



Faculty of Engineering and Technology

Computer Systems Engineering Department

MACHINE LEARNING AND DATA SCIENCE

ENCS5341

Assignment #3

Prepared By:

Hajar Salah 1191482

Sary Hammad 1192698

Instructor: Yazan Abu Farha

Section: (2)

BIRZEIT

January– 2024

Table of Contents

3.1. Introduction.....	1
3.2. Dataset.....	2
3.2.1. Overview.....	2
3.2.2. Quantitative Measures.....	3
3.2.3. Qualitative Measures.....	5
3.3. Experiments and Results.....	7
3.3.1. K-Nearest-Neighbour (KNN).....	7
3.3.2. XGBoost Classifier.....	9
3.3.3. Random Forest Classifier.....	11
3.4. Analysis.....	13
3.5. Conclusions and Discussion.....	14

3.1. Introduction

In this project, the realm of breast cancer diagnosis was delved into, utilizing a dataset that encapsulated crucial information. With 30 distinct measurements pertaining to the dimensions of tumors observed in 569 patients, along with their corresponding medical classifications, the dataset presented a powerful opportunity for a Binary Classification task. The focus was to recognize whether a tumor is cancerous (malignant) or non-cancerous (benign).

To accomplish this task, three distinct machine learning models—K-Nearest Neighbors (KNN), XGBoost, and Random Forest—were employed. These models were tasked with analyzing the complex patterns within the dataset to make accurate predictions regarding the nature of the tumors.

The evaluation process encompassed a comprehensive set of metrics to measure the performance of each model. Precision, recall, accuracy, F1 Score, Area Under the Curve (AUC), and the Receiver Operating Characteristic (ROC) curve were employed to provide a thorough assessment of the models' capabilities. Through this exploration, the project aimed to not only achieve accurate predictions but also gain insights into the detailed performance of each model across various evaluation criteria.

3.2. Dataset

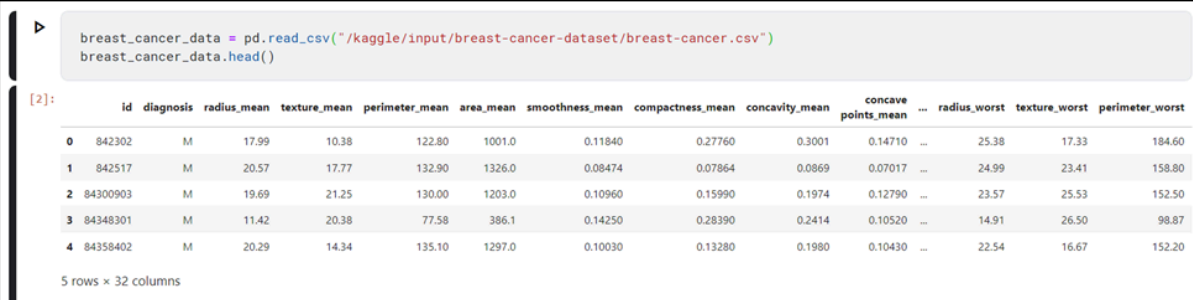
3.2.1. Overview

This project's chosen dataset is the following Breast Cancer Dataset acquired from Kaggle:

<https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset/data>

Breast cancer is the most common cancer amongst women globally, and accounts for 25% of all cancer cases. The data card mentions that over 2.1 million people were affected by this type of cancer in 2015 alone. This type of cancer starts with breast cells beginning to grow out of control, creating tumors that can be seen through X-rays or felt as lumps in the breast. The features in the dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass, and describe characteristics of the cell nuclei present in the image.

