2025

Cancer Classification Analysis Report

PDAN8411 PART 2

MAX WALSH – ST10203070

26 MAY 2025

Table of Contents

Introduction

The objective of this analysis is to develop a predictive model that can aid a medical scheme in identifying customers who may require cancer benefits. This effort is to streamline the benefit allocation and facilitate early interventions (OpenAI, 2025). The chosen dataset is a widely used and publicly available breast cancer dataset retrieved from Kaggle and originally compiled by the University of Wisconsin (OpenAI, 2025). The dataset contains diagnostic measurements derived from digitised images of fine needle aspirates (FNA) of breast masses (OpenAI, 2025). Of which, each observation is labelled benign (B) or malignant (M), based on diagnosis (Learning & Ovsen, 2016). The process of this project involved a full data science pipeline, including exploratory data analysis (EDA), feature selection, model training, evaluation, and interpretation of results. The goal is to produce a reliable classification model that can benefit the client in assessing cancer risk and enable data-driven decisions about benefit distribution (OpenAI, 2025).

Dataset Justification

Source: Kaggle – “Breast Cancer Wisconsin (Diagnostic) Data Set” by UCI Machine Learning & Ovsen (Learning & Ovsen, 2016).

Size: 569 records, 30 features.

Target variables: diagnosis.

The Breast Cancer Wisconsin (Diagnostic) dataset was chosen for its high relevance, data quality, and suitability for binary classification tasks involving cancer diagnosis (OpenAI, 2025). The following are some key reasons for selection:

- Clinical Relevance: Given the objective of identifying patients who may require cancer-related medical benefits, this dataset is suitable as it directly addresses breast cancer diagnosis (OpenAI, 2025).
- Quality and Completeness: The dataset contains 569 samples with no missing values in key features (Learning & Ovsen, 2016). Providing robust input for machine learning models.
- Balanced Target Distribution: The target variable (diagnosis) is reasonably balanced with 357 benign and 212 malignant (Learning & Ovsen, 2016). Supporting stable model training and a fair performance evaluation (OpenAI, 2025).
- Rich Feature Set: The dataset provides 30 continuous variables derived from diagnostic imaging, providing detailed and nuanced pattern recognition (Learning & Ovsen, 2016).
- Consistent Data Types: All predictors are ‘float64’ numeric values, providing compatibility with preprocessing and machine learning pipelines without additional type handling (Learning & Ovsen, 2016).
- Proven Benchmark: Given the frequent use of this dataset in academic and industrial settings, with well-documented behaviour across various algorithms, we can expect enhanced model benchmarks and reproducibility (OpenAI, 2025).

These stated attributes offer an ideal foundation for developing an interpretable, effective, and predictable model for the client’s cancer benefit management use case.

Algorithm Justification

Logistic Regression was this study's primary algorithm due to its interpretability, robustness, and suitability for binary classification problems (Muller & Guido, 2016). In a healthcare scenario like this, where transparency and explanation of results are crucial, Logistic Regression offers the advantage of clearly defined feature coefficients, enabling stakeholders to understand the influence of each variable on the model's predictions (Schober & Vetter, 2021).

Furthermore, Logistic Regression performs well when the relationships between features and the target variable are seen to be approximately linear (Thrane, 2023). Which, in this case, holds for many of the dataset predictors as seen in the correlation analysis (Learning & Ovsen, 2016). Logistic Regression is computationally efficient, handles large datasets well, and offers reliable probability estimates, which can be useful for risk stratification (OpenAI, 2025).

Model performance was compared before and after hyperparameter tuning to validate the algorithm choice. Previous analysis completed by others on Kaggle on this specific dataset showed that Logistic Regression offered high accuracy (98.25%) with hyperparameter tuning (Learning & Ovsen, 2016). Whilst more complex models like Random Forests and Gradient Boosting may yield marginally higher accuracy, interpretability would be compromised (OpenAI, 2025). Given the medical context of this application, this trade-off would be deemed unacceptable as decisions impact patient outcomes (Schober & Vetter, 2021).

Ergo, Logistic Regression offers a practical balance between predictive power and explainability, making it the preferred algorithm for this analysis.

Exploratory Data Analysis (EDA)

