Project 3 Milestone 3 - White Paper

Topic

Breast cancer is a common cancer that many people have encountered, whether they themselves have survived it or a loved one has been affected. This issue is personally relevant to me, as an aunt had a breast removed due to breast cancer, and a childhood friend has been affected by it because his mother was diagnosed. This project aims to analyze the characteristics of tumor cell nuclei and determine if these characteristics can be used to distinguish between benign and malignant tumors.

Background History

I found a dataset on Kaggle.com containing breast cancer diagnostic data from Wisconsin. Upon further research, I discovered the original source of the dataset and will be using the version downloaded directly from there.

The study the dataset is from was conducted in 1992 and aimed to diagnose breast tumors without requiring a full biopsy. Fine needle aspirations (FNAs) were used and digitized to analyze characteristics and features of individual cells and/or cell clumps (Street, 1993). The study used image processing and machine learning techniques.

Research Questions

The goal of this project is to analyze breast cancer diagnostic data to determine whether specific characteristics of tumor cell nuclei can help predict whether a tumor is benign or malignant. Specifically, I aim to answer the following research questions:

1. Can the characteristics of tumor cell nuclei (e.g., size, shape, texture) predict whether a tumor is benign or malignant?

2. Which variables in the dataset are the most significant indicators for distinguishing between benign and malignant tumors?
3. How accurate can a predictive model be in classifying breast tumors based on these characteristics?

Data Explanation

The dataset includes 32 variables, one of which will serve as the target variable, making it ideal for model development. This dataset will allow for an in-depth exploration of the relationship between tumor cell nuclei characteristics and diagnoses.

Methods

Data Preparation

I first checked for any null values in the dataset and handled them appropriately. The 'id' column, which was not needed for the analysis, was dropped. I then created a binary target column, 'd_bool', by converting the 'diagnosis' column into a binary format (malignant or benign). I checked the unique values in this new column to ensure correctness. Finally, I examined the data types of all variables in the dataset to ensure consistency and compatibility with the models.

Exploratory Data Analysis (EDA)

For the exploratory data analysis (EDA), I began by visualizing the distribution of the target variable using a simple countplot. I then performed a univariate analysis on the features by plotting histograms for each of the 30 variables, excluding the two target variables. To better understand the distribution of key features, I created boxplots for the radius, perimeter, and area variables with respect to the target column 'd_bool'. Following this, I calculated the correlation matrix and specifically focused on extracting the correlation values with respect to the target

column, 'd_bool', to assess which features were most strongly associated with tumor classification.

Model Building

I applied feature scaling to normalize the data and initially used an 80/20 train-test split. I tested three models: Logistic Regression, Random Forest Classifier, and Support Vector Machine (SVM). After running the models, I then re-ran the same three models with a 70/30 train-test split and further applied 5-fold cross-validation to the Logistic Regression model to assess its robustness and generalizability. During this process, I focused on analyzing model performance across different splits and evaluating the impact of cross-validation. Additionally, I extracted and displayed the coefficients of the logistic regression model to evaluate the importance of each feature in predicting the target variable.

Analysis

The countplot visualizing the target variable revealed around 200 malignant records and approximately 350 benign records. The histograms showed a decent distribution, but overall, the values tend to be on the smaller end, which is expected given the scale of nucleus measurements. Additionally, the box plots indicated that, while there is some overlap, malignant diagnoses generally have larger values in the 'radius', 'perimeter', and 'area' variables.

Correlation Matrix

The correlation values show the relationships between various features and the target variable, 'd_bool' (benign or malignant tumors). The strongest correlations are with 'concave_points3' (0.79), 'perimeter3' (0.78), 'concave_points1' (0.78), 'radius3' (0.78), and 'perimeter1' (0.74), highlighting that shape and size features like concave points, perimeter, and radius are key predictors. Other significant positive correlations include 'area3' (0.73) and 'area1'

(0.71), emphasizing the importance of size. In contrast, weak correlations with 'fractal_dimension2' (0.08) and 'symmetry2' (-0.01) suggest minimal impact on classification, while features related to texture and smoothness, like 'texture3' (0.46) and 'smoothness3' (0.42), show moderate influence but are less important than size and shape features.