Thus, through providing 30 different measurements regarding the dimensions of the tumors observed in 569 patients, and their medical classification, this dataset resembles the opportunity for a Binary Classification task for the type of tumor, whether it is cancerous (malignant) or non-cancerous (benign). In such a problem, it is crucial to accurately identify malignant cases as the cost of missing a malignant case (false negative) can be much higher than incorrectly identifying a case as malignant (false positive). Figure 2-1 shows the first 5 rows of data after reading it using pandas.



```
breast_cancer_data = pd.read_csv("/kaggle/input/breast-cancer-dataset/breast-cancer.csv")
breast_cancer_data.head()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	radius_worst	texture_worst	perimeter_worst
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	25.38	17.33	184.60
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	24.99	23.41	158.80
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	23.57	25.53	152.50
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	14.91	26.50	98.87
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	22.54	16.67	152.20

5 rows x 32 columns

Figure 2-1: first 5 rows of the dataset

Since the table does not fit all columns, figure (2-2) shows the info summary regarding the columns in the dataset.

```
breast_cancer_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 31 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   diagnosis                             569 non-null    object
1   radius_mean                           569 non-null    float64
2   texture_mean                           569 non-null    float64
3   perimeter_mean                         569 non-null    float64
4   area_mean                             569 non-null    float64
5   smoothness_mean                       569 non-null    float64
6   compactness_mean                      569 non-null    float64
7   concavity_mean                        569 non-null    float64
8   concave points_mean                   569 non-null    float64
9   symmetry_mean                         569 non-null    float64
10  fractal_dimension_mean                 569 non-null    float64
11  radius_se                             569 non-null    float64
12  texture_se                             569 non-null    float64
13  perimeter_se                           569 non-null    float64
14  area_se                               569 non-null    float64
15  smoothness_se                         569 non-null    float64
16  compactness_se                        569 non-null    float64
17  concavity_se                          569 non-null    float64
18  concave points_se                     569 non-null    float64
19  symmetry_se                           569 non-null    float64
20  fractal_dimension_se                   569 non-null    float64
21  radius_worst                          569 non-null    float64
22  texture_worst                         569 non-null    float64
23  perimeter_worst                       569 non-null    float64
24  area_worst                            569 non-null    float64
25  smoothness_worst                      569 non-null    float64
26  compactness_worst                     569 non-null    float64
27  concavity_worst                       569 non-null    float64
28  concave points_worst                   569 non-null    float64
29  symmetry_worst                        569 non-null    float64
30  fractal_dimension_worst                 569 non-null    float64
```

Figure 2-2: Dataset features

3.2.2. Quantitative Measures

The dataset is inspected for any missing values and figure 2-3 shows that there are none.

```
breast_cancer_data.isna().sum()
```

```
id                0
diagnosis         0
radius_mean       0
texture_mean      0
perimeter_mean    0
area_mean         0
smoothness_mean   0
compactness_mean  0
concavity_mean    0
concave points_mean 0
symmetry_mean     0
fractal_dimension_mean 0
radius_se         0
texture_se        0
perimeter_se      0
area_se           0
smoothness_se     0
compactness_se    0
concavity_se      0
concave points_se 0
symmetry_se       0
fractal_dimension_se 0
radius_worst      0
texture_worst     0
perimeter_worst   0
area_worst        0
smoothness_worst  0
compactness_worst 0
concavity_worst   0
concave points_worst 0
symmetry_worst    0
fractal_dimension_worst 0
```

Figure 2-3: Missing dataset features

Then the balance of the classes was inspected and is shown in figure 2-4, which shows that there are 212 malignant cases and 357 benign cases. Meaning this dataset is not severely imbalanced like other datasets out there, since 37% of the cases are malignant.

```
breast_cancer_data['diagnosis'].value_counts()

diagnosis
B      357
M      212
```

Figure 2-4: classes distribution

Three randomly selected features were analyzed for some descriptive statistics as shown in figure 2-5.

- **Radius Mean:** The skewness is close to 1, indicating a moderately right-skewed distribution. A **kurtosis** less than 3 suggests it is less peaked than a normal distribution.
- **Concavity Mean:** This feature is also right-skewed, more so than radius mean. The kurtosis greater than 3 indicates a more peaked distribution than normal.
- **Area Worst:** It exhibits a high degree of right skewness and has a kurtosis much greater than 3, indicating a very peaked distribution with heavy tails.

