Breast Cancer Prediction

A Data-Driven Approach for Improved Breast Cancer Predictions

Milestone 5 – Final Project Paper

Table of Contents

[Introduction 2](#_Toc142748209)

[Data source & Preparation 2](#_Toc142748210)

[Models 7](#_Toc142748211)

[Random Forest Classifier 7](#_Toc142748212)

[Interpreting the Results 7](#_Toc142748213)

[Decision Tree Classifier 8](#_Toc142748214)

[Interpreting the Results 9](#_Toc142748215)

[KNN - KNeighbors Classifier 10](#_Toc142748216)

[Interpreting the Results 11](#_Toc142748217)

[Comparing Models Performance 11](#_Toc142748218)

[Conclusion 12](#_Toc142748219)

[Recommendations 13](#_Toc142748220)

[Limitations 13](#_Toc142748221)

[Risk 13](#_Toc142748222)

[References 14](#_Toc142748223)

# Introduction

Breast cancer, a complex and formidable adversary, demands meticulous scrutiny and proactive interventions to minimize its impact. It stands as a significant health concern, affecting not only women but also a portion of men. In an age propelled by technological advancements, the realm of healthcare is undergoing a transformative shift with the integration of machine learning. This project embarks on a journey to leverage the power of predictive modeling to enhance the early detection of breast cancer, a pressing concern that deeply impacts the lives of countless individuals. Through rigorous analysis and comparison of predictive models, we aspire to uncover the most effective means of predicting this challenging adversary.

# Data source & Preparation

I obtained the breast cancer dataset from Kaggle([Pmotta](#_References) - 2021, June 6). The breast cancer dataset that has been chosen contains essential features that capture the fundamental health factors of breast cancer patients. These features have been determined to be sufficient for the purpose of prediction.

The selected dataset was relatively small and only a modest amount of cleaning and preparation steps was required. The attributes of the dataset are summarized in the table below.

|  |  |
| --- | --- |
| Clump Thickness: | (1-10). Benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayers. |
| Uniformity of Cell Size: | (1-10). Cancer cells tend to vary in size and shape. That is why these parameters are valuable in determining whether the cells are cancerous or not. |
| Uniformity of Cell Shape: | (1-10). Uniformity of cell size/shape: Cancer cells tend to vary in size and shape. That is why these parameters are valuable in determining whether the cells are cancerous or not. |
| Marginal Adhesion: | (1-10). Normal cells tend to stick together. Cancer cells tend to lose this ability. So, the loss of adhesion is a sign of malignancy. |
| Single Epithelial Cell Size: | (1-10). It is related to the uniformity mentioned above. Epithelial cells that are significantly enlarged may be a malignant cell. |
| Bare Nuclei: | (1-10). This is a term used for nuclei not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors. |
| Bland Chromatin: | (1-10). Describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells, the chromatin tends to be coarser. |
| Normal Nucleoli: | (1-10). Nucleoli are small structures seen in the nucleus. In normal cells, the nucleolus is usually very small if visible at all. In cancer cells, the nucleoli become more prominent, and sometimes there are more of them. |
| Mitoses: | (1-10). Cancer is essentially a disease of uncontrolled mitosis. |
| Class: | 1. or 4) (2) Benign (non-cancerous) or (4) malignant (cancerous) lump in a breast. |

The executed steps for data preparation are as follows:

1. Renaming columns to enhance accessibility and clarity.
2. Checking for null rows and columns. This step aimed to remove columns or rows with missing data. Notably, the dataset did not contain any empty rows or columns.
3. Detecting duplicates to eliminate redundant information. No duplicate entries were identified.
4. Identify data classification by “Class” columns. The dataset is divided into benign and malignant classes.

Visualizations

Visualizations were generated to uncover relationships among the variables to help get a better understanding of the available features.

HEAT MAPA screenshot of a computer

Description automatically generated

PAIRPLOT

A screenshot of a graph

Description automatically generated

PIE CHART

A pie chart was created to identify potential data imbalances within the 'Class' attribute. The analysis revealed an imbalance in the data, with 75.3% categorized as malignant and 24.7% as benign.

