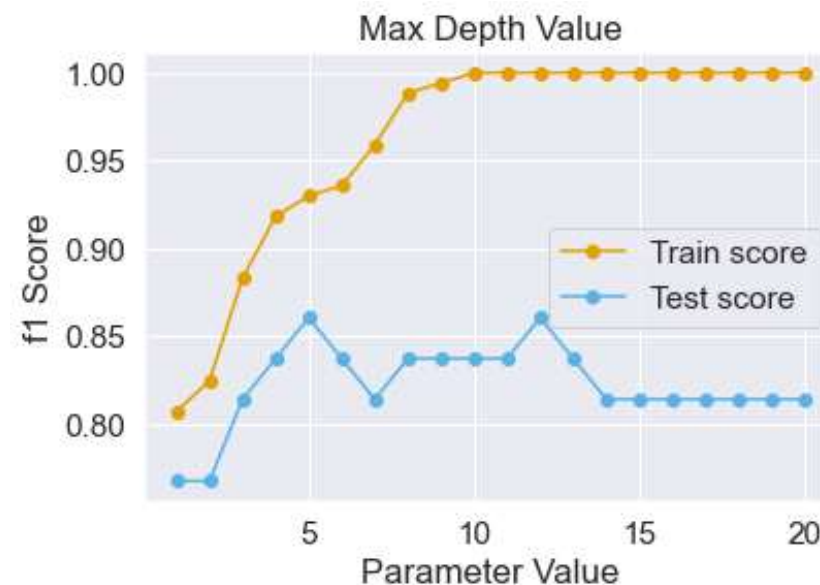
- Glass Identification dataset: **Can you correctly identify glass type?**
- Attribute Information: **RI, Na, Mg, Al, Si, K, Ca, Ba, Fe**
- Baseline models: **k-NN** and **Random Forest**
- Optimization models:
 - **Manual Hyperparameter** Training
 - **GridSearch Hyperparameter** Training
 - **Limited Input Columns** Training



The image features a complex network of nodes and edges on a black background. Nodes are represented by small white circles, some of which are double-outlined. The edges are thin lines in blue and orange, forming a dense web that curves across the frame. The word "BASELINE" is centered in a bold, white, sans-serif font.

BASELINE

	precision	recall	f1-score	support
Building	0.81	0.93	0.87	28
Container	1.00	0.50	0.67	2
Headlamp	0.86	1.00	0.92	6
Tableware	1.00	1.00	1.00	2
Vehicle	0.00	0.00	0.00	5
accuracy			0.81	43
macro avg	0.73	0.69	0.69	43
weighted avg	0.74	0.81	0.77	43