```
# Calculating descriptive statistics for selected columns
selected_columns = ['radius_mean', 'concavity_mean', 'area_worst']
selected_stats = breast_cancer_data[selected_columns].describe()

# Calculating additional metrics like skewness and kurtosis
skewness = breast_cancer_data[selected_columns].skew()
kurtosis = breast_cancer_data[selected_columns].kurtosis()

print(selected_stats)
print("Skewness:\n", skewness)
print("Kurtosis:\n", kurtosis)
```

	radius_mean	concavity_mean	area_worst
count	569.000000	569.000000	569.000000
mean	14.127292	0.088799	880.583128
std	3.524049	0.079720	569.356993
min	6.981000	0.000000	185.200000
25%	11.700000	0.029560	515.300000
50%	13.370000	0.061540	686.500000
75%	15.780000	0.130700	1084.000000
max	28.110000	0.426800	4254.000000

```
Skewness:
radius_mean      0.942380
concavity_mean    1.401180
area_worst        1.859373
dtype: float64
Kurtosis:
radius_mean      0.845522
concavity_mean    1.998638
area_worst        4.396395
dtype: float64
```

Figure 2-5: Descriptive statistics on random columns

These observations suggest that the distributions of these features are asymmetric, with a tendency to have more extreme values on the higher end (right tail). Which is expected in such dataset. For example, in the case of "radius_mean" and "concavity_mean", a right-skewed distribution suggests that most tumors have smaller radii and concavities, but

there are a number of cases with much larger values. This is typical in medical data where a majority of cases might be less severe, with fewer extreme but more severe cases.

In addition, the high kurtosis especially in "area_worst" indicates that the data has heavy tails or outliers. This is consistent with medical data where most patients might have similar characteristics, but a few cases can be significantly different and indicate a more aggressive form of tumor.

