Modeling and Analysis of the Wisconsin Diagnostic Breast Cancer Dataset

Raul “JR” Saenz

Austin Community College

COSC 3380-003

Professor Katrompas

September 2025

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset is well-known and often used for teaching and testing machine learning models. It includes data from 569 patients, with 30 different features that describe the cell nuclei in breast tissue samples. Each case is labeled as either malignant or benign, making it useful for studying how certain features can help distinguish between the two groups.

The purpose of this assignment was to use the dataset to practice basic data analysis and modeling with Python. Using Python, I generated descriptive statistics, created visualizations, and built a simple predictive model. The visualizations include box plots, histograms, and pair plots that illustrate how selected features are distributed and how they compare between benign and malignant tumors. I also examined whether there were any outliers in the data and why they matter.

Finally, I trained a logistic regression model and measured its performance using accuracy, precision, recall, and F1 score. These metrics are especially important in the context of breast cancer detection, because mistakes in prediction can have serious consequences. In the following sections, I present the analysis results and discuss their implications for the dataset and the model.

Question and Answer Section

The following section addresses the questions provided in the assignment. Each part includes a description of the data visualizations, an analysis of any outliers present, an evaluation of the model’s performance metrics, and a discussion of modeling choices.

Box Plots

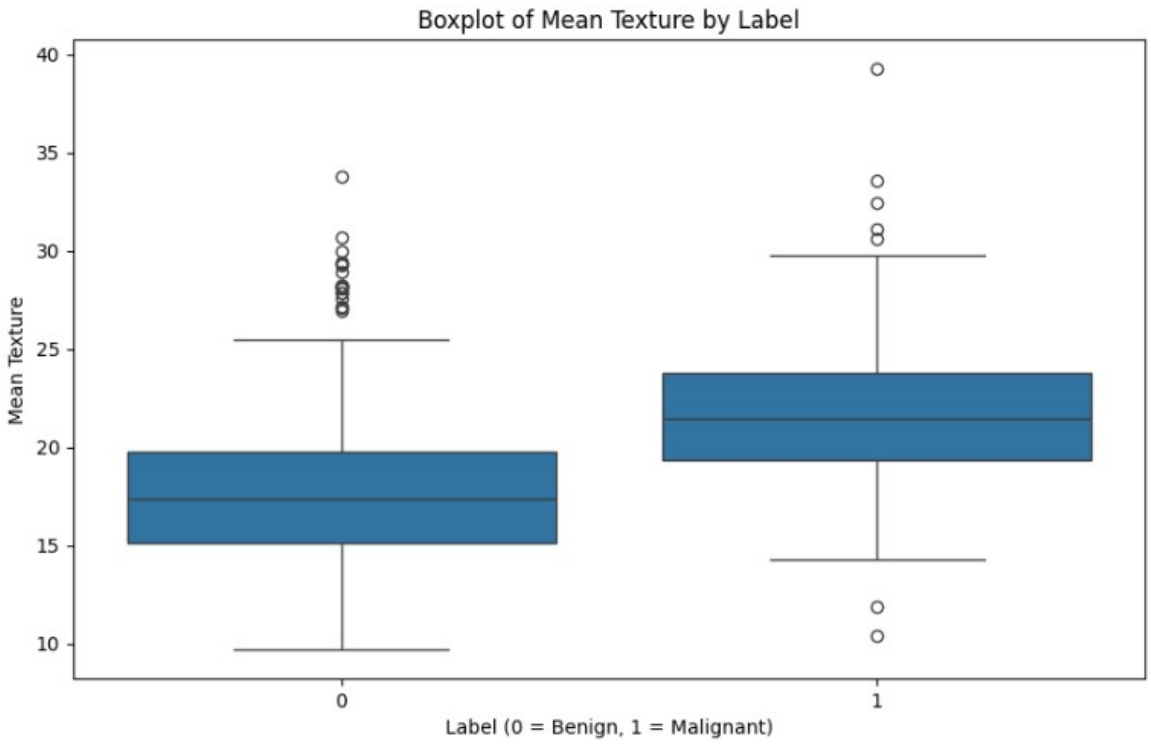

Figure 1. Box plot of mean texture values comparing benign (0) and malignant (1) tumors