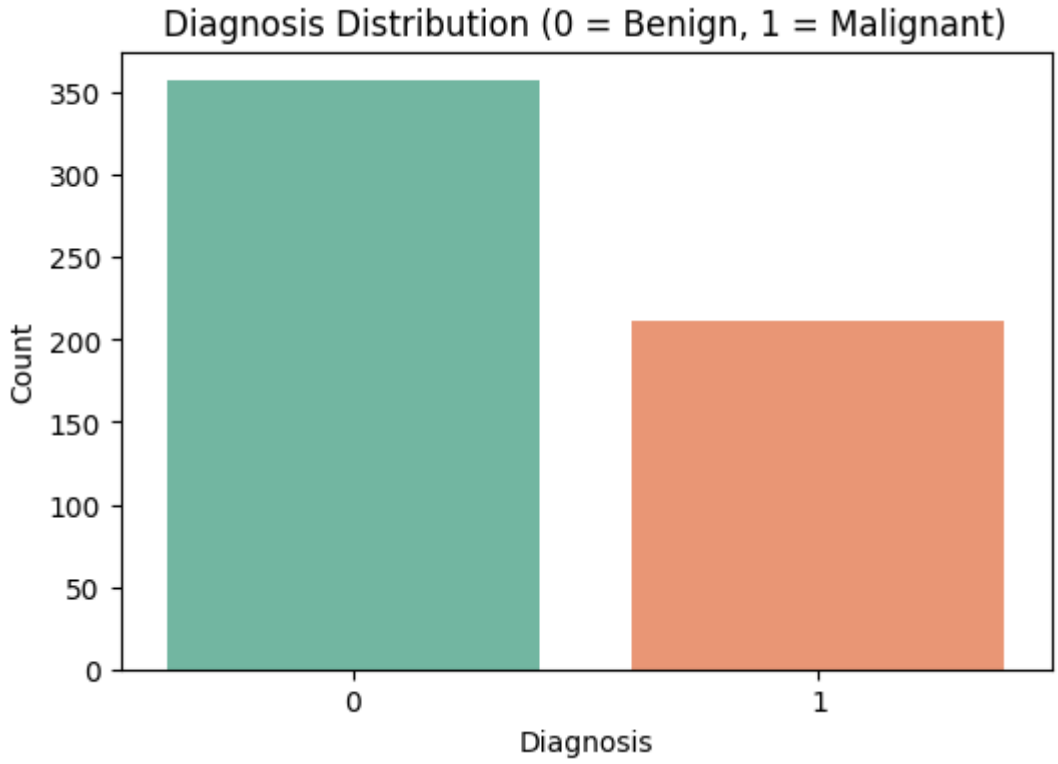

Figure 1: Diagnosis Class Distribution

The bar chart illustrated in Figure 1 shows the number of benign and malignant diagnoses within the dataset. Although benign vs malignant is slightly imbalanced, the class distribution is adequate for training a reliable classification model (OpenAI, 2025). At 357 benign and 212 malignant samples, the dataset supports both model generalisation and sensitivity to positive cancer cases (OpenAI, 2025).

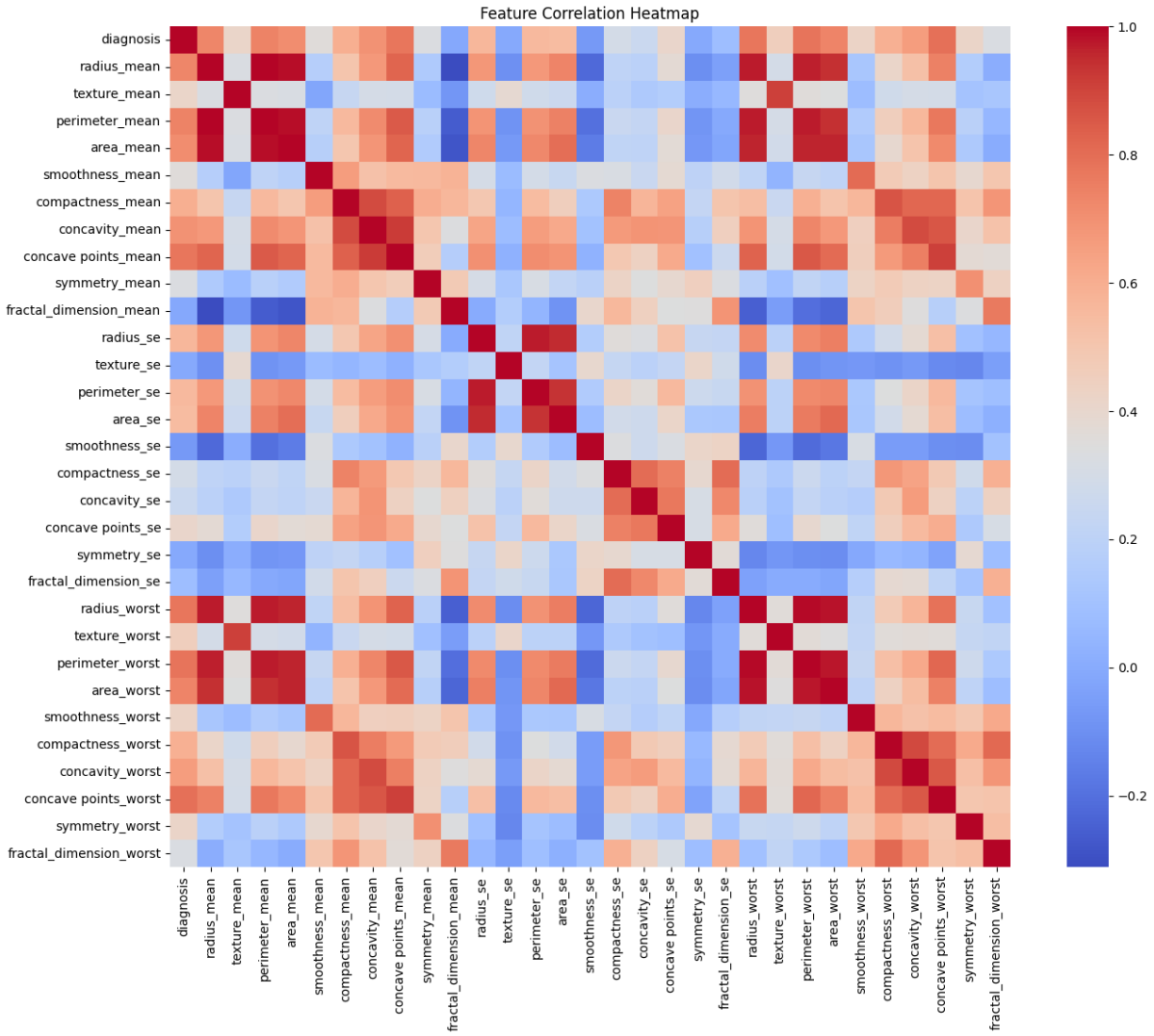

Figure 2: Feature Correlation Heatmap

The heatmap seen in Figure 2 reveals significant positive correlations between several features and the target diagnosis. Features that are seen to be strongly correlated with malignancy are radius_mean, concave points_mean, and perimeter_mean. Our data-driven feature selection is enhanced by these insights, whilst they also inform the dimensionality reduction strategy (OpenAI, 2025).