After this initial look at the dataset, the decision to use a label encoder to encode the target class in addition to removing the id column since it is not a feature was made.

3.2.3. Qualitative Measures

Starting off with a heat-correlation matrix for the variables in our dataset, shown in figure 2-6. It is noticed that the diagnosis has a relatively high correlation with many of the physical features in the dataset, like the radius, perimeter, area and concavity. Also, it is noticed that many of these features are actually related for example (radius, perimeter, and area).

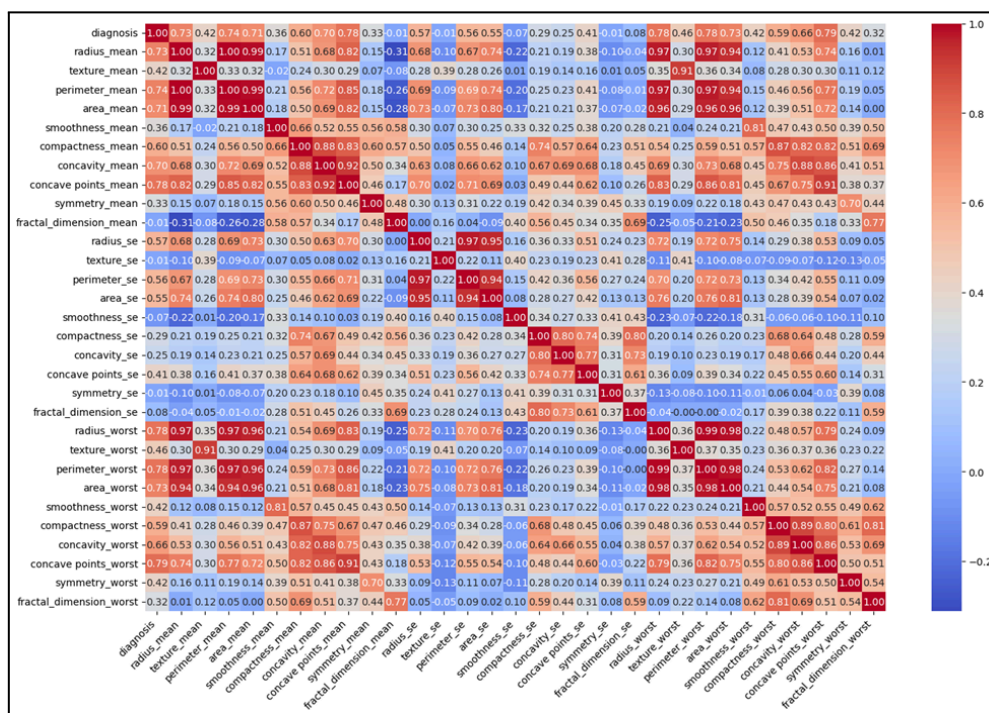


Figure 2-6: Heat-correlation matrix

Figure 2-7 shows the histograms of the 3 randomly selected features. The results shown confirm the right skewness as previously indicated by the skewness and kurtosis coefficients.

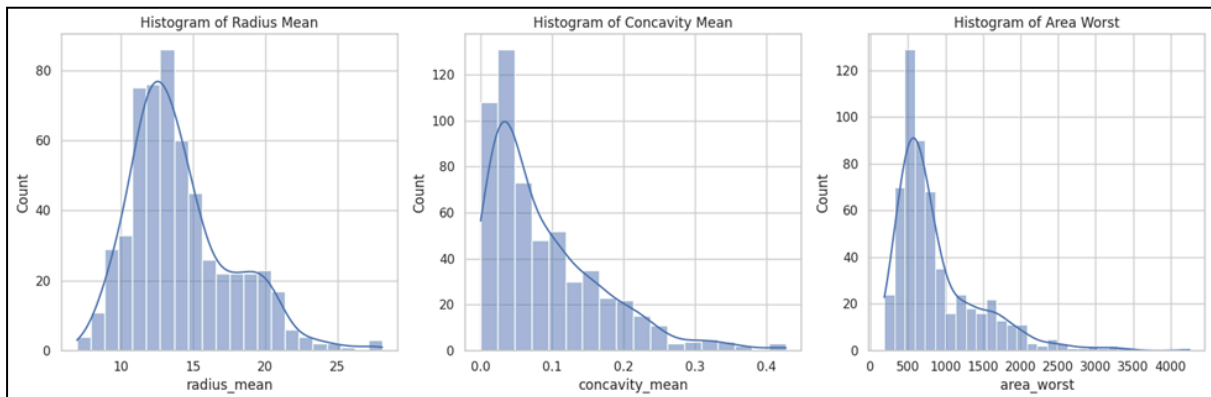


Figure 2-7: Histograms of the three features

3.3. Experiments and Results

3.3.1. K-Nearest-Neighbour (KNN)

Starting with a baseline KNN model, figure 3-1 shows the classification report for this model.

Accuracy for k = 1: 0.93				
Classification Report for k = 1:				
	precision	recall	f1-score	support
0	0.92	0.97	0.95	71
1	0.95	0.86	0.90	43
accuracy			0.93	114
macro avg	0.93	0.92	0.92	114
weighted avg	0.93	0.93	0.93	114
Accuracy for k = 3: 0.93				
Classification Report for k = 3:				
	precision	recall	f1-score	support
0	0.93	0.96	0.94	71
1	0.93	0.88	0.90	43
accuracy			0.93	114
macro avg	0.93	0.92	0.92	114
weighted avg	0.93	0.93	0.93	114

Figure 3-1: Classification Report of KNN

1. For K = 1:

- **Recall:** The recall is 0.86, which indicates that the model correctly diagnoses only 86% of the people as malignant. However, it assumes 14% to be benign when they are actually malignant. This is a dangerous measure and can't be relied on for accurate diagnoses.
