**Business Analytics Capstone**

**BA723**

**Breast Cancer Predictive Modelling Documentation**

**By**

**Aisha Akter Anamika**

**301439588**

**Executive Summary**

**Executive introduction**

Breast cancer is not only a critical public health concern but also a significant driver of healthcare costs and operational strain worldwide. Despite advancements in diagnostic imaging and pathology, the challenge of accurately identifying malignant tumors, particularly avoiding false negatives persists, with direct consequences for patient survival rates and treatment efficiency.

This project presents a predictive modeling framework built on fine needle aspiration (FNA) image-derived features, designed to maximize recall for malignant cases. By applying and comparing multiple statistical and machine learning techniques including forward, backward, and stepwise regression, decision trees, random forest, and XGBoost. The research identifies the most effective approach for reliable tumor classification.

The implications of this work extend beyond academic research. A model of this nature can be integrated into existing medical imaging workflows, embedded into diagnostic software, or licensed to healthcare technology companies. Such integration could reduce diagnostic turnaround times, lower the incidence of missed malignancies, and contribute to significant cost savings for hospitals, insurers, and public health systems. Furthermore, by making the dataset and methodology accessible, this project invites collaboration, inspires further innovation, and fosters the development of more affordable and scalable diagnostic tools, bridging the gap between cutting-edge research and practical healthcare solutions.

**Executive Objective**

The primary objective of this research is to develop a predictive model capable of accurately classifying tumors as malignant or benign based on fine needle aspiration (FNA) imaging-derived features, with a clear priority on maximizing recall for malignant cases to minimize false negatives and reduce the risk of undetected cancers. The medical implication of this objective is critical: in oncology, even a single missed malignant diagnosis can have life-threatening consequences.

**Specific objectives include**

1. **Accessibility and Knowledge Sharing** – This project is intentionally designed to be readily accessible to individuals and organizations interested in machine learning, AI in healthcare, and medical diagnostics innovation. By making both the methodology and dataset available, I aim to encourage further exploration, refinement, and inspiration for more advanced predictive modeling. My broader goal is to raise awareness of the potential for AI to contribute to affordable and accessible healthcare solutions, particularly in diagnostics, so that individuals and communities even outside professional medical environments can benefit from more cost-effective tools for early disease detection.
2. **Model Evaluation and Continuous Improvement** – By comparing multiple machines learning models and arriving at a final recommended solution, I aim to present my findings to both technical and medical audiences for feedback. This iterative process will allow further refinement, with the ultimate ambition of achieving the highest possible diagnostic accuracy while using minimally processed data and reducing complexity or scope for error.
3. **Clinical Impact** – The goal of this project is to deliver a model that can reliably and efficiently identify malignant cases, providing healthcare professionals with an additional decision-support tool. While this current implementation is at a research stage, it lays the groundwork for more advanced systems that could be integrated into clinical practice to improve early detection rates and reduce diagnostic uncertainty.

**Executive Model Description**

To ensure robustness, multiple modeling approaches were developed, tested, and compared across statistical, tree-based, and ensemble learning paradigms:

1. **Forward Selection (Logistic Regression)** – Incrementally added features based on statistical significance until no further improvement was achieved.
2. **Backward Elimination (Logistic Regression)** – Started with all features and systematically removed those with the least statistical contribution, producing the lowest multicollinearity set.
3. **Stepwise Selection (Logistic Regression)** – Combined forward and backward strategies to create a compact feature set with strong predictive accuracy; ultimately selected as the final recommended model.
4. **Single Decision Tree** – Provided an interpretable, rule-based classification structure useful for visualizing decision boundaries.
5. **Random Forest** – Leveraged multiple decision trees for improved accuracy and robustness against overfitting.
6. **XGBoost** – Applied gradient boosting for high-performance classification, demonstrating strong recall for malignant cases.
7. **Comparison of Models** – All models were evaluated on accuracy, precision, recall, F1-score with special focus on recall for malignant predictions, given the medical priority of minimizing false negatives.

This comprehensive modeling strategy ensured that the final recommendation was based on empirical performance, interpretability, and operational feasibility rather than a single modeling philosophy.

**Executive Recommendations**

The final selected model for this project is the **Stepwise Regression Model**, chosen for its ability to maintain strong predictive performance while using a reduced set of predictors, resulting in a simpler, more efficient framework for implementation. Although other models, such as Random Forest and XGBoost, demonstrated competitive results, stepwise regression was ultimately preferred for its balance between accuracy, interpretability, and operational efficiency.