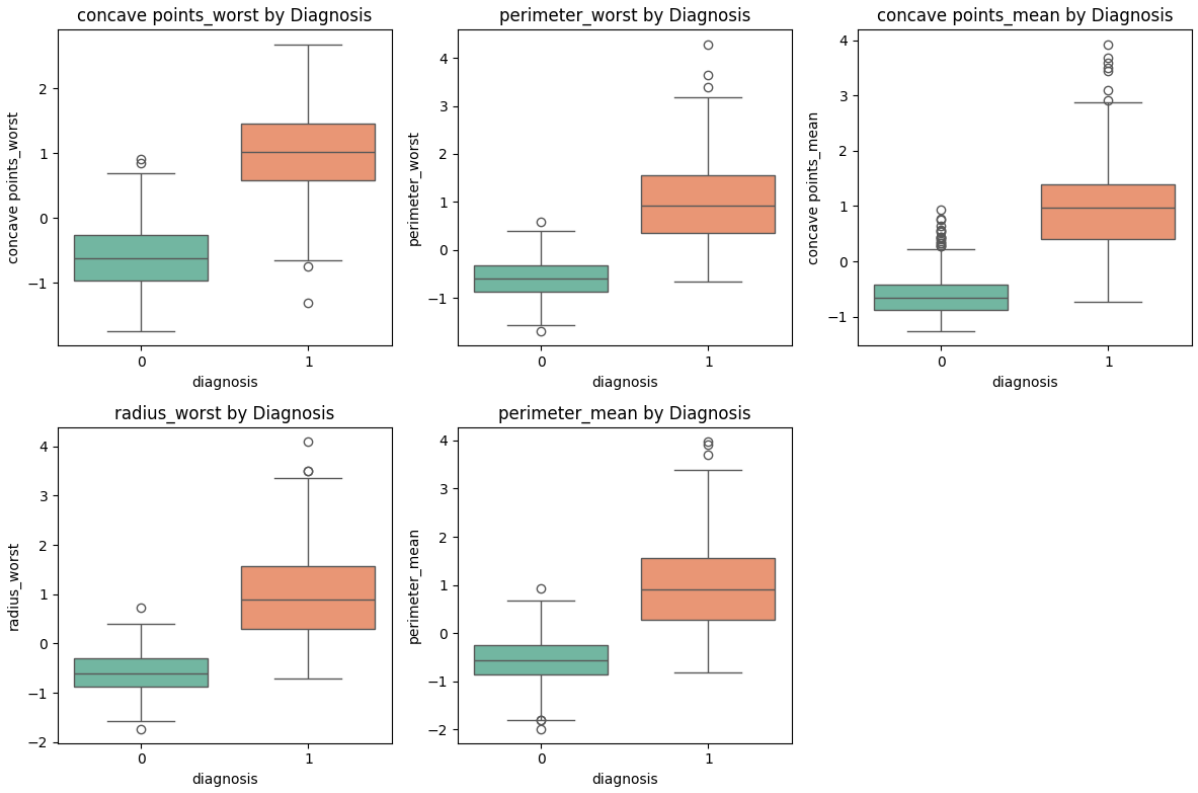

Figure 3: Box plots of top Features by Diagnosis

Figure 3 shows the box plots displaying the distribution of the top five most correlated features against the diagnosis variable. Malignant cases show notably higher values for features like concave points_worst, radius_worst, and perimeter_worst, whilst benign cases tend to cluster lower (OpenAI, 2025). Class separability is highlighted from these distinctions, confirming the diagnostic value of the selected features (OpenAI, 2025).

Modelling Process

The modelling process involved data cleaning, feature selection, and model development using logistic regression. Each step in this process was designed to ensure accuracy, reduce multicollinearity, and improve interpretability (OpenAI, 2025).

Data Cleaning

After the data inspection, the following decisions were made and executed to ensure clean data. Firstly, it was confirmed that the dataset contained no missing values, which is ideal as missing values can cause bias and lead to misinterpretation of the data (Ault, et al., 2025). Secondly, two irrelevant columns within the dataset – id and Unnamed: 32 – were removed. Thirdly, the target variable diagnosis was converted to binary to support supervised learning, where M (malignant) represented a 1 and B (benign) represented a 0 (OpenAI, 2025). Finally, all feature values were numeric and scaled using standardisation to ensure compatibility with logistic regression (Suzanne, 2023).

Feature Selection

The following steps were taken during the feature selection process to reduce the dimensions of the model and improve the model's efficiency and interpretability without compromising predictive power (Suzanne, 2023). To identify and remove highly collinear features to avoid redundancy and inflated variance, a correlation matrix was used, as seen in Figure 2 (Suzanne, 2023). Features with a correlation value greater than 0.9 were considered highly collinear (Suzanne, 2023). Correlation coefficients range from -1 to +1, where -1 indicates a perfect negative correlation, 0 means no correlation, and +1 indicates a perfect positive correlation (Wagavkar, 2024). After this, Recursive Feature Elimination (RFE) with logistic regression was applied to the 10 most important features: radius_mean, concavity_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, compactness_se, concavity_se, compactness_worst, and concavity_worst. RFE aims to identify the most relevant subset of features through iteratively removing features based on their importance (Vidhya, 2025).