- **Precision:** Precision is 0.95, meaning the system correctly classified malignant cases, but mistakenly classified 5% of the population as malignant when they were benign.

2. For K = 3:

- **Recall:** It increases by only 2%, resulting in a recall of 88%, which is still considered a very low value.
- **Precision:** Precision decreases to 0.93, indicating an increased risk of classifying a non-patient as a patient.

3. Regarding the accuracy:

The accuracy is 93%, meaning the system correctly classifies 93% of the cases. In the case of breast cancer recognition, a 93% accuracy is considered low and needs improvement.

4. ROC Curve:

The Receiver Operating Characteristic (ROC) curve for this model was also plotted, as shown in figure 3-2.

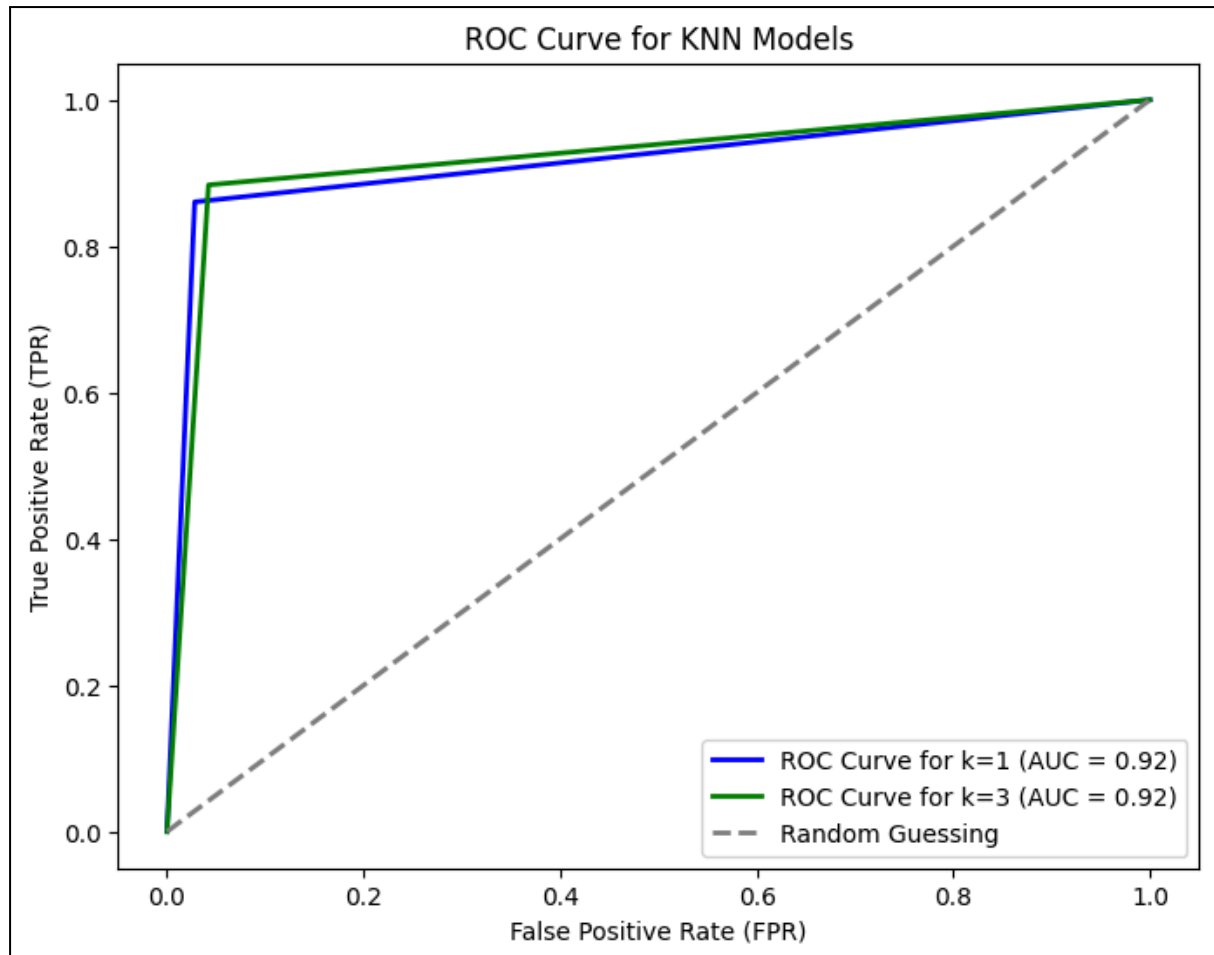


Figure 3-2: ROC curve of KNN

The AUC is a single value that represents the overall performance of the model. The high AUC (0.92) indicates the model's ability to better discriminate between positive and negative classes.

In summary, the model has limitations in correctly identifying malignant cases (low recall) and may classify some non-patients as patients (decreasing precision as K increases). The overall accuracy is also considered low for breast cancer recognition, highlighting the need for improvement in the model, or picking a more appropriate model.

3.3.2. XGBoost Classifier

Starting with a baseline XGBoost classifier, figure 3-3 shows the classification report for this model.

0.956140350877193				
	precision	recall	f1-score	support
B	0.96	0.97	0.97	71
M	0.95	0.93	0.94	43
accuracy			0.96	114
macro avg	0.96	0.95	0.95	114
weighted avg	0.96	0.96	0.96	114

Figure 3-3: Classification Report of baseline XGBoost