**Recommendations for future work include:**

1. **Exploring Organic Datasets** – This project was developed using a processed dataset in which certain correlated variables were removed to reduce multicollinearity. Future researchers may consider working with a more “organic” version of the dataset, preserving these correlated variables, and investigating ways to refine regression models so they remain stable and accurate even when multicollinearity is higher. This could expand the model’s ability to incorporate a wider range of predictors without sacrificing reliability.
2. **Further Evaluation of Tree-Based Models** – While stepwise regression was selected as the final model, both Random Forest and XGBoost showed promising results and robustness against correlated predictors. Future iterations of this project could explore tuning and refining these models further, potentially achieving even higher predictive accuracy and recall for malignant cases.
3. **Responsible Individual Use** – If adapted for broader use, individuals and community health programs could potentially leverage the model’s predictive reliability for preventive screening purposes. However, such usage must be approached with caution, ensuring the model is applied in conjunction with professional medical guidance to reduce the scope of error and avoid misinterpretation of results.

These recommendations aim to encourage both methodological refinement and responsible application, ensuring the continued advancement of AI-assisted breast cancer diagnostics while prioritizing accuracy, accessibility, and patient safety.

**Introduction**

Breast cancer is the most often diagnosed cancer among women aged 40 to 60. According to the World Health Organization there are 7.6 million deaths worldwide due to cancer each year, out of which 502,000 are caused by breast cancer alone. With such a high rate, breast cancer also is one of the deadliest cancers. For many years researchers have been trying to find the best way to cure breast cancer. To successfully cure a patient with breast cancer we need to diagnose it as early as possible. The most common diagnostic tools are mammography and a fine needle aspiration biopsy (FNA). This study aims to classify breast cancer lesions which have been obtained from fine needle aspiration (FNA) procedure using predictive modelling and machine learning algorithms. The stage of cancer depends on the malignancy factor that is assigned during an FNA examination. In this paper, we aim to build a classification model for predicting breast cancer cells and dive deeper into why and how.

**Problem statement** – Making cancer diagnosis accessible and affordable, starting at a primary level using a predictive mechanism which will detect breast cancer cells and can be further be built as a diagnostic tool.

**Key Objectives**

Improve early detection of malignant tumors to reduce the risk of late-stage cancer diagnosis.

Ensure the model remains clinically and statistically interpretable so healthcare professionals/public can trust and use it.

Make the solution lightweight and resource-efficient so it can be deployed in low-resource environments like small clinics or telehealth platforms.

**Measurements**

* Recall
* F1 score
* Precision
* VIF
* PSI
* Number of variables
* Performance Metrics for the Final Product can be considered such as Ease of Use / Time / Engagement Rate, User Confidence / Satisfaction Rate, rate of consistency with different data and systems etc.

**Assumptions**

* The dataset used is representative of real-world breast cancer cases.
* Users of the tool will have basic digital literacy and access to smartphones or computers.
* Predictive modeling, even if not perfect, can still assist in early self-assessment and trigger timely medical consultation.
* Clinical diagnosis will still be required for confirmation. This tool acts as a support system, not a replacement.

**Limitations**

* the dataset only contains the current state of the cancerous cell and does not include additional patient information such as genetic history, previous health records, or other relevant medical background.
* The model predicts benign vs. malignant only; it does not account for tumor grades, subtypes, or risk levels.
* The dataset is relatively small sample size and based on women in the U.S, which may affect generalizability.

**Data Sources**

Data Introduction - The dataset used is the Breast Cancer Wisconsin (Diagnostic), consisting of 569 observations and 32 features. From the total sample, 357 are benign tumors and 212 are malignant tumors. Each record corresponds to an image-derived sample and contains features such as radius, texture, perimeter, area, smoothness, concavity, symmetry, and fractal\_dimension, measured as mean, standard error (se), and worst-case values. The target variable diagnosis indicates malignancy (M = 1) or benignity (B = 0).

**Exclusions**

The “id” column was removed as it serves only as a unique identifier and carries no predictive value. Other than this no other columns were removed.