Model Training

The steps taken during the model training process are as follows: The dataset was split into training and test sets, using an 80/20 ratio, respectively, with random sampling. This is essential to ensure generalisability, a term used to describe a model's ability to perform well on new, unseen data, rather than just trained data (Muller & Guido, 2016). A logistic regression model was trained using the default Scikit-Learn parameters – solver='liblinear', penalty='l2',

C=1.0, and max_iter=10000. Scikit-Learn is a Python data science library for machine learning (Muller & Guido, 2016). I'll now explain the meaning of these default parameters for the logistic regression model. The solver chosen, 'liblinear', was selected for its efficiency for small datasets and support for L1 and L2 regularisation (Muller & Guido, 2016). The penalty was set to 'l2' as L2 regularisation penalises large coefficients, aiding in reducing overfitting (Muller & Guido, 2016). Regularisation refers to techniques like L1 and L2 regularisation, to reduce the model from overfitting – a term describing a case when the model learns the training data too well (Muller & Guido, 2016). The regularisation parameter value, C, was set to 1.0, a higher value that recues regularisation, allowing the model to fit the training data more closely (Muller & Guido, 2016). Finally, the max_iter was set to 10000, which is an increased value from the default to ensure convergence (OpenAI, 2025). Convergence is essential as it means the model has reached a point of stability, and accurate predictions can be made, and it is especially important when applying recursive feature elimination (OpenAI, 2025). After this, the model's performance was evaluated using a range of metrics, including accuracy, precision, recall, F1 score, AUC (Area Under the Receiver Operating Characteristic Curve), and a confusion matrix (OpenAI, 2025). Later, a second training iteration was conducted with a regularisation parameter adjustment of C=0.1, to evaluate the effect of increased regularisation on the model's generalisation and performance (Muller & Guido, 2016).

These modelling process steps guided the way for a final logistic regression classifier that was both robust and interpretable (OpenAI, 2025). The combination of dimensionality with a stable training pipeline ensured a model that was optimised for accuracy and transparency, two necessary priorities in clinical decision support (OpenAI, 2025). After a trained and validated model was in place, the proceeding step was to assess its performance on unseen data using appropriate classification metrics and visual diagnostics.

Model Evaluation

The trained logistic regression model was evaluated on the 20% test set using a suite of performance metrics. Metrics used to evaluate include accuracy, precision, recall, F1 score, AUC, and the confusion matrix. To enhance interpretability, visuals like the Receiver Operating Characteristic (ROC) curve and confusion matrix heatmap were used.

Initial Model Performance (C=1.0)

- Accuracy: 96.5% - This metric shows the proportion of all correct predictions, reflecting the overall effectiveness, but may be misleading in an imbalanced dataset (Muller & Guido, 2016).
- Precision 95.35% - Showing that of all patients predicted as malignant, 95.35% were truly malignant (OpenAI, 2025). This impressive precision metric reduces unnecessary anxiety and procedures from false positives in the medical setting.
- Recall: 95.35% - Indicating that of all malignant cases, 95.35% of them were correctly identified (OpenAI, 2025). To ensure no critical cases are missed, a high recall is critical.
- F1 Score: 95.35% - The F1 score represents a harmonic mean of precision and recall (Muller & Guido, 2016). Offering a balanced view of the model, especially when false positives and false negatives are critical (OpenAI, 2025).
- AUC Score: 0.9971 – This score represents how well the model distinguishes between classes across all thresholds (Muller & Guido, 2016). The score reflects a near-perfect separation, suggesting exceptional robustness and flexibility (OpenAI, 2025).

Threshold Consideration:

These results are calculated using a default decision threshold of 0.5. Meaning that any instance with a predicted probability above 50% is classified as malignant (OpenAI, 2025). In clinical deployment, this threshold can be adjusted to prioritise sensitivity (recall) or specificity, depending on the use case, for example, screening vs diagnostic (OpenAI, 2025).

Confusion Matrix:

- True Positives: 41
- True Negatives: 69
- False Positives: 2
- False Negatives: 2