All features with less than 0.4 correlation were attempted to be removed, resulting in a somewhat improved result as shown in figures 3-4.

Accuracy of the model: 0.9649122807017544				
	precision	recall	f1-score	support
B	0.96	0.99	0.97	71
M	0.98	0.93	0.95	43
accuracy			0.96	114
macro avg	0.97	0.96	0.96	114
weighted avg	0.97	0.96	0.96	114

Figure 3-4: Classification Report of XGBoost

The results show that removing features with less correlation to the target variable has had a small yet positive effect on the model's performance. The improvement in precision for the malignant class is particularly important in such a clinical dataset, as it implies fewer false positives for the most critical class. The slight increase in recall for the benign class also suggests fewer false negatives, which is equally important.

Now PCA (components=2) is applied to XGBoost and the results are shown in figure 3-5.

0.9824561403508771				
	precision	recall	f1-score	support
B	0.99	0.99	0.99	71
M	0.98	0.98	0.98	43
accuracy			0.98	114
macro avg	0.98	0.98	0.98	114
weighted avg	0.98	0.98	0.98	114

Figure 3-5: Classification Report of XGBoost with PCA

The following are the most notable improvements:

- Accuracy: There is a notable improvement in the overall accuracy of the model after applying PCA, from 95.61% to 98.25%.
- Class B (Benign): Both precision and recall have improved from 0.96 and 0.97 respectively to 0.99, meaning that the model's ability to identify benign cases correctly has improved.
- Class M (Malignant): There is an improvement in both precision and recall, from 0.95 and 0.93 to 0.98. This indicates a significant increase in the model's ability to correctly identify malignant cases.

Hyperparameter tuning was given a shot on this model with PCA, however the results remained the same, indicating that the default values for the XGB were good enough, this is shown in figure 3-6.