Model Analysis

The results showed that with the 80/20 split, Logistic Regression achieved an accuracy of 96.49%, Random Forest Classifier achieved 93.86%, and Support Vector Machine reached 96.49%. When using the 70/30 split, the accuracy for Logistic Regression increased to 97.66%, Random Forest Classifier improved to 95.91%, and Support Vector Machine reached 97.08%. Finally, after applying 5-fold cross-validation on the Logistic Regression model with the 70/30 split, the accuracy further increased to 97.74%.

Feature Importance. The feature importance from the logistic regression model reveals that characteristics related to the size, shape, and texture of the cell nuclei are most influential in distinguishing between benign and malignant tumors. Key features include 'texture3' (1.47), 'radius3' (1.31), and 'area3' (1.16), which reflect the size and surface texture of the cell nuclei. Features like 'concavity3' (0.94) and 'concave_points3' (0.94), which measure the irregularities or indentations in the cell nuclei, also play an important role. In contrast, features such as 'compactness3' (0.01) and 'smoothness1' (0.13) have minimal impact, indicating that smoothness and compactness are less important for classification. Overall, size and shape characteristics are the primary factors driving the classification, with less emphasis on smoothness and texture properties.

Conclusion

This project successfully demonstrated that characteristics of tumor cell nuclei, particularly related to size, shape, and texture, are strong indicators for distinguishing between benign and malignant tumors. The logistic regression model achieved the highest accuracy, especially after applying 5-fold cross-validation, highlighting the importance of size and shape features, such as 'radius3', 'area3', and 'concave_points3', in tumor classification. Although features related to smoothness and compactness had minimal impact, the analysis revealed that these morphological characteristics can be effectively leveraged to predict tumor malignancy. This approach provides a robust framework for breast cancer diagnosis and further model refinement.

Assumptions

There were no significant assumptions to consider. As always, it is assumed that all information and data were ethically sourced and that sensitive patient data was excluded during the data collection process.

Limitations

The dataset contains 569 records, which is a reasonable size, but it may still be somewhat small for more complex analyses. While a larger dataset would have been preferable, it appears that this was the maximum available from the original study team. There were no major challenges or issues that stood out during the analysis.

Future Uses/Additional Applications

I believe the dataset could be expanded to include more patient records, which would improve the analysis. A larger dataset would provide greater robustness to the model, and ideally, I would aim for a highly accurate model with an accuracy around 98-99%. This or an even more refined model could assist doctors in making accurate diagnoses for their patients.

Recommendations

I recommend increasing the record size of this dataset. While the data is accurate and free from errors, it would benefit from a larger number of records to strengthen the model's performance. Additionally, leveraging AI could improve model accuracy further. AI has proven to be an invaluable tool, and its integration into healthcare could help create models that accurately and safely diagnose patients.