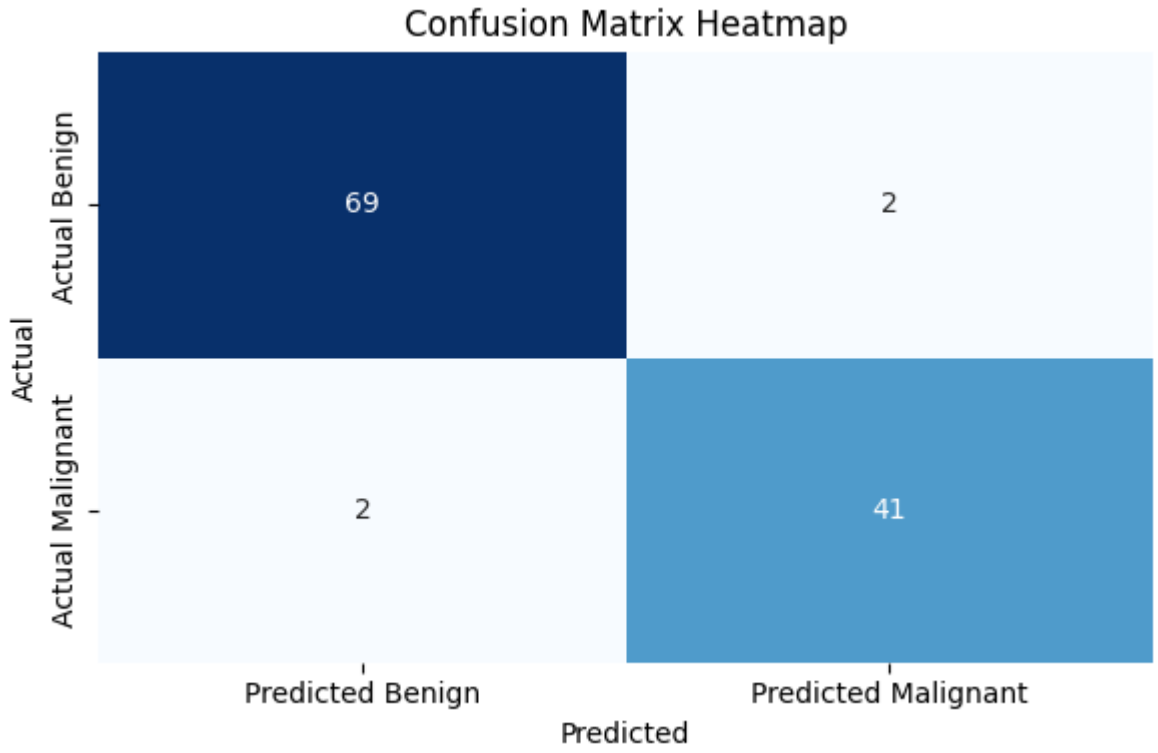

Figure 4: Confusion Matrix Heatmap – Initial Model

The confusion matrix seen in Figure 4 demonstrates excellent balance between sensitivity and specificity (OpenAI, 2025). With only two false positives and two false negatives recorded, there is a clear indication of a strong ability to distinguish between malignant and benign cases, whilst showing near symmetry in prediction accuracy (OpenAI, 2025).

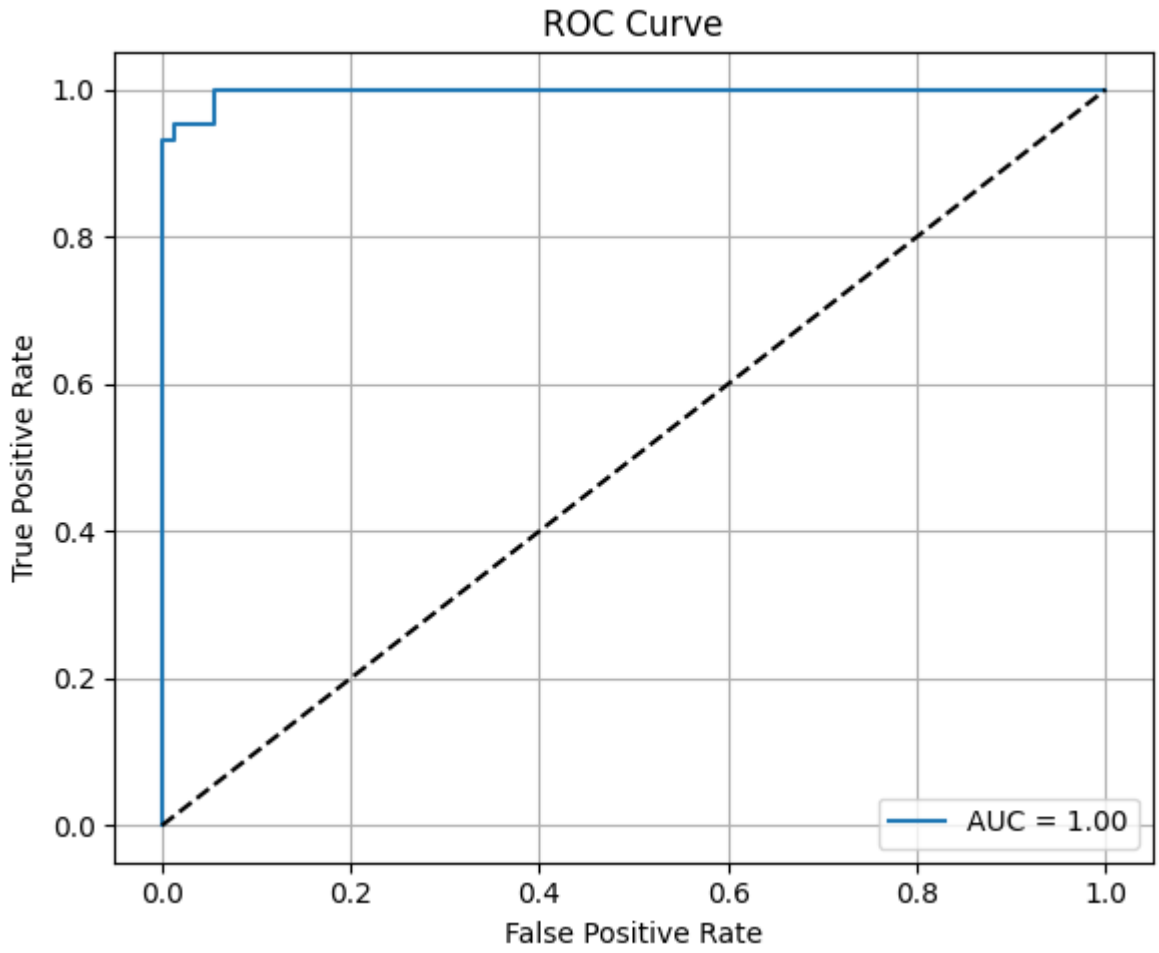

Figure 5: ROC Curve – Initial Model

The ROC curve seen in Figure 5 shows near-perfect separation between classes. With a steep rise in the curve and an AUC of 0.9971, high sensitivity and specificity are demonstrated, this being the true positive rate and true negative rate, respectively (OpenAI, 2025). An AUC score of 0.9971 indicates that across all possible classification thresholds, the model maintains strong discriminative power (Muller & Guido, 2016). This offers clinicians flexibility to adjust sensitivity and specificity based on use case, for example, prioritising recall in early detection settings (OpenAI, 2025). Confirming the model's suitability for real-world clinical screening support.