**Data dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Data Type | Description | Contribution to Malignancy |
| id | int64 | Unique identifier for each sample | –Irrelevant |
| diagnosis | object | Diagnosis of breast tissue (M = malignant, B = benign) | – (target variable) |
| radius\_mean | float64 | Mean of distances from center to points on the perimeter | ↓ Decreases malignancy |
| texture\_mean | float64 | Mean of gray-scale values (texture) | ↓ Decreases malignancy |
| perimeter\_mean | float64 | Mean size of the perimeter of the nucleus | ↑ Increases malignancy |
| area\_mean | float64 | Mean area of the cell nuclei | ↑ Increases malignancy |
| smoothness\_mean | float64 | Mean smoothness (local variation in radius lengths) | ↓ Decreases malignancy |
| compactness\_mean | float64 | Mean compactness (perimeter² / area - 1.0) | ↑ Increases malignancy |
| concavity\_mean | float64 | Mean concavity (severity of concave portions of the contour) | ↑ Increases malignancy |
| concave points\_mean | float64 | Mean number of concave portions of the contour | ↑ Increases malignancy |
| symmetry\_mean | float64 | Mean symmetry of the cell nucleus | ↑ Increases malignancy |
| fractal\_dimension\_mean | float64 | Mean fractal dimension (complexity of contour) | ↑ Increases malignancy |
| radius\_se | float64 | Standard error of radius | ↑ Increases malignancy |
| texture\_se | float64 | Standard error of texture | ↑ Increases malignancy |
| perimeter\_se | float64 | Standard error of perimeter | ↑ Increases malignancy |
| area\_se | float64 | Standard error of area | ↑ Increases malignancy |
| smoothness\_se | float64 | Standard error of smoothness | ↓ Decreases malignancy |
| compactness\_se | float64 | Standard error of compactness | ↑ Increases malignancy |
| concavity\_se | float64 | Standard error of concavity | ↑ Increases malignancy |
| concave points\_se | float64 | Standard error of concave points | ↑ Increases malignancy |
| symmetry\_se | float64 | Standard error of symmetry | ↓ Decreases malignancy |
| fractal\_dimension\_se | float64 | Standard error of fractal dimension | ↓ Decreases malignancy |
| radius\_worst | float64 | Worst (largest) value of radius | ↑ Increases malignancy |
| texture\_worst | float64 | Worst value of texture | ↑ Increases malignancy |
| perimeter\_worst | float64 | Worst value of perimeter | ↑ Increases malignancy |
| area\_worst | float64 | Worst value of area | ↑ Increases malignancy |
| smoothness\_worst | float64 | Worst value of smoothness | ↓ Decreases malignancy |
| compactness\_worst | float64 | Worst value of compactness | ↑ Increases malignancy |
| concavity\_worst | float64 | Worst value of concavity | ↑ Increases malignancy |
| concave points\_worst | float64 | Worst value of concave points | ↑ Increases malignancy |
| symmetry\_worst | float64 | Worst value of symmetry | ↑ Increases malignancy |
| fractal\_dimension\_worst | float64 | Worst value of fractal dimension | ↓ Decreases malignancy |

**Data Cleansing**

For data cleansing no rows were excluded due to missing or null values, as the dataset was found to be complete with no null entries.

**Data exploration**

The dataset was examined thoroughly by checking the number of rows, summary statistics, and diagnosis distribution, along with the data types of each column. This initial exploration helped to understand the overall structure of the dataset and identify any potential irregularities that might require attention during preprocessing. The target variable was converted from categorical to numerical to assign numbers 0 and 1 to 'B' and 'M', respectively. A descriptive Summary statistic (mean, median, standard deviation) was generated to understand the distribution and spread of the features. A box plot, histogram, and heat map were then created to examine the correlation between variables, identify which features stood out, and assess the extent of their influence. This visualization provided valuable insights into how each of these key variables should be addressed during the preprocessing phase.

**Histogram for feature distribution**

A group of blue and white graphs

AI-generated content may be incorrect.

**Boxplot to detect outliers**

A graph with lines and dots

AI-generated content may be incorrect.

**Data Exploration Techniques**

From our observation of the data, there were many key issues that needed to be solved to prepare the data before model development. For which we applied various data exploration techniques as discussed below-

**Data Preprocessing and feature engineering:**

1. Firstly, the outliers. Outliers are extreme values that differ significantly from many data points. In machine learning and statistical modelling, they can have negative impacts such as pulling predictions toward extreme values, especially in linear models like logistic regression. That’s why if outliers were to be present in the dataset, they need to be dealt with first. In this dataset **perimeter\_mean, area\_mean, area\_se, perimeter\_worst, area\_worst** were some notable outliers. Simply removing the outliers is a risky choice because sometimes they contain useful insights/values, so capping and flooring method was used to cap the extreme values at 99th/1st percentile. The reason why cap and floor were done first was simply to understand how much the outliers were impacting the data such as skews, correlation etc. The core idea was to deal with the outliers first and then look at the rest. After capping and flooring the outliers, there were certain changes that were noticed in the dataset. For example, many of the numerical features were found to be highly skewed. Skewed distributions can bias models like regression, which perform better when features resemble a normal distribution. To address this, we planned to apply log transformation, so the values are more evenly distributed. However, after capping and flooring, the skews of the data significantly stabilized (between 0-1, with only one exception of 2.03). As a result, log transformation was not performed.

A screenshot of a computer

AI-generated content may be incorrect. Before capping and flooring

A screen shot of a diagram

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect. After capping and flooring

Highly Correlated pairs

A screenshot of a computer

AI-generated content may be incorrect.

A colorful grid with red and blue squares

AI-generated content may be incorrect.