A blue and red pie chart

Description automatically generated

To construct an effective model, it was essential to address the dataset's imbalance. To achieve this, the Synthetic Minority Oversampling Technique (SMOTE) was employed. SMOTE generates synthetic instances for the minority class, thus balancing the dataset. With the dataset now balanced, the subsequent step was to partition the data into training and testing datasets.

The following models were developed, with their respective outcomes recorded.

1. Random Forest Classifier
2. Decision Tree Classifier
3. KNN Classifier

Following the initial plan, I assessed the model's performance using metrics like accuracy, precision, recall, F1-score and the confusion matrix. These insights guided decisions to enhance its predictive abilities.

# Models

## Random Forest Classifier

The outcome of the Random Forest Classifier are as follows:

A screenshot of a computer

Description automatically generated

### Interpreting the Results

For predicting cases of benign tumors (class 2), the model has a high precision (0.99), meaning that when it predicts a tumor as benign, it is correct 99% of the time. The recall (sensitivity) is 0.95, indicating that the model captures 95% of the actual benign cases. The F1-score (a balance between precision and recall) is 0.97.

While the model's precision in predicting malignant cases (class 4) is slightly lower (0.94) than that for benign cases (class 2), its recall (0.99) indicates significant effectiveness in capturing actual malignant cases.

The F1-score for class 4 is also 0.97. The confusion matrix indicates that there were 5 false positive predictions and 1 false negative prediction. The accuracy of 0.97 indicates that the model correctly predicts around 97% of the instances in the dataset.

Overall, the model appears to perform well with high precision and recall for both classes, as well as a strong F1-score. This suggests that the Random Forest Classifier is effective for predicting breast cancer, distinguishing between benign and malignant tumors based on the provided metrics.

## Decision Tree Classifier

Initially, the model was configured with a max\_depth of 5, which did not yield the optimal accuracy. However, by conducting cross-validation across various max\_depth values, it became evident that the best accuracy was achieved when using a max\_depth of 3. This iterative process of exploring different max\_depth values through cross-validation led to the identification of an improved accuracy level.

A graph with blue and orange dots

Description automatically generated

The above plot shows that the best accuracy for the model is when the parameter max\_depth is 3.

The results of the Decision Tree Classifier model, optimized with the specific parameters, are as follows:

A screenshot of a computer

Description automatically generated

### Interpreting the Results

The model exhibits commendable performance for both classes, displaying relatively high precision, recall, and F1-score values. It correctly identifies around 96% of benign cases (class 2) and 93% of malignant cases (class 4). The confusion matrix indicates that there were 6 false positive predictions and 4 false negative predictions. While the model is performing well overall, these misclassifications should be considered in the context of the application. The ROC-AUC score of 0.944 and accuracy score of 0.944 demonstrate the model's effectiveness in distinguishing between benign and malignant cases.

The Decision Tree map is as follows:

A diagram of a mathematical algorithm

Description automatically generated with medium confidence

From the decision tree above, we can see how the model classifies instances and the sequence of decisions it makes based on different features.

## KNN - KNeighbors Classifier

Initially, using the default parameters for the KNeighborsClassifier resulted in an approximate accuracy of 60%. However, by employing GridSearchCV to determine the optimal value for the n\_neighbors parameter, a substantial enhancement was achieved. Specifically, upon adopting the optimal n\_neighbors value of 5, the model's accuracy experienced a notable surge to 95.5%.

Following is the KNN model outcome:

A screenshot of a computer program

Description automatically generated

### Interpreting the Results

The KNN model performs well for both classes, with high precision, recall, and F1-score values. It correctly identifies around 96% of benign cases (class 2) and 95% of malignant cases (class 4). The confusion matrix indicates that there were 4 false positive predictions and 4 false negative predictions. The ROC-AUC score of 0.955 and accuracy score of 0.955 demonstrate the model's effectiveness in distinguishing between benign and malignant cases.

