Report

Background:

The database I choose is Breast Cancer Dataset. The dataset consists of 30 features, which are different factors of breast nuclei. The target is a little counter-conventional, Malignant as 0 and Benign as 1. The classifiers are Decision Tree, Random Forest, and Naïve Bayes.

1. Compare all three models in a table

Model	Target	Accuracy	Precision	Recall	F1-Score
Decision Tree	0	0.94	0.93	0.91	0.92
	1		0.94	0.96	0.95
Random Forest	0	0.96	0.98	0.93	0.95
	1		0.96	0.99	0.97
Naïve Bayes	0	0.97	1	0.93	0.96
	1		0.96	1	0.98

2. Identify the best model and discuss

The best model is **Naïve Bayes** based on F1-Score. Naive Bayes achieved the best balance between precision and recall, making it the top performer by this metric.

2.1 Impact of hyperparameters:

For Decision Tree:

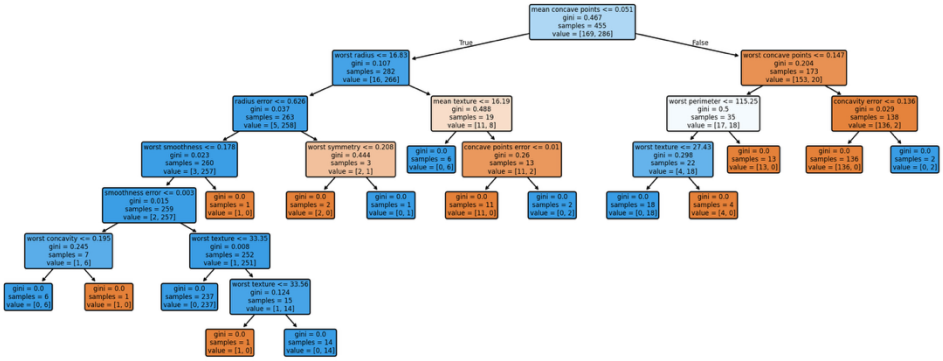
The default max_depth=None, means it will find the final leaf node until the end no matter how deep the tree is. The classification report will be like this:

```
Decision Tree:
              precision    recall  f1-score   support

           0       0.93      0.91      0.92        43
           1       0.94      0.96      0.95        71

    accuracy                           0.94       114
   macro avg       0.94      0.93      0.93       114
weighted avg       0.94      0.94      0.94       114
```

The final tree is:



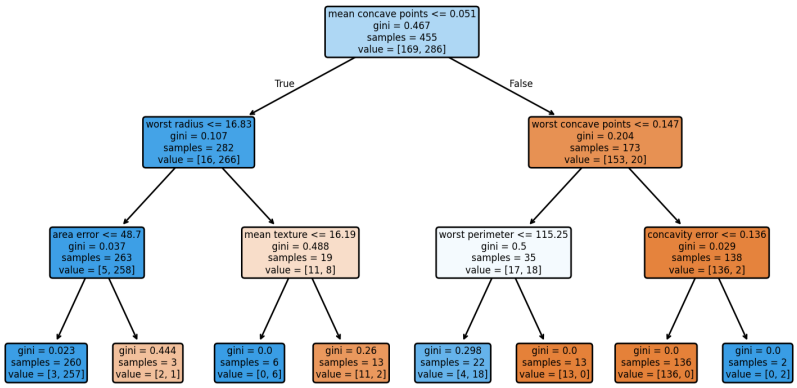
When we manipulate the max_depth=3, output is:

```
Decision Tree:
              precision    recall  f1-score   support

           0       0.95      0.91      0.93        43
           1       0.95      0.97      0.96        71

    accuracy                           0.95       114
   macro avg       0.95      0.94      0.94       114
weighted avg       0.95      0.95      0.95       114
```

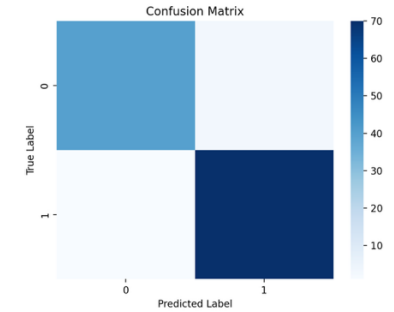
Figure 61



We can see the leaf nodes are fewer, and training for large dataset will be shorter.