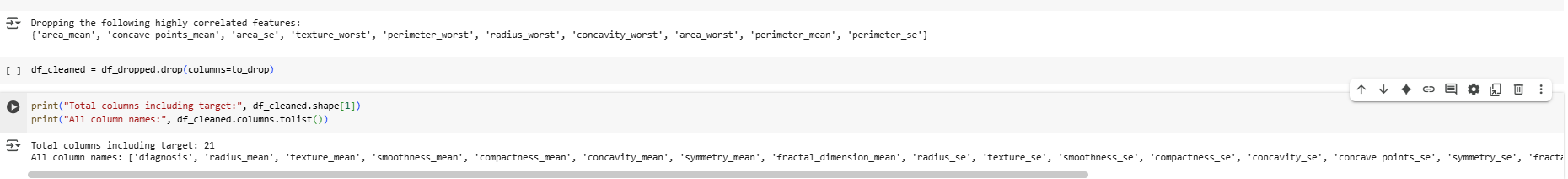
An important part of the dataset was to observe the correlation. Using the correlation heatmap, it was identified how correlated were the variables to each other and which variables may cause multicollinearity for the model. Although the initial plan was to not drop any variables since out of 3 models that was chosen, only one of them could not deal with correlated variables and the intention was to keep the dataset intact. However, while running forward, backward and stepwise regression; backward regression could not be completed due to redundant variables. Hence, certain variables had to be dropped. The decision to drop certain highly correlated variables in the dataset was based on two key reasons. First, it was unable to run the backward elimination model successfully without addressing multicollinearity. Second, many of the variables were not only highly correlated but also conveyed very similar information, making them redundant.

For example, variables like area\_mean, area\_se, and area\_worst were all highly correlated with one another, as well as with other features such as perimeter and radius. All these essentially represent the size of the cell. To avoid redundancy and multicollinearity, it was decided to retain only one representative variable from each group of similar features. In this case, radius (e.g., radius\_mean) was chosen to keep as it effectively captures the size-related information and dropped area and perimeter variables. Similarly, key variables from each unique category that represented different geometric or textural properties of the cell were retained. For example:

Texture (e.g., texture\_mean) reflects the smoothness variations on the cell surface. Smoothness, concavity, and compactness each represent distinct biological characteristics of the tumor cells, even though they may appear similar at a surface level. It was crucial to keep variables from each of these unique categories because they capture different dimensions of how tumor cells behave and change. These nuances are crucial for accurately classifying tumors as benign or malignant.

Finally, each feature's correlation with the target variable was also examined. This helped guide the final decision on which variables to retain or remove. Features that had both high relevance to the target and low redundancy with others were prioritized, ensuring a more efficient and interpretable model. After dropping correlated variables, the dataset was left with 21 features including the target variable from 31 features. Since apart from the regression models the other models were correlation friendly, this cleaning of variables was performed on a copy of the main dataset and used this version to only perform the regression model. It was named df\_cleaned. The purpose of this is to make sure that the tree models are being run on the raw/original dataset and the integrity of the model remains same.

**Final dataset overview**



**A diagram of different colors

AI-generated content may be incorrect.Derived Variables Explanation**

For the derived variable and as part of the feature engineering phase of this project, the potential for creating a derived variable was explored, which is basically a new feature formed by combining or transforming existing features to uncover deeper insights. Derived variables can sometimes help the model capture more complex patterns in the dataset that can potentially identify malignant cases better, or sometimes they can bring out useful insights to understand what factors are causing tumors to be identified as benign or malignant. So, to further explore, attempts were taken to create a derived variable and came up with two variables. One of them was radius growth rate, which was designed to measure the growth of the tumor's radius across different stages. This would basically track how rapidly the structure of a cell is expanding, giving us a clue about whether the tumor is behaving abnormally, whether it is growing larger at a normal rate or abnormally fast. That would indicate that the larger it gets after a certain point, it's malignant. Similarly, it was also considered incorporating convexity as another geometric indicator. In medical imaging, convexity refers to how closely the shape of a tumor resembles a convex form, meaning one with no inward curve. A perfectly convex shape is smooth, outward facing, and regular. Tumors, on the other hand, that have a lot of concave regions like a lot of inward curves that look more structurally complex, and they are possibly malignant. So, the simpler and smoother the shape is, the more likely it is to be benign, and the more complex and inward facing it is, the more likely it is to be malignant. High convexity is generally a sign of a more regular, benign structure, but lower convexity, which has more indentations, could suggest malignancy. Both variables were created to gain insights from them, but it was only possible to proceed with one. While trying to incorporate convexity, it turned out to be very redundant and was creating multicollinearity with concavity, and it just wasn't performing well in the model. As a result, the decision was to only move forward with radius growth rate. Radius growth was incorporated rate into the model, in the dataset, which had gone through some EDA. For example, it did go through some outlier fixations like capping and flooring, some transformations, and some variables were dropped. When the derived variable was incorporated, it did not show much contribution to the dataset. As it can be seen from the image that “radius\_growth\_rate’’ which was the derived variable has only a correlation score of 0.31 with the target diagnosis. **0.31 is not very** strong, it suggests that the variable has some predictive value, but on its own, it is not a powerful predictor. Similarly, as a variable, it didn’t provide any new insights that we hadn’t already seen; instead, it was creating redundancy, so it was decided to proceed without it.