A box plot is a graphical representation that summarizes the distribution of a dataset using five main descriptive statistics. Those are the minimum, the first quartile (Q1), the median, the third quartile (Q3), and the maximum. The central box represents the interquartile range (IQR). The line inside the box indicates the median. The “whiskers” extend to values within 1.5 × IQR. Any points beyond the whiskers are considered outliers and are plotted individually. This makes box plots especially useful for identifying differences between groups and spotting unusual values in the data.

In the figure above, the distribution of mean texture values is shown separately for benign tumors (label 0) and malignant tumors (label 1). The median texture value for malignant tumors is significantly higher than that of benign tumors, indicating that texture is a key distinguishing feature between the two classes. The boxes also indicate that malignant tumors tend to have a wider interquartile range, suggesting greater variability in texture measurements compared to benign tumors.

Several outliers are visible for both groups. For benign cases, most outliers occur on the higher end of the distribution, while for malignant cases, they appear at both extremes. These outliers could represent tumors with unusual cellular structures and may influence the performance of statistical models if not handled properly. Overall, the box plot indicates that mean texture is a significant feature for distinguishing between malignant and benign tumors.

Histograms

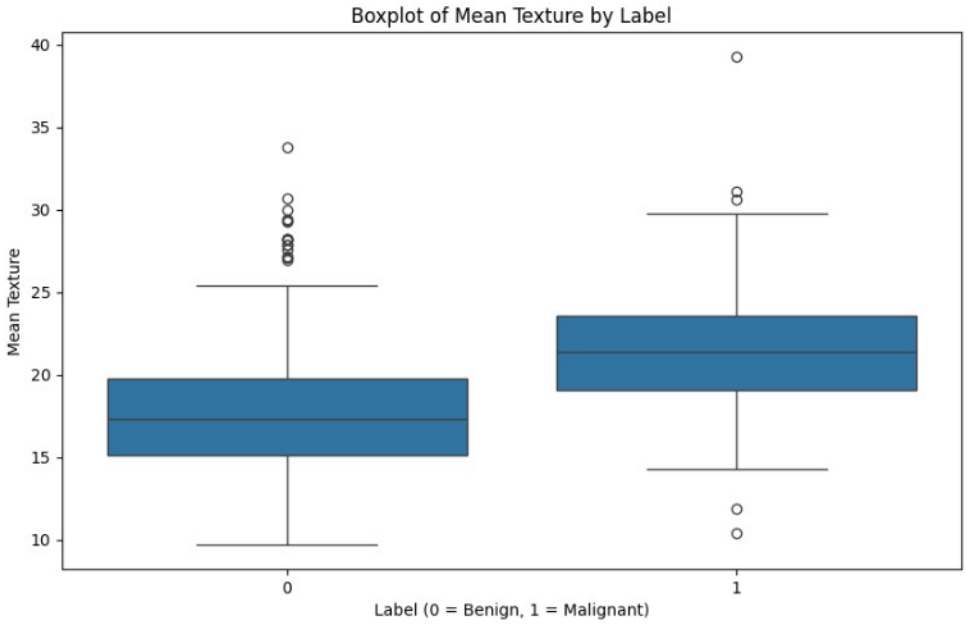

Figure 2. Histogram of mean radius values across all 569 patients in the WDBC dataset. This histogram shows the frequency distribution of tumor nuclei radii with an overlaid density curve.

A histogram is a type of graph that displays the distribution of a numerical variable by grouping values into intervals, or buckets, showing how many observations fall into each bucket. Making it easy to see the shape of the data, including its central tendency, spread, and whether the distribution is symmetric or skewed.