```
Best Hyperparameters: {'colsample_bytree': 1.0, 'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 200, 'subsample': 1.0}
Accuracy of the model: 0.9824561403508771
```

Figure 3-6: Hyperparameter tuning for XGBoost with PCA

The Receiver Operating Characteristic (ROC) curve for this model was plotted:

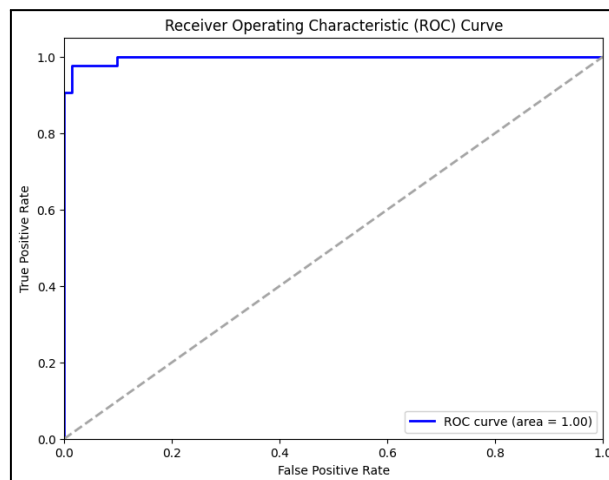


Figure 3-7: ROC cure of the XGBoost with PCA

3.3.3. Random Forest Classifier

Starting with a baseline RF classifier, figure 3-8 shows the results obtained:

Accuracy: 0.9649122807017544					
	precision	recall	f1-score	support	
B	0.96	0.99	0.97	71	
M	0.98	0.93	0.95	43	
accuracy			0.96	114	
macro avg	0.97	0.96	0.96	114	
weighted avg	0.97	0.96	0.96	114	

Figure 3-8: Classification Report of RF

The Random Forest classifier demonstrates excellent performance metrics on this dataset. The high recall for the benign class is particularly encouraging since it suggests that the classifier is very capable of identifying benign instances, which is critical in a medical diagnosis context to avoid false negatives. Similarly, the precision for the malignant class is also excellent, which is important to avoid false alarms and unnecessary treatments.

Next, an attempt was made to reduce the dimensionality of the data by removing some features based on their correlation coefficient, and a threshold was used to manipulate this. It was found that the optimal threshold value at 0.4 yielded the exact same results as the baseline RF. Therefore, another approach for dimension reduction was used, which is principal component analysis (PCA). After using PCA and finding the optimal number of components to be 2, figure 3-9 shows the results of this model:

Accuracy: 0.9824561403508771					
	precision	recall	f1-score	support	
B	0.97	1.00	0.99	71	
M	1.00	0.95	0.98	43	
accuracy			0.98	114	
macro avg	0.99	0.98	0.98	114	
weighted avg	0.98	0.98	0.98	114	

Figure 3-8: Classification Report of RF with PCA

It was noticed that there is an improvement in the overall accuracy of the model after applying PCA from 96.49% to 98.25%. In addition, the precision for the benign class (B) remains roughly the same, but the precision for the malignant class (M) has improved to 1.00, indicating that every instance predicted as malignant is actually malignant. Also, the recall for the benign class has improved to 1.00, meaning all benign instances were correctly identified. The recall for the malignant class has slightly improved as well. Thus, causing an overall improvement in F1 score, reflecting an overall better performance across both classes after PCA probably due to the removal of noisy features.

Hyperparameter tuning on was given a shot on this model with PCA, however the results remained the same, indicating that the default values for the RF were good enough, this is shown in figure 3-9.

Fitting 3 folds for each of 144 candidates, totalling 432 fits				
Best Parameters: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}				
Accuracy: 0.9824561403508771				
	precision	recall	f1-score	support
B	0.97	1.00	0.99	71
M	1.00	0.95	0.98	43
accuracy			0.98	114
macro avg	0.99	0.98	0.98	114
weighted avg	0.98	0.98	0.98	114

Figure 3-9: Classification Report of RF with PCA and hyperparameter tuning

The Receiver Operating Characteristic (ROC) curve for this model:

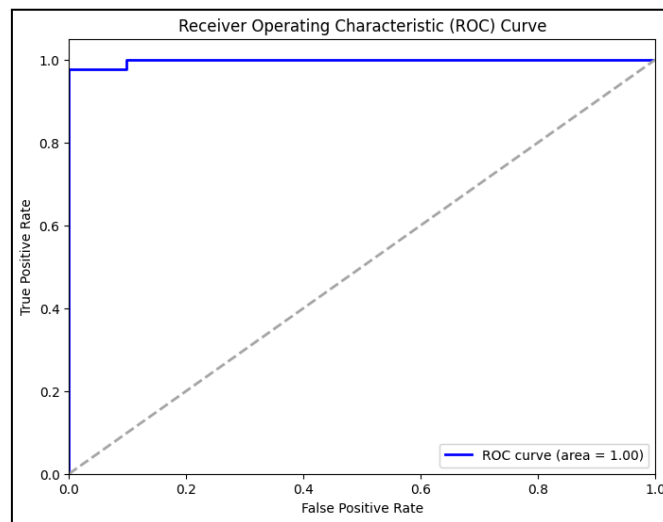


Figure 3-10: ROC of RF with PCA and hyperparameter tuning

3.4. Analysis