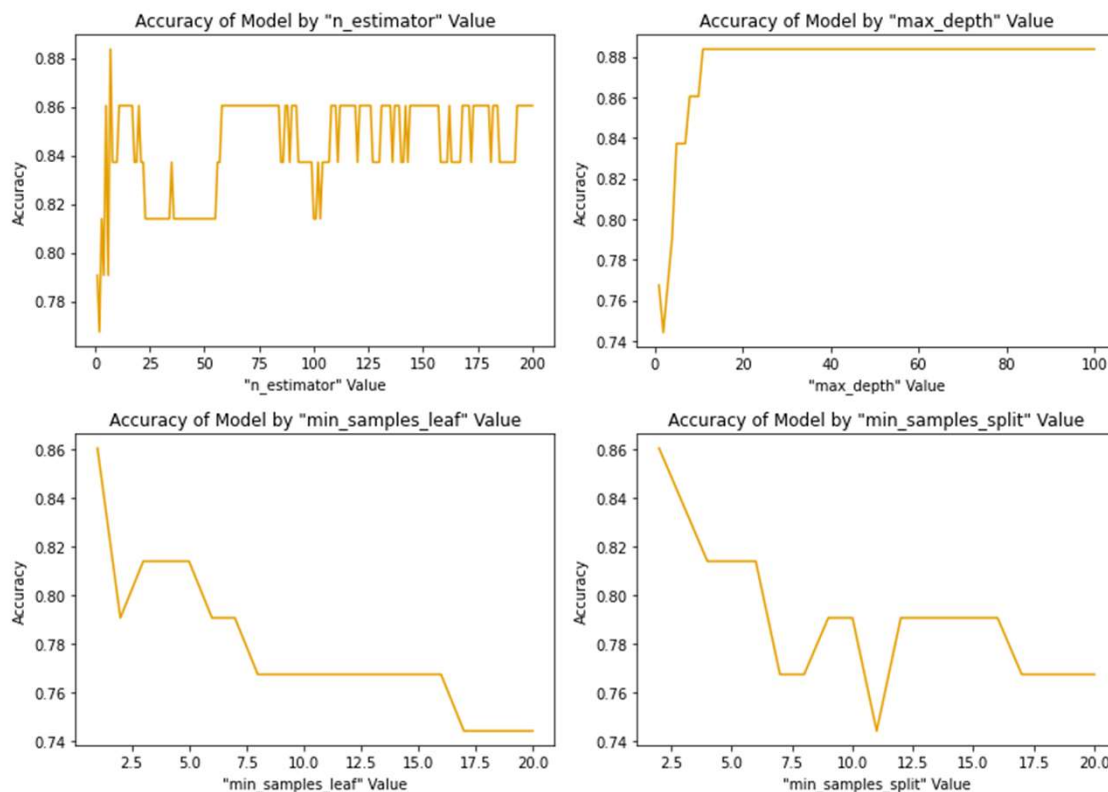
BASLINE: UNALTERED RANDOM FOREST

- Default hyperparameters set `n_estimators = 100`; `random_state = 0`; `max_depth = 2`
- Compare max depth 2 on x axis

The background is a dark, almost black, field filled with a complex network of thin, glowing lines in shades of blue and orange. These lines are interconnected by small, white, circular nodes or dots. The lines and nodes are arranged in a way that suggests a dynamic, evolving system, possibly representing a network or a process of optimization. The lines are more concentrated on the right side of the image, where they form a dense, curved structure that resembles a stylized 'C' or a series of overlapping arcs. On the left side, the lines are more sparse and form a more open, star-like pattern. The overall effect is one of a complex, interconnected system that is constantly changing and evolving.

OPTIMIZATION

MODEL 1: MANUAL HYPERPARAMETER TRAINING



- Held **other hyperparameters constant** while **varying one** (n_estimator, max_depth, min_samples_leaf, & min_samples_split)
- Evaluated for **highest accuracy**
- **More time efficient** (~1 min) than **GridSearch** (2 min to days)

MODEL 2: GRIDSEARCH HYPERPARAMETER TRAINING

```
# Add tuned random forest model using grid search

# List Hyperparameters that we want to tune.
n_estimators = range(10, 251, 10)
max_features = ['sqrt'] # Not explored
max_depth = list(range(1, 21))
min_samples_leaf = [2] # Not explored
min_samples_split = [2] # Not explored

# Convert to dictionary
hyperparameters = dict()
hyperparameters['n_estimators'] = n_estimators
hyperparameters['max_features'] = max_features
hyperparameters['max_depth'] = max_depth
hyperparameters['min_samples_leaf'] = min_samples_leaf
hyperparameters['min_samples_split'] = min_samples_split
hyperparameters['criterion'] = ["gini"]

# Create new KNN object
rf_opt = RandomForestClassifier(random_state=0)

# Use grid search to find the ideal hyperparamters
grid_search_rf = GridSearchCV(rf_opt, hyperparameters, cv=2, verbose=5)

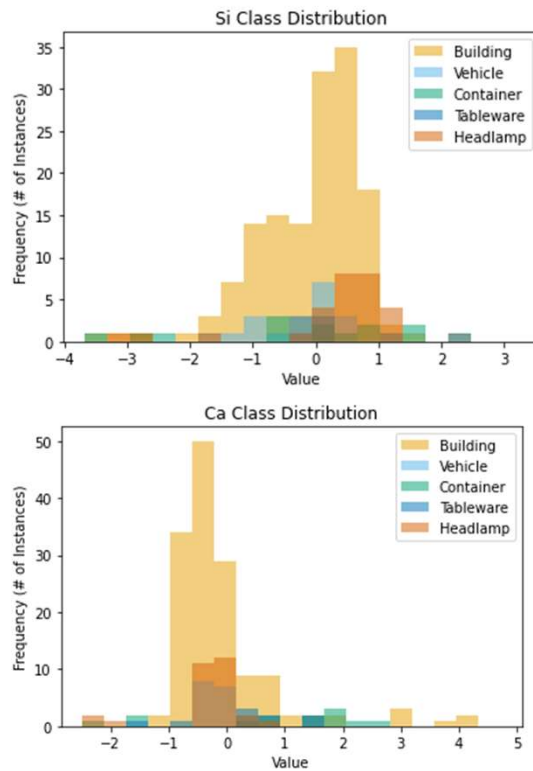
# Fit the model
grid_search_rf.fit(X_train, y_train)

# Add model to dictionary
models['Random Forest (Grid Search Tuned)'] = grid_search_rf
```