Ethical Assessment

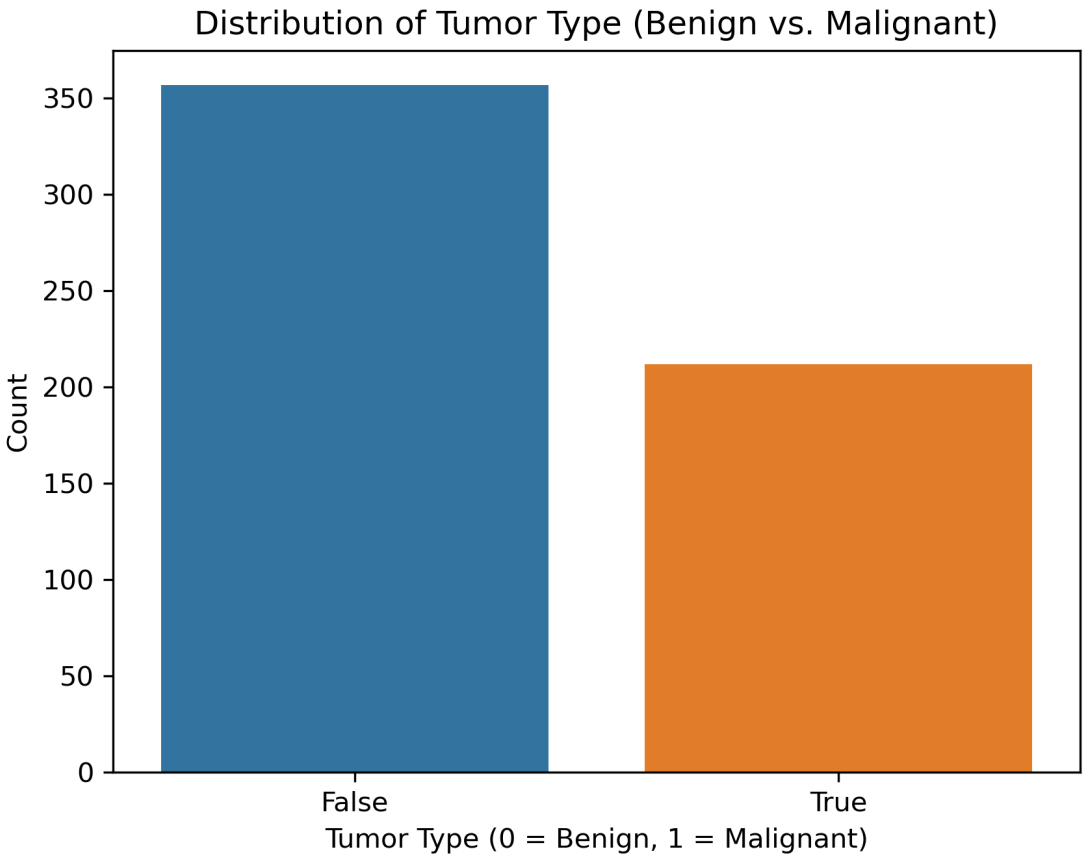
There were no major ethical concerns in this analysis. I dropped the 'id' column and created a binary target variable from the 'diagnosis' column. The dataset was ethically sourced, and I do not claim ownership of it. No personal data is included in the dataset, so there are no security concerns.