Lastly, no Undersampling/Oversampling or SMOTE was done because in medical dataset it’s important to preserve the integrity and avoid unnecessary synthetic noise.

**Modelling**

Before looking at the modeling and the results and trying to understand which model performed the best, it’s essential to understand some of the key result indicators that will be prioritized, which are called, in this case, Precision, Recall, and F1 Score for the Decision Tree, Random Forest, and XGBoost model. And for my Forward, Backward, and Stepwise models, along with these scores, some other factors also need to be looked at such as odds ratio, p-value, R-square scores, and for overall all of the models, there will also be observations to understand which variables contributed the most—basically the top five variables or features that contributed the most to the results. So, before getting into the results, let’s see what each of these things means. Top variables are basically the highlight of which variables are contributing the most and performing the most for each of the models. For example, in Forward model, out of the 20 variables and one target, the model might work with 10 or 15, whereas Backward model might only take 8, or Stepwise model might only pick out 10 variables to predict which cases are malignant and which cases are benign. The other models, which are Random Forest, XGBoost, and Decision Tree, are going to work with all of the variables, and they're going to determine out of all of these variables which are powerful and which ones shine through to help the case better and associate a ranking or number, a numeric value, for each of them, which is supported justification. The explanation of all these model indicators and performance result metrics is provided down below.

**Key Evaluation Metrics in Classification**

1. Recall (Sensitivity / True Positive Rate)

What it means: Out of all the actual malignant cases, how many did the model correctly identify?

**Formula:**

Recall= True Positive / True positive + False negative

**Why it matters**:  
The model is predicting whether a tumor is malignant (1). A false negative (missed malignant case) is very dangerous, as it could mean a patient doesn’t get timely treatment.  
So, maximizing recall ensures the model catches as many malignant cases as possible, even if it occasionally flags a benign tumor as malignant.

Precision

What it means: Out of all the cases the model predicted as malignant, how many were malignant?

**Formula:**

Precision= True Positive / True positive + False Positive

**Why it matters:**Precision is about avoiding false alarms (false positives). A low precision means the model raises too many red flags (benign cases wrongly flagged as malignant), which could lead to unnecessary tests or stress.  
But in healthcare, false positives are less dangerous than false negatives, which is why recall over precision is prioritized.

F1 Score (Harmonic Mean of Precision and Recall)

What it means: A balanced metric that combines both precision and recall into a single number.

**Formula:**

F1= 2× (Precision \* Recall) / (Precision + Recall)​

**Why it matters:**  
F1 is helpful to balance both concerns, not missing malignant cases (recall) and not over-predicting them (precision). It’s especially useful when the classes are imbalanced (like benign vs malignant).

**Why Recall is most important?**

The goal is early detection of malignant tumors, which can be lifesaving.

Missing a malignant case (false negative) is far more dangerous than a false positive.

Therefore, high recall is critical, even if it slightly compromises precision**.**

After creating a dataset with the necessary variables dropped, at this stage, the dataset is fully prepared for machine learning. It includes 20 features, a target variable, and clean, consistent data split into training and test sets. When work began on the regression models, thorough research was conducted to determine which type of predictive modeling would be most effective for a medical dataset specifically for distinguishing between benign and malignant cancer cells. Since the goal was to build a binary classification model, logistic regression was initially one of the top choices. It is widely recognized for its interpretability and suitability for binary outcomes, making it a standard approach in healthcare analytics. However, after discussions with professors and after implementing logistic regression on the dataset, several challenges were encountered. One significant issue was that logistic regression required extensive data transformation, such as log transformations, normalization, and dealing with multicollinearity, to ensure that model assumptions were met. These preprocessing steps added complexity and time to the workflow. Moreover, even after these adjustments, the logistic regression model did not yield performance metrics especially accuracy and recall as strong as those produced by other regression approaches. In contrast, when forward, backward, and stepwise regression were performed, the process felt more efficient and better aligned with the dataset. These methods offered more flexibility in automated variable selection, allowing the model to iteratively identify the most relevant predictors based on statistical significance. This helped in reducing noise, avoiding overfitting, and improving interpretability. Additionally, these selection techniques allowed retention of important predictors while excluding redundant or less informative ones, leading to better model generalization. This is particularly critical in a medical context where the goal is to minimize false negatives and hence maximize recall a metric where the stepwise and backward models performed particularly well. Given these advantages, it was concluded that forward, backward, and stepwise regression were more effective for the breast cancer predictive modeling task than logistic regression. Therefore, these models were chosen for further analysis. The modeling process began with regression, with the initial plan being not to remove any of the variables. However, with all variables present in the dataset, backward regression could not be performed due to the problem of singular metrics and multicollinearity. As a result, some variables had to be dropped before running the models.

Forward, backward, and stepwise regression were performed on the dataset, which had been capped and floored and had some of the correlated variables removed. The dataset was split into training and test sets using a 70-30 split, where 30% was allocated as the test set. Each model was trained on the training dataset and then evaluated on the test dataset.