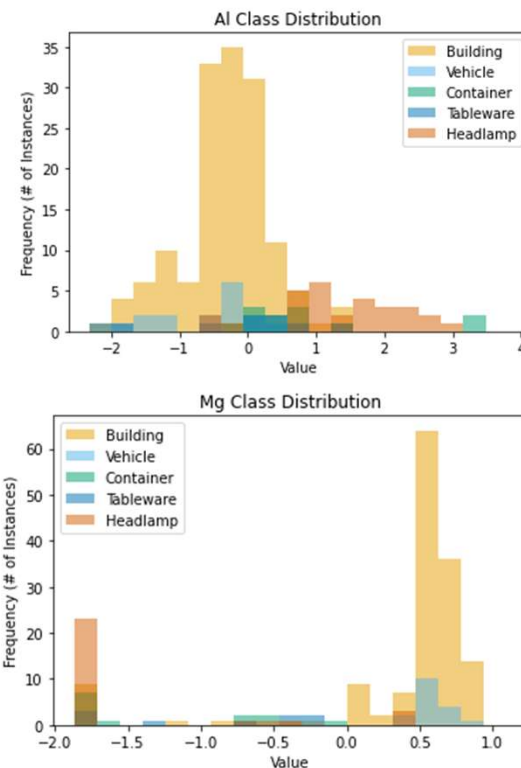
- Execute **every variation of model** (n_estimator & max_depth)
- Evaluated for **highest GINI index**
- Gets **highest possible accuracy** at the **cost of time**

MODEL 3: LIMITED INPUT COLUMNS TRAINING

Example Dropped



Example Retained



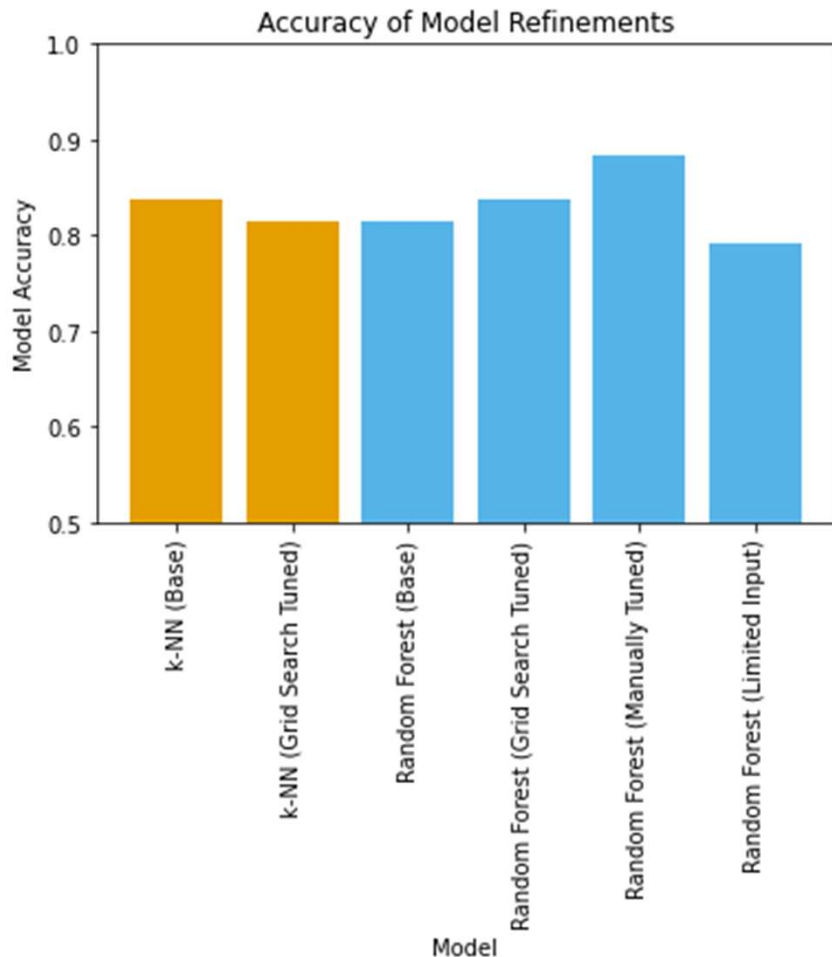
- **Dropped all variables** that do not have **sufficient distinction between classifications**
- Trained using **Model 1 hyperparameters**
- Potentially **remove variables** that contribute **unhelpful noise**