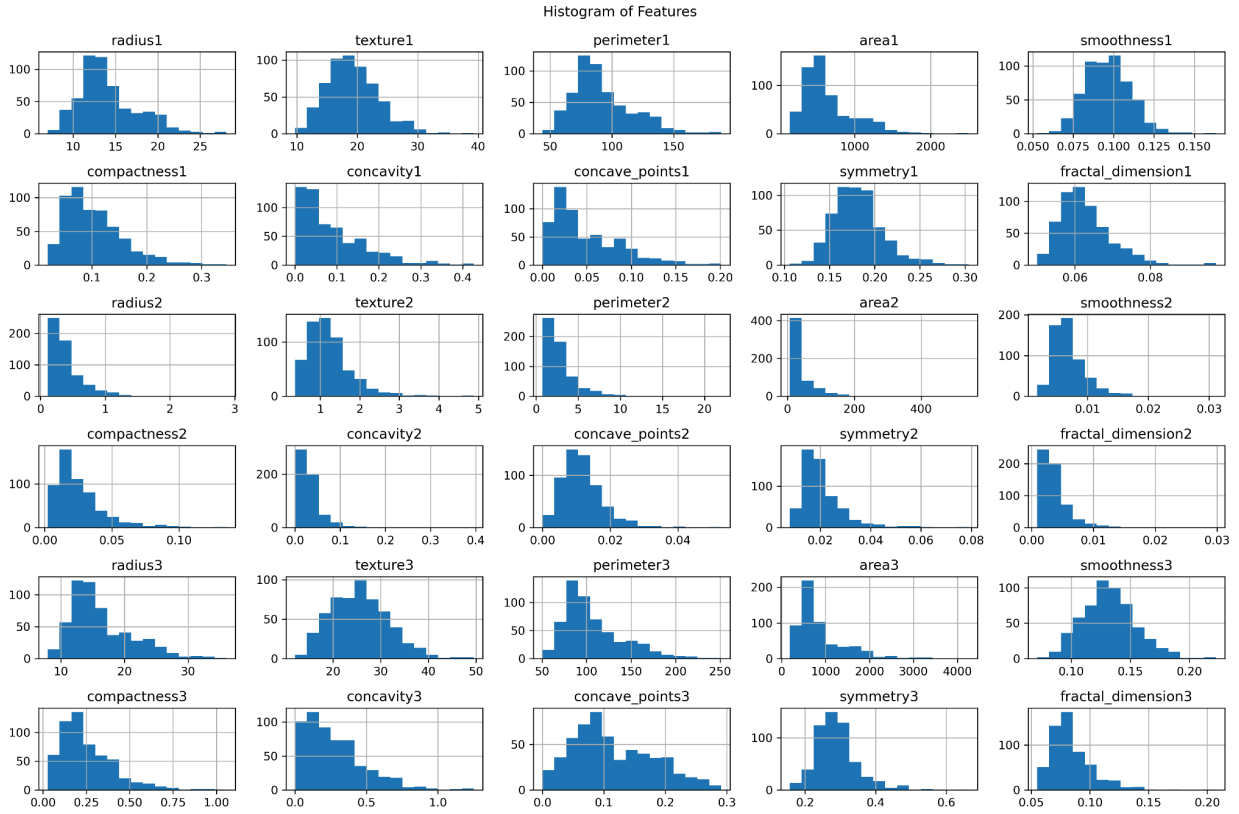
References

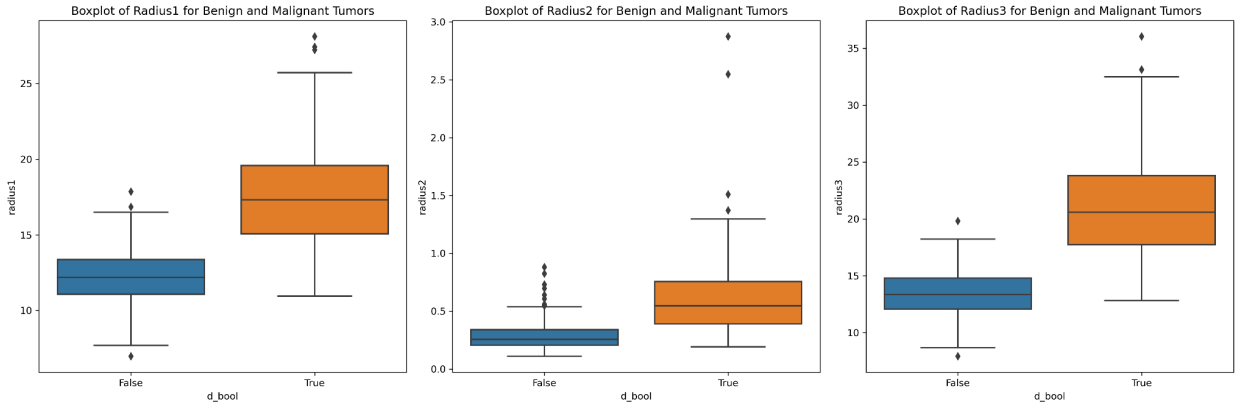
Breast cancer wisconsin (Diagnostic). UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