# Comparing Models Performance

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metric** | **Random Forest classifier** | | **KNN – KNeighbors classifier** | | **Decision Tree classifier** | |
|  | Benign | Malignant | Benign | Malignant | Benign | Malignant |
| Precision | 0.99 | 0.94 | 0.96 | 0.95 | 0.96 | 0.93 |
| Recall | 0.95 | 0.99 | 0.96 | 0.95 | 0.93 | 0.95 |
| F1-score | 0.97 | 0.97 | 0.96 | 0.95 | 0.95 | 0.94 |
| Roc-Auc score | 0.96 | | 95.5 | | 94.4 | |
| Accuracy score | 0.96 | | 95.5 | | 94.3 | |

This table will provide a clear overview of key metrics such as accuracy, precision, recall, and F1-scores for each model—Random Forest, Decision Tree, and KNN. This side-by-side comparison will shed light on how each model excels in different aspects of breast cancer prediction, enabling us to make informed conclusions about their effectiveness.

# Conclusion

The primary objective of this project was to analyze the breast cancer dataset and develop a predictive model for breast cancer. By constructing and comparing different models, we aimed to identify the most effective approach for this prediction task. Three models were implemented: RandomForest, DecisionTree, and KNN, in order to determine the optimal model for breast cancer prediction. The **Random Forest Classifier** exhibited the highest accuracy of 96.6%, showcasing its proficiency in distinguishing between benign and malignant cases. It demonstrated robust precision, recall, and F1-scores for both classes, underscoring its overall effectiveness. In contrast, the **Decision Tree** achieved an accuracy of 94%, displaying slightly lower performance compared to the Random Forest. The **KNN** model, after fine-tuning its parameters, achieved an accuracy of 95.5%.

The ROC-AUC score and accuracy score of the models demonstrate the model's effectiveness in distinguishing between benign and malignant cases. All three models perform well, with accuracy scores in the range of 94 to 97%, indicating strong predictive capabilities. The Random Forest and KNN models show slightly higher precision, recall, and F1-scores compared to the Decision Tree. The ROC-AUC scores for all models are above 94, indicating their capacity to discriminate between classes.

Overall, each model demonstrates effectiveness in predicting breast cancer based on the provided metrics. While the Random Forest and KNN models exhibit slightly superior performance, the Decision Tree model remains competitive.

Collectively, these insights underscore the pivotal role of predictive analytics in advancing breast cancer detection, with the Random Forest and KNN models poised as prime candidates for further exploration and application in real-world scenarios.

### Recommendations

My recommendation is to develop an API/model that enables patients to input their symptoms for predicting the likelihood of a benign or malignant tumor.

Considering the Random Forest Classifier's superior accuracy, it would be prudent to conduct an in-depth examination of feature importance and the potential for overfitting. Rigorous regression testing is essential prior to deploying the API/model to minimize false positives and false negatives.

### Limitations

It's important to acknowledge the study's limitations due to its reliance on a potentially outdated and constrained dataset. A more recent and comprehensive dataset could provide a more accurate representation of current trends and factors influencing breast cancer prediction. Additionally, the dataset's features might not capture all relevant factors that contribute to breast cancer prediction. To enhance the model's accuracy, considering the incorporation of additional clinical, genetic, or lifestyle-related features could be valuable.

### Risk

Predictive models in the medical domain pose a risk of false positives (classifying benign as malignant) and false negatives (missing malignant cases). False positives could trigger unnecessary distress and invasive procedures, straining healthcare resources. False negatives may lead to delayed treatment, harming patient outcomes and intervention efficacy. Thorough testing is imperative to mitigate these risks and ensure the well-being of patients.

# References

1. Pmotta. (2021, June 6). Breast cancer prediction. Kaggle  <https://www.kaggle.com/code/pmotta/breast-cancer-prediction/input>
2. Patel, J., Patel , U., Patel, R., & Shah, P. (2019, April 28). Breast Cancer Analysis. <https://rstudio-pubs-static.s3.amazonaws.com/491489_b86f191488ab4ed0a37e7a95c839a8f4.html>