The background is a solid black field. It is populated with a network of thin, curved lines in two colors: a vibrant blue and a warm orange. These lines originate from various points and converge towards specific nodes. The nodes are represented by small white circles, some of which are simple outlines, while others are filled with a solid white color. The lines and nodes are distributed across the frame, with a higher density of lines and nodes on the right side, where they appear to form a more complex, web-like structure. The overall effect is one of dynamic movement and interconnectedness, reminiscent of a data visualization or a stylized celestial map.

RESULTS

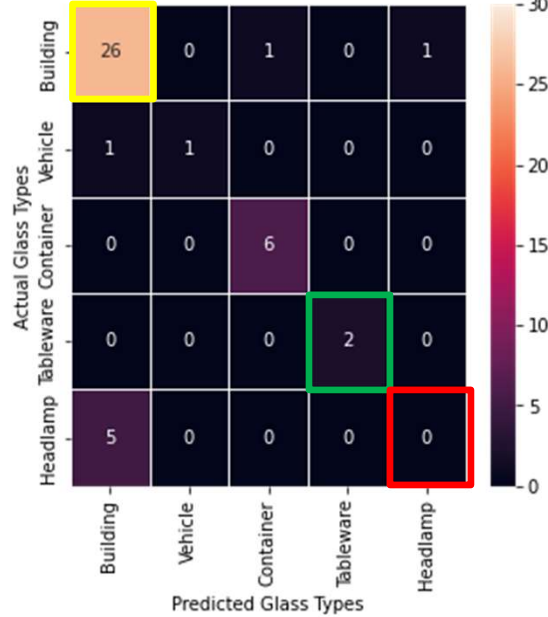
GridSearch and **manually tuned hyperparameters** of **Random Forest** models produced the **highest accuracy**