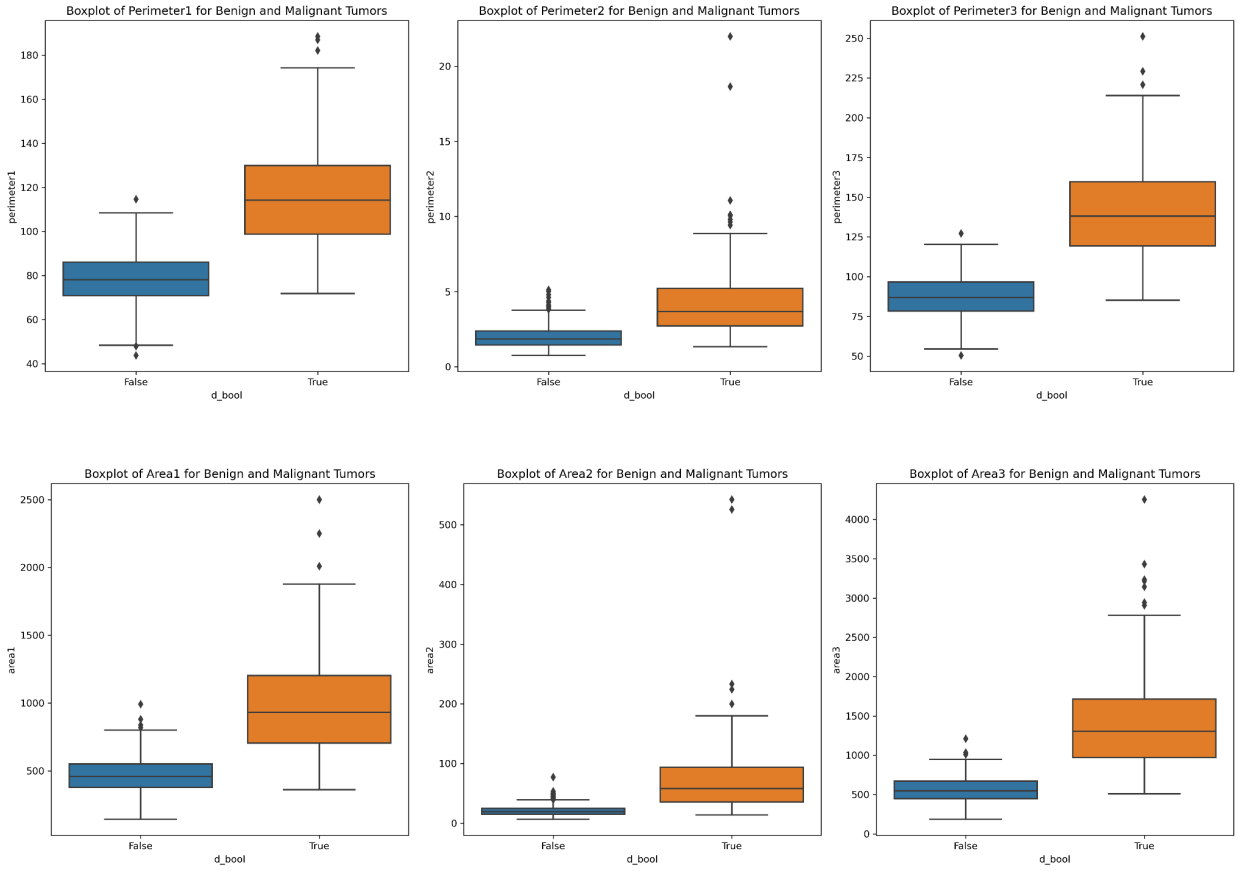
Street, W., Wolberg, W., & Mangasarian, O. (1993, July 29). [PDF] nuclear feature extraction for breast tumor diagnosis | semantic scholar. https://www.semanticscholar.org/paper/nuclear-feature-extraction-for-breast-tumor-Street-Wolberg/53f0fbb425bc14468eb3bf96b2e1d41ba8087f36