Model Robustness and Limitations

Although the model shows exceptional results on the test set, it must be noted that these results were achieved on a carefully curated noise noise-free academic dataset. In the context of deployment in a real-world environment, performance may be impacted by variability in patient data, imaging methods, or missing values (OpenAI, 2025). To ensure generalisability, additional validation would be required (Muller & Guido, 2016).

Conclusion

Considering the combination of these results, it can be said that the model consistently identified malignant cases without over-classifying benign ones. An element particularly critical in clinical applications, where false positives may lead to unnecessary stress and procedures, and false negatives could delay life-saving interventions (OpenAI, 2025).

Model Retraining

Despite the strong results obtained from the initial model, model retraining was conducted with an adjusted regularisation parameter value of 0.1 from the initial 1.0 to meet two objectives: To test the model's robustness and satisfy the evaluation protocol. With an increased regularisation parameter of 0.1 in the second model, an assessment could be made to see whether a simpler decision boundary could retain high performance (OpenAI, 2025). This form of penalisation aids in detecting and reducing overfitting, especially in high-dimensional data (Muller & Guido, 2016). To validate the stability and consistency of the model's performance under alternative configurations, project guidelines require retraining, regardless of the exceptional performance of the initial model (OpenAI, 2025). To ensure the retraining isolates the effect of the regularisation parameter adjustment, no changes were made to the feature set – the top 10 recursive feature elimination selected features were reused, outlined under Feature Selection of the Modelling Process section earlier in this report. Additionally, the same default threshold of 0.5 was used in the model retraining as was done in the initial model training. The same metrics and visuals, Receiver Operating Characteristic (ROC) curve, and confusion matrix heatmap were used to interpret the retrained model.

Retrained Model Performance (C=0.1)

- Accuracy: 88.6% - Although the accuracy value is relatively high, the score can be inflated from the high number of true negatives (benign cases) at 71. The retrained model performs well, but doesn't offer much insight into how well it can detect cancer specifically (OpenAI, 2025).
- Precision 100% - The retrained model never misclassified a benign case as malignant, resulting in zero false positives, which is key for healthcare, as unnecessary panic and testing for benign patients is avoided (OpenAI, 2025).
- Recall: 69.77% - This value portrays the model's sensitivity (Muller & Guido, 2016). The model missed over 30% of actual cancer cases with 13 false negatives, a statistic showing a serious limitation in a clinical setting.
- F1 Score: 82.19% - The F1 score represents a good balance between not missing positives (recall) and not falsely alarming (precision), but also shows that recall is the limiting factor in this retrain (OpenAI, 2025).
- AUC Score: 0.9574 – Even though the retrained model's AUC score is lower than the initial model's, this score of 0.9574 still shows excellent overall discrimination capability, even with a lower recall (Muller & Guido, 2016).

Confusion Matrix:

- True Positives: 30
- True Negatives: 71
- False Positives: 0
- False Negatives: 13