In regards to the 3 models employed on the dataset, it appears that the KNN model has limitations in correctly identifying malignant cases (low recall 0.86) and may classify some non-patients as patients (decreasing precision as K increases). The overall accuracy (0.93) is also considered low for breast cancer recognition. Thus, these results pushed to use the other models; the top 2 models, which also pretty much have the same accuracy are the XGBoost and the Random Forest classifiers. Given the nature of our selected dataset, the incredible results are anticipated and are confidently realistic figures that stand as proof of no overfitting on the data. In this dataset, the cost of a false negative is very high and dangerous, therefore, recall is our most critical metric here. XGBoost has a higher recall for malignant class (0.98) compared to the Random Forest classifier (0.95). Meaning that the XGBoost model is less likely to miss a case of malignancy, making it favorable. Even though it was noticed that the precision for the RF is perfect (1.00), in the medical context of this task, the slightly lower precision of XGBoost (0.98) can be acceptable if it means capturing more true positive cases.

However, looking at the Benign class, and a case is incorrectly predicted to be benign when it is actually malignant (false negative), then that is a problem. Thus, comparing the recall for this class for the two models is important. The RF had a perfect recall score (1.00) for the Benign class meaning that it correctly identified all benign cases, and also that it did not incorrectly classify any malignant tumor as benign, which is what is aimed for. On the other hand, XGBoost had a slightly lower recall (0.99), meaning that it did have a small rate of false negatives on the benign class. This presents us with a trade-off that will be discussed in the conclusion.

3.5. Conclusions and Discussion

In conclusion, after evaluating 3 models on the selected dataset, it turns out that the KNN was not a good model for this dataset, as it wasn't accurate enough to provide true classifications, thus an alternative model was needed. As a result, diving deeper into analyzing the RF and XGBoost, with exact same F1 scores, this presented a trade off. The XGBoost model has a better recall for the critical Malignant class, making it less likely to miss actual cases of cancer and therefore is more sensitive to detecting cancer. On the other hand, the RF model had a better recall for the Benign class, ensuring no cancer cases are misclassified as benign meaning it was more specific in confirming non cancer cases.

The choice between the two models depends on our priority, if it was to ensure that all malignant cases are caught, even at the risk of more benign cases being misclassified as malignant (leading to additional tests), then the XGBoost is preferred. However, if the priority was to minimize false negatives, as they could lead to further invasive testing, and potentially harmful treatment that is not needed in cancer diagnosis, then RF is the better choice.

It is true that both models provided very high impressive results on the dataset, which was predictable, since doctors also rely on FNAs of a breast mass for their diagnosis. This does not mean that the models can replace a professional Doctor's diagnosis, or at least not yet, but both can be employed by such Doctors in order to give them insights and perhaps for double checking their diagnosis with the diagnosis of both models. In order to make this work, a much larger dataset needs to be available so that many more cases are fed into the ML models. Humans are diving quickly into a world of ML and Big Data, where machines and humans collaborate towards a better future, and hopefully not weaponizing AI to hurt each other.