Looking at the results, the forward regression performed comparatively lower than both backward and stepwise regression. Given that the dataset is medical in nature, the most important evaluation metric is recall. In this context, minimizing false negatives is key meaning it is more acceptable for the model to incorrectly flag someone as having cancer (false positive) than to miss someone who has cancer (false negative). More specifically, if the model mistakenly classifies a benign cell as malignant (precision), it is not as critical as the model missing a malignant cell and classifying it as benign (recall). Therefore, recall is the top priority, followed by F1 score, and lastly, precision.

When comparing the recall scores for malignant cases on the test set:

Forward selection: 97%

Backward elimination: 97%

Stepwise regression: 97%

So, in terms of recall, all three models perform equally well.

Next, considering the F1 score for malignant cases:

Forward: 95%

Backward: 95%

Stepwise: 95%

Again, all three models are equal on this metric.

When looking at precision, some differences were found:

Forward: 92%

Backward: 94%

Stepwise: 94%

This indicates that backward and stepwise regression are performing slightly better than forward selection in terms of precision. Given these results, forward selection appears to be the weakest of the three. Between backward and stepwise regression—the top two models—the performance is nearly identical across recall, F1, and precision. Since the goal is to narrow it down to one model, the next step is to proceed with developing other models and compare all their performances comprehensively. A final model will be chosen based on that full comparison.

A screenshot of a computer

AI-generated content may be incorrect.**Forward Selection Result**

**A screenshot of a computer

AI-generated content may be incorrect.**

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.**Backward Selection Result**

**A screenshot of a computer

AI-generated content may be incorrect.**A screenshot of a computer

AI-generated content may be incorrect.**Stepwise Selection Result**

For regression models, it's important to see which variables are coming off as important from the dataset, and what some of the model results are such as the odds ratios, how many variables the model is using, the p-values for those variables, and so on. All these details are attached in the screenshot given above. As it can be seen, there are some top features across the regression models. For forward selection, it's concavity mean, concave points worst, and radius mean. At the same time, for stepwise regression, it's concave points worst, radius, and texture mean. For the backward model, it's radius mean, texture mean, concave points worst, and symmetry mean. So, it’s quite clear how these models are picking out different variables to make their predictions and justify their results. From that, it can be interpreted how each model affects and determines the possibility of a cancer cell being malignant. After variables, comes the numbers and the first thing to observe is the odds ratio. Now a question can be: what is the role that odds ratio is playing here? So, in logistic regression or the regression models that were done here, the log odds of an event, in this case, it's the cancer being malignant, which is equal to one. Each odds ratio shows how a one-unit increase in the predictor variable affects the odds of the outcome being malignant. So, for example, an extremely large odds ratio such as concave points worst means even a small increase in this feature is associated with a huge increase in the odds of malignancy. So, if a variable like concave points worst has an extremely high odds ratio, it may genuinely be one of the strongest indicators of malignancy in the dataset. So more concave points typically signal irregular, aggressive tumor growth, which is a sign of malignancy. The same goes with an extremely small odds ratio as well. The smaller the odds ratio is, the less the odds of malignancy get. So, for example, concave points SE has a very extremely small odds ratio. That means it is associated with an extreme reduction in the odds of malignancy, that this variable doesn't indicate or has very low contribution to identifying malignant cells. So, looking at the odds ratios of the mentioned models, Stepwise Regression’s almost all variables are statistically significant, and the odds ratios are the most stable compared to the other models. In the Forward model, the odds ratios are a bit more extreme, and in the Backward model, they are relatively more stable, but still not as consistent as in the stepwise model. On the other hand, the role of the p-value is that for each variable, it tests the null hypothesis that the coefficient; meaning the impact of the variable is equal to zero, meaning the variable has no effect on predicting malignancy. So, if the p-value is lower than 0.05, the lower it gets, the stronger it gets, which means if the p-value is lower than 0.05, the variable significantly contributes to predicting malignancy. But on the other hand, if the p-value of a variable is high, which is more than 0.05, the evidence is that it's a weak variable, because it's not meaningful enough and it probably even can be removed. So now, comparing Forward, Backward, and Stepwise regression models, the Stepwise regression had five variables in the model, and they all have p-values well below 0.01, which means that almost every predictor is statistically significant. It also shows that the model has low multicollinearity, is highly interpretable, and achieves a good balance of simplicity and significance, because it works with very few variables but still delivers very good results. At the same time, Forward Selection includes 10 predictors, but a lot of them have high p-values, and some variables show a risk of overfitting. As for Backward model, it used nine predictors, and while most p-values were below 0.05, some were close to the threshold.

**VIF for Forward Selection Model**

A screenshot of a computer program

AI-generated content may be incorrect.