For Random Forest:

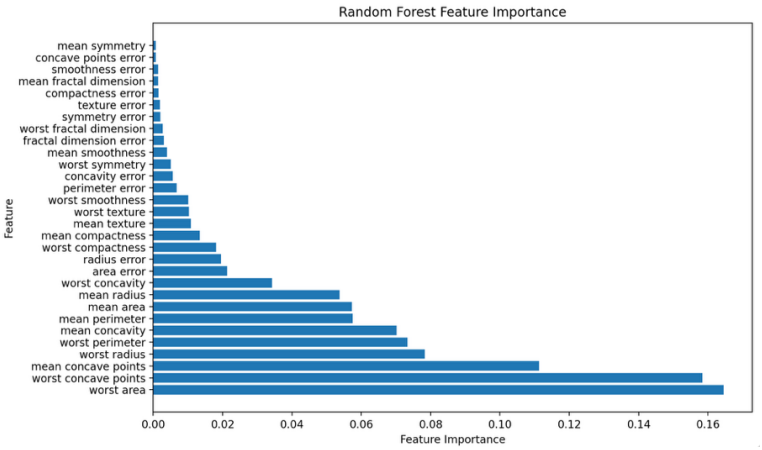
When n_estimators=100, max_depth=3



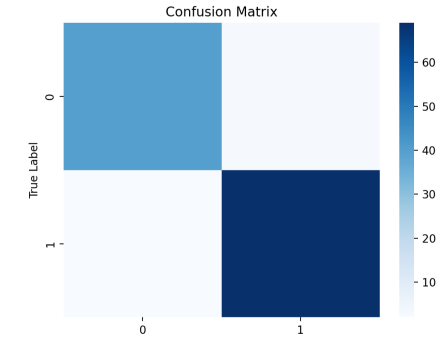

```
Random Forest:
              precision    recall  f1-score   support

           0       0.98      0.93      0.95        43
           1       0.96      0.99      0.97        71

    accuracy                           0.96       114
   macro avg       0.97      0.96      0.96       114
weighted avg       0.97      0.96      0.96       114
```



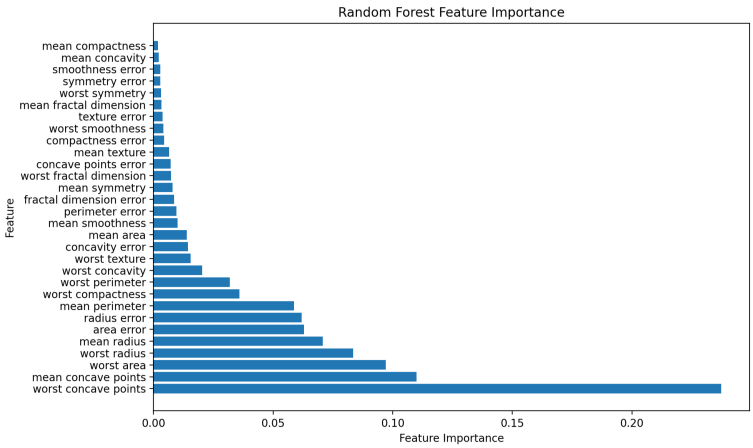
When n_estimators=10 (max_depth=None)




```
Random Forest:
              precision    recall  f1-score   support

           0       0.95      0.93      0.94        43
           1       0.96      0.97      0.97        71

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114
```



From the comparison we can see that classification report only change slightly even we switch n_estimators from 100 to 10. But the feature importance in random forest changed significantly. Due to the randomness nature of this classifier, bootstrapping sample 10 time is much different with 100 time.

For Naïve Bayes Classifier:

Unlike Decision Tree and Random Forest, there are not too many parameters that can fine-tune with.

```
Naive Bayes:
              precision    recall  f1-score   support

           0       1.00      0.93      0.96        43
           1       0.96      1.00      0.98        71

    accuracy                           0.97       114
   macro avg       0.98      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114
```

2.2 Bias-variance trade-off observations:

Decision Tree:

Training accuracy is higher than test accuracy. This means the model memorizes the training data too well and doesn't predict new data as accurately. Different training sets produce very different trees—the model is unstable.

Random Forest:

Training and test accuracy are close. The model learns real patterns without memorizing noise. Different training sets produce similar predictions—the model is stable and reliable."

2.3 Why performance differs among models:

Decision Tree (94%): Makes hard splits, greedy algorithm finds locally optimal solutions, sensitive to small data changes

Random Forest (96%): Combines many trees, errors cancel out, more robust to noise

Naive Bayes (97%): Simple probabilistic model, fits this dataset well because features are roughly independent and normally distributed

2.4 Which model performed worst and why?

Decision Tree (94%). Single trees are unstable—small data changes create different structures. High variance leads to overfitting. Can only make rectangular decision boundaries.

2.5 Why does random forest reduce variance?

Bootstrap sampling: Each tree sees different data

Random features: Each split uses random feature subset

Averaging: 100 trees vote, individual errors cancel out

Math: Averaging independent predictions reduces variance by ~1/n

Diversity: Trees make different mistakes, ensemble corrects them

2.6 Why is Naïve Bayes “naive”?

Assumes all features are independent given the class: P(features|class) = P(f_1|class) × P(f_2|class) × ...

In reality, features correlate (e.g., tumor radius and area are related). The "naive" assumption ignores these correlations, making calculations simple but theoretically

wrong. It works anyway because classification only needs correct ranking, not accurate probabilities.