Appendix A: Data Visualizations




Histogram of Features
radius1
texture1
perimeter1
area1
smoothness1
compactness1
concavity1
concave_points1
symmetry1
fractal_dimension1
radius2
texture2
perimeter2
area2
smoothness2
compactness2
concavity2
concave_points2
symmetry2
fractal_dimension2
radius3
texture3
perimeter3
area3
smoothness3
compactness3
concavity3
concave_points3
symmetry3
fractal_dimension3


Boxplot of Radius1 for Benign and Malignant Tumors
radius1
d_bool
False
True
Boxplot of Radius2 for Benign and Malignant Tumors
radius2
d_bool
False
True
Boxplot of Radius3 for Benign and Malignant Tumors
radius3
d_bool
False
True


Boxplot of Perimeter1 for Benign and Malignant Tumors
perimeter1
d_bool
False
True
Boxplot of Perimeter2 for Benign and Malignant Tumors
perimeter2
d_bool
False
True
Boxplot of Perimeter3 for Benign and Malignant Tumors
perimeter3
d_bool
False
True
Boxplot of Area1 for Benign and Malignant Tumors
area1
d_bool
False
True
Boxplot of Area2 for Benign and Malignant Tumors
area2
d_bool
False
True
Boxplot of Area3 for Benign and Malignant Tumors
area3
d_bool
False
True

Appendix B: Potential Audience Questions

1. How did you handle missing values in the dataset, and what impact might missing data have on the results of your analysis?
    a. Fortunately, there were no missing values in the dataset.
2. You mentioned that most of the records are labeled as 'benign.' Did you consider using techniques like oversampling or undersampling to balance the dataset, or did you account for any potential class imbalance in your models?
    a. I hadn’t considered an imbalance or oversampling in the dataset.
3. How did you determine the choice of models (Logistic Regression, Random Forest, and Support Vector Machine)? Were there any other models you considered, and why were these selected over others?
    a. I originally tested these three models along with KNN, but KNN yielded a very low accuracy. Since the three models I selected yielded accuracies in the high 90s, I decided to stick with them.
4. You indicated that the Logistic Regression model achieved 93.37% accuracy. Did you conduct any hyperparameter tuning to optimize the models?
    a. I did try hyperparameter tuning, but it yielded the same accuracy.
5. Given the strong performance of your models, how do you ensure that the model doesn't overfit to the training data, and what steps would you take to mitigate overfitting in future analyses?
    a. I utilized cross-validation and tried both 80/20 and 70/30 train-test splits, which help ensure the model generalizes well to unseen data. In future analyses, I would explore additional techniques, such as regularization or early stopping, to further

reduce the risk of overfitting. I would also experiment with different models, like Gradient Boosting, which can be more robust against overfitting.

6. What steps have you taken to evaluate the generalizability of your models? Do you plan to test your models on an external dataset or use cross-validation techniques?
    a. Unfortunately, I don't have an external dataset to test on, but I did apply cross-validation to the Logistic Regression model to achieve a slightly higher accuracy.
7. You also ran the models using a 70/30 split. What do you anticipate this change would've brought, and why do you think it would've impacted the results?
    a. I wasn't looking for anything specific, but I wanted to try different approaches to achieve the highest possible accuracy. The difference was minimal.
8. Given the relatively small size of the dataset, how confident are you in the robustness of your results? How might the models perform with a larger dataset?
    a. I believe the model would have performed relatively well with a larger dataset. While the 97% accuracy may not be as high with a larger dataset, I'm confident it would still maintain a relatively high accuracy.
9. What other potential applications could this analysis have beyond predicting benign vs. malignant tumors, and how could these models contribute to medical decision-making or patient outcomes in practice?
    a. I believe this analysis could be applied to other areas in the medical field. For example, women struggling with PCOS could benefit from similar technology. There are many medical conditions where technological advancements could help predict causes and improve diagnostics.

10. How do you plan to validate the performance of your models beyond the training and test split, and what additional techniques might you use to ensure robustness?
    a. I would have liked to test the model on an external dataset. I also would have preferred conducting the analysis on a larger dataset, but I'm still happy with the results.