In the figure above, the histogram shows the distribution of the mean radius feature for all 569 tumors in the dataset. The x-axis represents the mean radius values, while the y-axis shows how often each range of values occurs. Most tumors fall in the 10–15 range, with the peak around 12–13. The

distribution is not perfectly symmetrical; instead, it is slightly right-skewed, meaning there are fewer tumors with very large radii, but they still occur.

This skewness is essential for modeling because it suggests that while most tumors share a fairly typical size, there are some with much larger radii that could affect averages and influence how a model learns. The density curve overlaid on the histogram helps highlight this trend, showing a single central peak with a long tail to the right.

Pair Plots

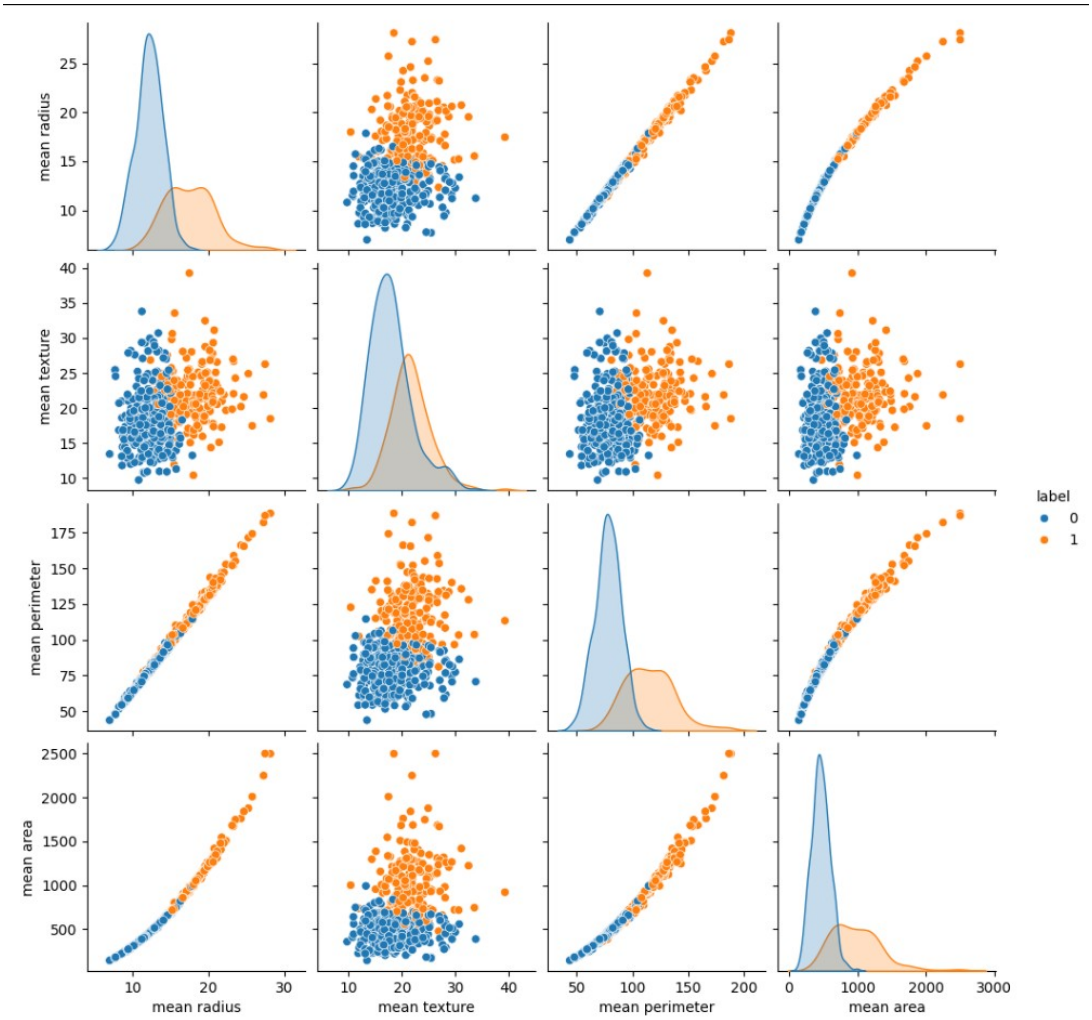

Figure 3. Pair plot of selected features (mean radius, mean texture, mean perimeter, and mean area) from the WDBC dataset. Here 12 scatter plots and four distribution plots that compare two features and their distribution.