**A screenshot of a computer

AI-generated content may be incorrect.VIF for Backward Selection Model**

**A screenshot of a computer code

AI-generated content may be incorrect.VIF for Stepwise Regression Model**

Another factor in result with the models is VIF which is Variance Inflation Factor (VIF). VIF was calculated for each model’s final feature set. VIF measures the extent to which a predictor is correlated with the other predictors in the model, with values above 5 indicating moderate to high multicollinearity and values above 10 suggesting severe collinearity. The forward selection model showed the highest degree of multicollinearity, particularly among shape-related variables such as *concavity\_mean* (VIF = 9.80), *concave\_points\_worst* (VIF = 8.25), and *concavity\_se* (VIF = 7.23). The stepwise model displayed a similar pattern, retaining high-VIF variables, which suggests that while these predictors are highly correlated, they may each contribute uniquely to predictive performance. In contrast, the backward elimination model resulted in the lowest overall VIF values, with the highest being *concave\_points\_worst* (VIF = 5.88) and *compactness\_se* (VIF = 5.67). This indicates that backward elimination naturally filtered out more redundant variables, producing a statistically more stable set of predictors. Overall, these findings highlight a trade-off between interpretability and predictive accuracy. The primary objective of this project is accurate and efficient prediction of malignant cases rather than coefficient interpretability which is why all factors considered between the three of them, the Backward and Stepwise models performed much better than forward, when looking at a more detailed view of the result set. That said, between Backward and Stepwise, it's always more practical to go with a model that is simpler, less complicated, and highly interpretable. So, from that perspective Stepwise regression model performed the best among the three.

**Random Forest Model Explanation**

The next model in the project was Random Forest because it is one of the models that performs well with a complete set of variables and does not require extensive preprocessing. Since the dataset is medical data, the aim was to keep it as intact as possible and obtain more organic results from the predictive modeling. Therefore, Random Forest was a natural second choice.

To preserve the integrity of the original dataset, Random Forest was performed on a copy of the dataset rather than modifying the original. The same train-test split of 70% training and 30% testing used for the regression models was applied. After training the model, its performance was tested on the test set.

For the Random Forest model:

* The recall score for malignant cases was 94%.
* The F1-score was 96%.
* The precision for malignant cases was 98%.

Overall, the model performed quite well. The confusion matrix results were also satisfactory.

Next, the top 10 important features identified by the Random Forest model were reviewed and compared with the predictors selected by previous models — forward, backward, and stepwise regression — to check for consistency in the most influential variables. Several of the top features overlapped, indicating some level of agreement across models. Additionally, the feature importance scores for the Random Forest model were quite reasonable, further supporting its reliability.

A screenshot of a computer

AI-generated content may be incorrect.**Random Forest Test Results**

A screenshot of a computer

AI-generated content may be incorrect.**Top 10 most important features for Random Forest**

From the screenshot, it can be seen that some of the features are ranked among the top 10 most important, based on their feature importance scores. These scores reflect how much each variable contributes to the prediction accuracy of the model, the higher the score, the more influential the feature. For example, concave points\_mean has an importance score of 0.1419, making it the highest and most important predictor. This variable measures the average number of concave portions of the tumor contour. Higher values for this variable are strongly associated with malignant tumors, meaning that the greater the average concave points observed, the more concerning the tumor is likely to be.

Similarly, area\_worst has a score of 0.1182, representing the largest area observed in the tumor. Malignant tumors generally have larger areas, so an increase in this value indicates that tumors are growing and are more likely to be malignant. These examples show how each score aligns with the variable’s role in the prediction.

A key takeaway from these results is that tumor shape and size, particularly in their most extreme forms are critical in predicting malignancy. The model highlights these morphological characteristics as highly influential in differentiating between benign and malignant cases.

**Decision Tree Explanation**

The next model in the project was the Decision Tree. A copy of the original dataset was used because Decision Tree models perform well on raw data and do not require extensive preprocessing. The same 70-30 train-test split used in previous models was maintained for consistency.

After training the model, it was evaluated on the test set. The Decision Tree delivered strong results, with a recall of 95% and an F1-score of 92% for the malignant class. The overall accuracy was 93%. The tree depth was 7 with 16 leaves, meaning the tree concluded in 16 final classification points, making it relatively interpretable and not overly complex in structure.

To better understand its decision-making process, the tree splits were visualized. This binary classification tree represents:

* Class 1 (Malignant) in blue
* Class 0 (Benign) in orange

Each box (node) in the tree makes a decision based on a specific feature and threshold. For example, at the root node, the split occurs on concave points\_mean ≤ 0.051. From top to bottom, each path represents a sequence of decisions leading to a final classification:

* The leftmost path predicts benign cases.
* The rightmost path predicts malignant cases.

Each node also shows a Gini index, which measures purity:

* Gini = 0 → all samples in the node belong to a single class (pure)
* Gini = 0.5 → samples are evenly split between classes (impure)

In the visualization, darker shades of orange or blue indicate more confident and purer predictions.