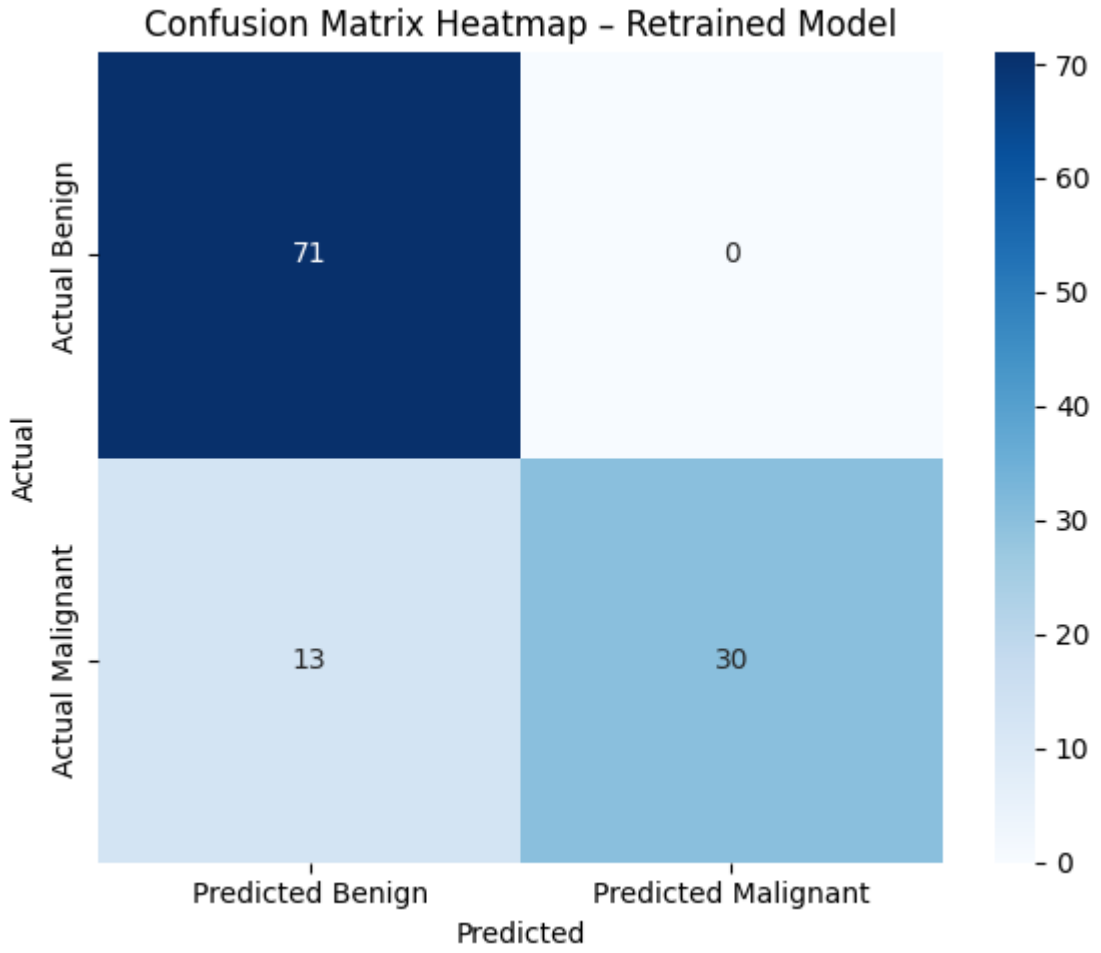

Figure 6: Confusion Matrix Heatmap – Retrained Model

From the heatmap seen in Figure 6, we can conclude that the retrained model perfectly identified all benign cases, giving zero false positives. Although the misclassification of 13 malignant cases as benign resulted in a drop in recall (OpenAI, 2025). It can be said that this is a characteristic of a more conservative model, as it is cautious of labelling a case as malignant, prioritising specificity (true negative rate) over sensitivity (true positive rate) (Muller & Guido, 2016).

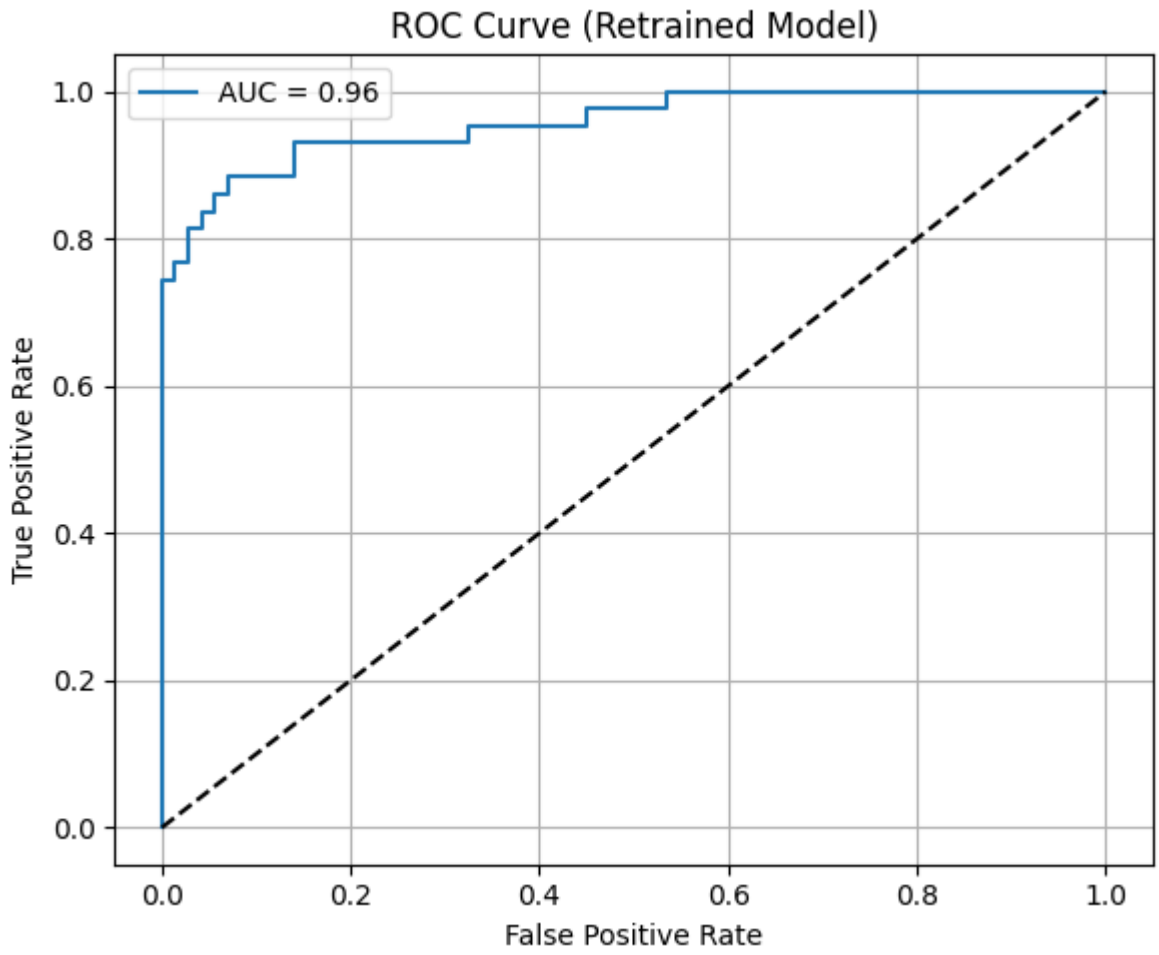

Figure 7: ROC Curve – Retrained Model

Despite the reduced recall metric in the retrained model, Figure 7 shows the AUC remains high at 0.9574, indicating strong overall discriminative power (Muller & Guido, 2016). The ROC curve maintains a steep incline, however, the area under the curve (AUC) is slightly reduced compared to the initial model's AUC of 0.9971. Showing a trade-off in performance (Muller & Guido, 2016).