A pair plot displays the relationship between several features in a single figure. The off-diagonal parts of the grid are scatter plots that compare two features at a time, while the diagonal parts show the

distribution of a single feature. This allows for the visualization of both the distribution of features and their interconnections.

In this pair plot, the features' mean radius, mean texture, mean perimeter, and mean area are compared. The scatter plots show strong positive relationships between mean radius, perimeter, and area. This makes sense because as a tumor gets larger in radius, its perimeter and area also increase. Malignant tumors (orange, label 1) are grouped more toward the higher values, while benign tumors (blue, label 0) are grouped lower.

The distributions on the diagonal confirm these trends. Malignant tumors tend to have larger radius, perimeter, and area values, while benign tumors mostly cluster on the smaller side. For mean texture, there is more of an overlap, but malignant tumors still lean higher overall. These patterns suggest that certain features, such as radius and area, are stronger indicators of malignancy than others.

Outliers

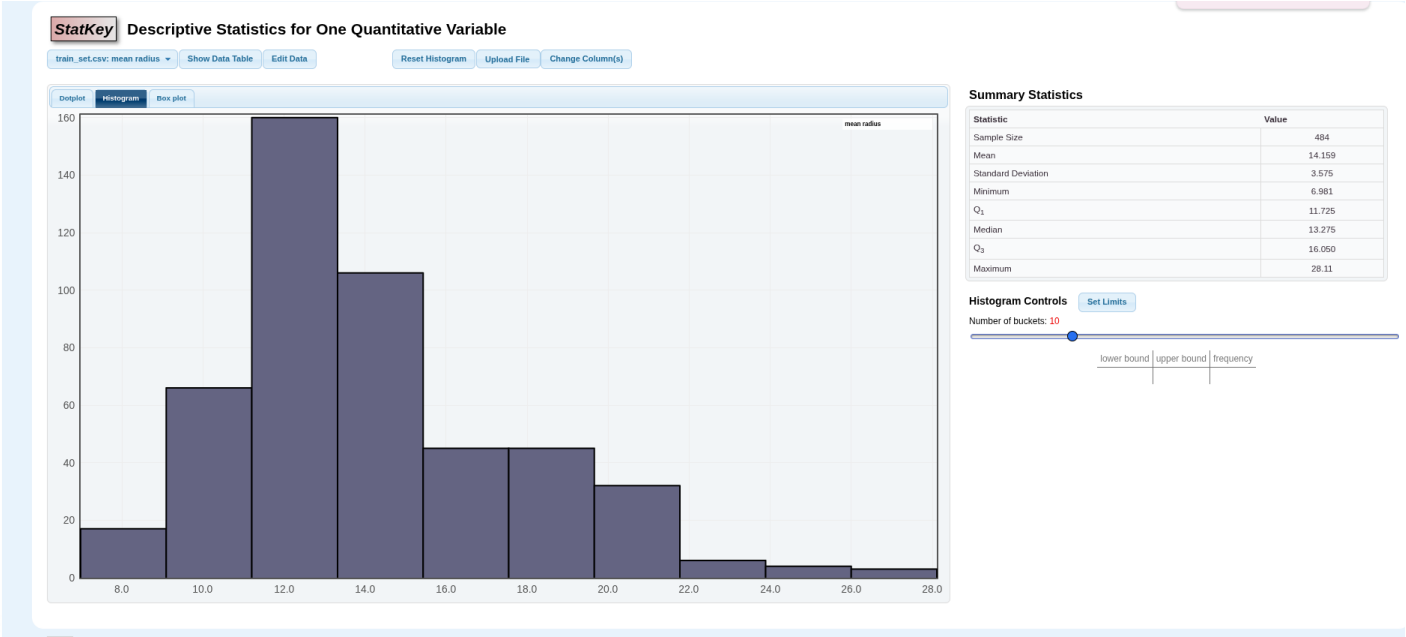

Figure 4. Histogram and summary statistics for mean radius generated using Lock5 StatKey. The mean (14.159) and the standard deviation (3.575) confirm the presence of outliers, with the maximum value (28.11) exceeding the upper cutoff of 24.884 under the standard deviation method.

Outliers are data points that fall far outside the typical range of values in a dataset. Because outliers can significantly impact averages and model performance, it is essential to identify and understand them before drawing conclusions from the presented data.

To check for outliers, I applied the standard deviation method to the feature mean radius. The average was 14.159 with a standard deviation of 3.575, according to the StatKey summary statistics. Using the formula mean ± 3 × standard deviation, the cutoff values were 3.434 on the low end and 24.884 on the high end. Any tumor with a mean radius outside this range is considered an outlier. StatKey showed that the maximum observed value was 28.11, which is above the cutoff, confirming that there are unusually large tumors in the dataset.

The histogram and boxplots also visually support this finding, showing that most tumors cluster between 10 and 15 in mean radius, while a handful extend much further to the right. These outliers represent rare cases with unusually large cell nuclei. Outliers are important because if left unchecked, they can distort statistics like the mean and potentially reduce the accuracy of models trained on the data. In the context of breast cancer detection, however, such extreme cases may also carry crucial medical meaning, and removing them could risk losing valuable information.

Model Performance Metrics

Table 1