Removing data columns **reduced accuracy**



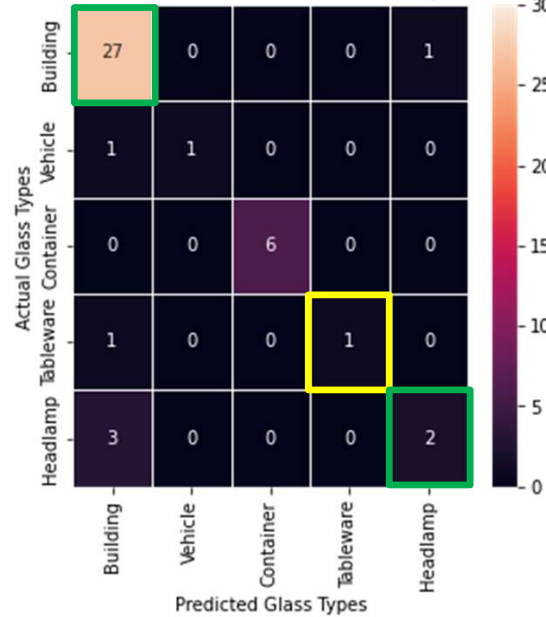
Base

Confusion Matrix for Random Forest (Base) Model



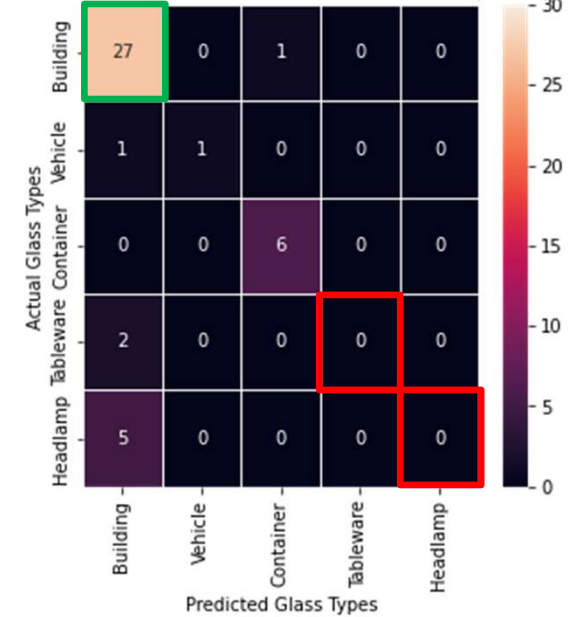
Manually Tuned

Confusion Matrix for Random Forest (Manually Tuned) Model



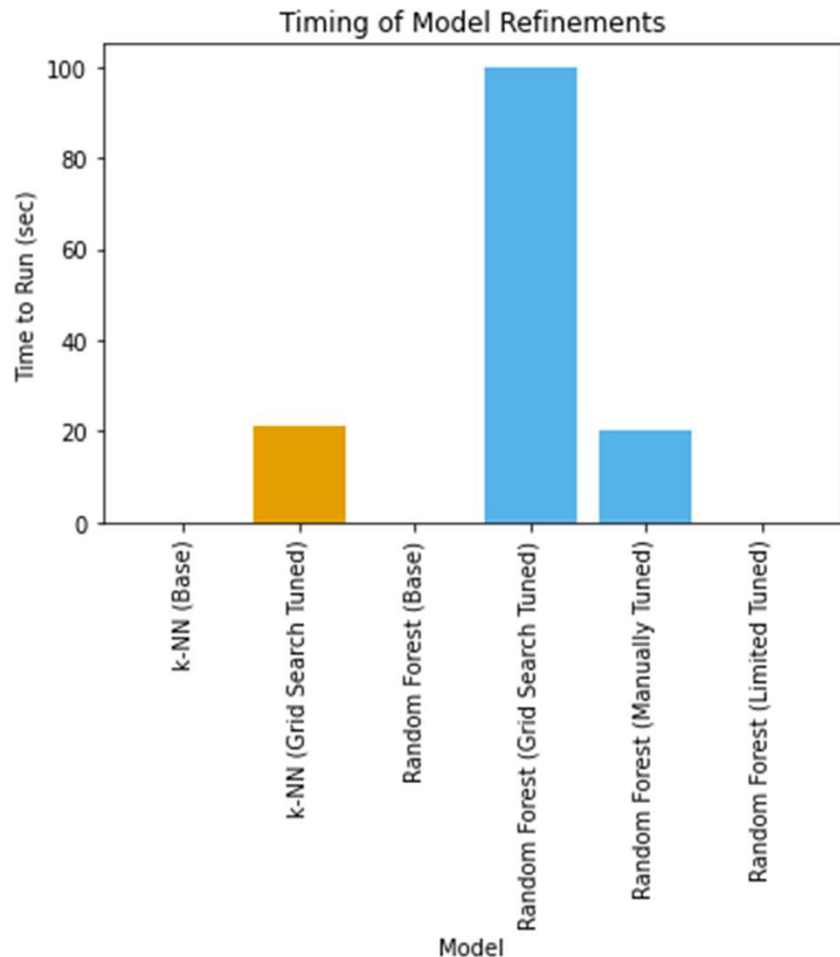
Limited

Confusion Matrix for Random Forest (Limited Input) Model



GridSearch often takes **too long** to be **practically useful**

Manual hyperparameter search dramatically **reduces computation time** with minimal accuracy loss



LIMITATIONS

- **Small** dataset
- **Older** data
 - Glass manufacturing evolves
- **Moderate variability**
 - Likely due to small dataset

An abstract network diagram on a black background. It features numerous nodes, represented by small white circles with black outlines, some of which are larger and have concentric circles. These nodes are interconnected by a dense web of thin lines. The lines are primarily blue and orange, with some white lines visible. The lines are mostly straight, but there are several prominent, thick, curved lines in blue and orange that sweep across the lower right portion of the image, suggesting a dynamic or evolving network structure. The overall composition is asymmetrical, with a higher density of nodes and lines in the upper left and lower right areas.

QUESTIONS?