Model Robustness and Limitations

Although the retrained model achieved perfect precision, it suffered from reduced recall (69.77%), indicating it failed to detect 13 malignant cases. This trade-off resulted from stronger regularisation (C=0.1), which simplified the model but led to underfitting (Muller & Guido, 2016). Whilst the default threshold of 0.5 was maintained, this contributed to missed cancer cases (OpenAI, 2025). In the clinical context, missed cancer diagnoses are a serious risk, although high specificity is considered desirable (OpenAI, 2025). Limitations in this model being suitable for high-stakes environments stem from the model's conservative nature.

Threshold adjustments may be needed for a more balanced and clinically acceptable performance (Schober & Vetter, 2021).

Conclusion

To conclude, valuable insights have been gathered from this model retraining, including a deeper understanding of the trade-offs between model complexity, precision, and recall. With a failed ability to identify multiple malignant cases, the clinical importance of recall in cancer detection scenarios is highlighted (OpenAI, 2025). Retraining showed that simplification may have reduced overfitting but hindered the diagnostic capabilities of the model (Schober & Vetter, 2021). Threshold tuning should be considered to enhance the alignment between model behaviour and real-world clinical risks (OpenAI, 2025).

Recommendations

Recommendations based on findings are as follows:

1. Deploy Initial Model (C=1.0) for Screening: The initial model was better suited as a clinical decision support tool for cancer screening. With a strong balance between sensitivity and specificity.
2. Avoid Over-regularisation: Given the performance results of the retrained model and its failure in identifying several malignant cases, it can be said that over-penalising complexity (C=0.1) should be avoided unless combined with threshold adjustments or ensemble methods (OpenAI, 2025).
3. Implement Threshold Tuning: To fine-tune sensitivity and specificity based on clinical risk tolerance, it is recommended that the model performance be evaluated across a range of probability thresholds (Schober & Vetter, 2021).
4. Incorporate Clinical Context: To align with the critical nature of cancer diagnoses, future versions should consider cost-sensitive learning, prioritising the minimisation of false negatives (OpenAI, 2025).
5. Continuous Model Monitoring: Integrate a monitoring pipeline to retrain the model periodically with new data, helping maintain relevance as diagnostic standards and patient demographics evolve (Schober & Vetter, 2021).

Disclosure of AI Use

Sections: Part 2.

Name of the tool used: ChatGPT4.

Purpose behind use: Outlines, summaries, Python analysis code, queries, evaluations, and suggestions.

Date used: 07/05/2025.

Link to chat: https://chatgpt.com/share/6800b44d-98b4-8004-a81b-7855cb0b5000

References

Ault, D. S. V., Liao, D. S. N. & Musolino, L., 2025. *Principles of Data Science.* Houston: OpenStax.

Learning, U. M. & Ovsen, 2016. *Kaggle - Breast Cancer Wisconsin (Diagnostic) Data Set.* [Online]
Available at: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
[Accessed 14 May 2025].

Muller, A. C. & Guido, S., 2016. *Introduction to Machine Learning with Python.* 1st ed. Sebastopol: O'Reilly Media.

OpenAI, 2025. *Open AI ChatGPT4.* [Online]
Available at: https://chatgpt.com/share/6800b44d-98b4-8004-a81b-7855cb0b5000
[Accessed 10 April 2025].

Schober, P. & Vetter, T. R., 2021. *Logistic Regression in Medical Research.* [Online]
Available at:
https://pmc.ncbi.nlm.nih.gov/articles/PMC7785709/#:~:text=The%20regression%20coefficients%20represent%20the,interpreted%20as%20an%20odds%20ratio.
[Accessed 14 May 2025].

Suzanne, 2023. *Data Pre-Processing for Linear Regression in Machine Learning.* [Online]
Available at: https://medium.com/@sds152/data-pre-processing-for-linear-regression-in-machine-learning-4b73ec48392a
[Accessed 17 April 2025].

Thrane, C., 2023. *The normality assumption in linear regression analysis — and why you most often can dispense with it.* [Online]
Available at: https://medium.com/@christerthrane/the-normality-assumption-in-linear-regression-analysis-and-why-you-most-often-can-dispense-with-5cedbedb1cf4
[Accessed 17 April 2025].

Vidhya, A., 2025. *Recursive Feature Elimination (RFE): Working, Advantages & Examples.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2023/05/recursive-feature-elimination/#:~:text=deal%20with%20multicollinearity.-,Avoid%20Overfitting%20or%20Underfitting,%2Drested%20and%20well%2Dfitted.
[Accessed 19 May 2025].

Wagavkar, S., 2024. *Introduction to the Correlation Matrix.* [Online]
Available at: https://builtin.com/data-science/correlation-matrix#:~:text=For%20example%2C%20let's%20say%20you,the%20relationship%20is%20positively%20strong.
[Accessed 20 April 2025].