Logistics Regression Model Performance Metrics on WDBC Dataset

Metric	Value
Accuracy	0.9647
Precision	0.9714
Recall	0.9444
F1 Score	0.9577

Note. Values are rounded to three decimal places. Metrics were computed on the test set using logistics regression.

To evaluate the performance of the logistic regression model, I measured four key metrics: accuracy, precision, recall, and the F1 score. These metrics offer different perspectives on how the model distinguishes between malignant and benign tumors.

The model achieved an accuracy of 96.5%, meaning most predictions were correct. While accuracy gives a good overall sense of model quality, it does not always reveal whether mistakes are more common in one class than the other. In medical contexts, such as breast cancer detection, it is essential to look beyond accuracy because the cost of different types of errors is not equal.

Precision was 97.1%, showing that when the model predicted a tumor as malignant, it was almost always correct. High precision is essential because it means fewer patients would be incorrectly told they have cancer (false positives).

Recall was 94.4%, meaning the model successfully identified nearly all malignant cases. This metric is especially critical in cancer detection because a false negative (failing to detect an actual malignant tumor) could lead to a missed diagnosis and delayed treatment. Although the recall is slightly lower than the precision, it remains powerful, indicating that the model is susceptible to malignant cases.

The F1 score, which balances precision and recall, was 95.8%. This demonstrates that the model strikes an excellent balance between minimizing false alarms (false positives) and maximizing the detection of actual cancer cases (true positives).

All four metrics indicate that the logistic regression model performs very well. Recall may be the most critical in this context, because missing a cancer diagnosis can have the most serious consequences. Precision also matters because false positives can cause unnecessary stress and medical procedures. The balance reflected in the F1 score shows that the model handles both concerns effectively.

Model Stability on Repeated Runs

To test the stability of the logistic regression model, I ran the model five times using the same training and testing data files. Each run produced identical metrics: an accuracy of 96.5%, a precision of 97.1%, a recall of 94.4%, and an F1 score of 95.8%.

The reason the results did not change is that the model was trained and evaluated on the same train and test sets each time. Logistic regression is deterministic, meaning that given the same input data and parameters, it will always converge to the same solution. Since the data split into training and testing sets was already fixed when the data.py script was first executed, rerunning the model on these files does not introduce any randomness.

However, if data.py were run again before model.py, the dataset would be reshuffled and new train and test files would be created. In that case, the metrics would most likely change from run to run because the model would be trained on different data.

This demonstrates that the model is consistent when applied to the same data. Any changes in the metrics would only occur if the data were split differently, if cross-validation were used, or if randomness were introduced into the training process.

Variation Across Different Data Splits

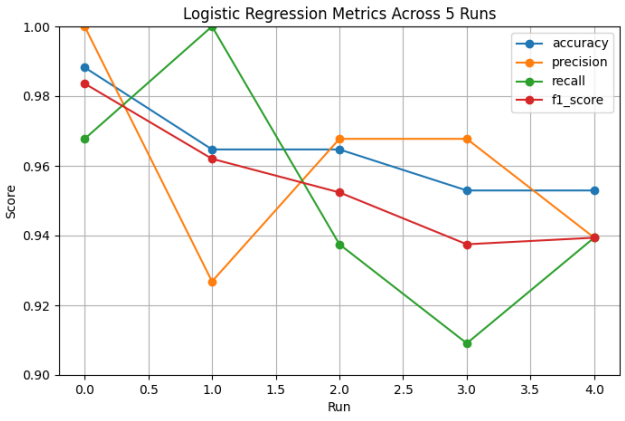

Figure 5. Logistic regression performance across five runs with different training and testing splits of the WDBC dataset.

The metrics do change slightly from one run to the next. This happens because each time the data is shuffled and split into training and testing sets, different patients end up in the test set. Since the dataset is relatively small (569 patients), even small shifts in which cases are included can change the precision, recall, and other values. Across the five runs, however, the scores remain consistently high, showing that the model is stable. The averages of the five runs are approximately 0.96 for accuracy, 0.97 for precision, 0.94 for recall, and 0.96 for F1 score. This strategy of repeating the process with different splits is valid and is called cross-validation. It helps confirm that the model is not overly dependent on one particular train/test split and that its performance is generalizable across different subsets of the data.

Suitability of Linear Regression for Breast Cancer Classification

Linear regression is not the best choice for the WDBC dataset. The WDBC involves a binary classification problem where tumors are either malignant or benign. Linear regression is designed for predicting continuous outcomes, not categorical ones. As shown in Katrompas and Yan's study, decision trees and random forests often outperform linear regression in this context. Logistic Regression, which aligns with the approach used in our assignment, would be a better choice because it demonstrates strong performance across accuracy, precision, recall, and F1 score.

References

Aurélien Géron. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. "O'Reilly Media, Inc."

Curve Fitting. (n.d.). Retrieved from www.originlab.com website: https://www.originlab.com/index.aspx?go=Products/Origin/DataAnalysis/CurveFitting

GeeksforGeeks. (2017, November 21). Cross Validation in Machine Learning. Retrieved from GeeksforGeeks website: https://www.geeksforgeeks.org/machine-learning/cross-validation-machine-learning/

GeeksforGeeks. (2024, May 7). What is Outlier Detection? Retrieved from GeeksforGeeks website: https://www.geeksforgeeks.org/data-analysis/what-is-outlier-detection/

Katrompas, A., & Yan, Y. (n.d.). Empirical Analysis and Comparison of Machine Learning Methods and Methodologies (pp. 1–12). Texas State University.

Relative Frequency Histogram: Definition and How to Make One. (2023, January 29). Retrieved September 22, 2025, from Statistics How To website: https://www.statisticshowto.com/relative-frequency-histogram/

Box Plot (Box and Whiskers): How to Read One & Make One in Excel, TI-83, SPSS. (2024, October 13). Statistics How To. https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/box-plot

wdbc function - RDocumentation. (2024). Retrieved September 22, 2025, from Rdocumentation.org website: https://www.rdocumentation.org/packages/mclust/versions/6.1/topics/wdbc