From the analysis, concave points\_mean emerged as the most important root feature. As the tree branches deeper, variables like radius\_worst, texture\_worst, and area\_worst further refine the classifications. Some leaf nodes had a Gini index of 0, indicating perfectly pure classifications containing only malignant or only benign cases.

**Decision Tree Test Result**

A screenshot of a graph

AI-generated content may be incorrect.A screenshot of a computer code

AI-generated content may be incorrect.

A diagram of a company

AI-generated content may be incorrect.

**XGBoost Explanation**

For the final model, XGBoost was selected based on recommendations from both peers and professors, who suggested it as a powerful classifier worth exploring. XGBoost was applied to the dataset where only the outliers had been capped and floored; no transformations or feature drops were performed beyond that. The same 70-30 train-test split used in previous models was maintained to ensure consistency. Upon evaluation, XGBoost achieved a recall score of 94% for malignant cases, along with an impressive F1-score of 97%. These results were closely aligned with the performance of earlier models, demonstrating comparable predictive strength. When examining feature importance, many of the top predictors identified by XGBoost overlapped with those from other models. This overlap reinforced the consistency and reliability of certain variables across different modeling techniques, further validating their significance in predicting malignancy.

**A screenshot of a computer

AI-generated content may be incorrect.XGBoost Test Result**

A screenshot of a computer program

AI-generated content may be incorrect. **XGBoost Top 10 variables**

**17.0 Model Comparison**

Based on the comparison table of all the models, it is evident that the Forward, Backward, and Stepwise regression models performed the best in terms of recall and F1 score, which are the top two priorities for this project. Precision is also a key metric, but secondary to recall and F1 in this case. From this perspective, the regression models outperformed the Decision Tree, Random Forest, and XGBoost models.

Among the three regression models, the Stepwise model is the most suitable choice. It achieved results comparable to the Backward model in terms of recall and F1 score, while being simpler and more interpretable. The Forward model did not perform as well due to several factors, including the inclusion of variables with weaker statistical significance.

The regression models were trained on a dataset that had undergone capping and flooring, with certain variables removed to address multicollinearity. In contrast, the Decision Tree and Random Forest models were applied to the original, unprocessed dataset, while the XGBoost model was run on the version that had only been capped and floored.

This raises an important consideration when working with medical datasets: should we prioritize results derived from an untouched, organic dataset, or aim for the highest possible recall, even if that involves preprocessing? The former approach supports transparency and raw data integrity, while the latter enhances predictive performance particularly in identifying malignant cases. For the purposes of this project, the primary goal is to maximize recall and correctly identify as many malignant cases as possible. Therefore, the Stepwise Regression model has been selected as the final model due to its strong performance, simplicity, and interpretability. After the model selection the top five features across all the models have also been attached below. This is basically a breakdown of which features are prominent when it comes to identifying how malignant a cancerous cell is. As it can be seen, regression models all have the same variables as their top features, except Stepwise includes concave\_points\_se, which is another interesting finding, highlighting why Stepwise specifically selected concave\_points\_se. concave\_points\_se was likely chosen by stepwise regression because it provided statistically significant, non-redundant information when considered alongside the other selected features, even though its standalone predictive strength wasn’t enough for tree-based models to rank it in their top features.

**Correlation and VIF score for concave\_points\_se to show its importance-**

A screenshot of a computer program

AI-generated content may be incorrect.

The Decision Tree, Random Forest, and XGBoost models also have many variables in co common with each other. Most of these variables revolve around concavity, texture, and aspects related to the size of the cells, such as area or perimeter.

From this list, it can be identified that size, texture, and concavity are three very important factors when it comes to identifying benign and malignant cells.

A screenshot of a computer

AI-generated content may be incorrect.**Scores for all the models -**

A close-up of a white background

AI-generated content may be incorrect.**Top 5 features across all models -**

**18.0 Model Selection**

After evaluating multiple models, logistic regression, forward selection, backward elimination, decision trees, random forest, and XGBoost. The stepwise regression model was chosen as the final model for this project as discussed above.

**19.0 Model Theory**

Stepwise regression is a variable selection technique used in regression modeling to identify a subset of predictors that contribute most to the model. It systematically adds and removes features based on statistical significance (commonly using p-values) and model fit criteria such as AIC or BIC. In this case, regression was used due to the binary nature of the target variable (malignant vs. benign tumor), and stepwise selection helped optimize variable inclusion while avoiding overfitting.

**20.0 Model Assumptions and Limitations**

**Assumptions:**

* **Independence of observations**: Each data point is assumed to be independent.
* **Linearity of logit**: There must be a linear relationship between the log odds of the outcome and the continuous predictors.
* **No multicollinearity**: Strong correlations between predictors can distort the model’s accuracy and interpretability.
* **Large sample size**: Logistic regression performs best with sufficient data, especially when predictors are numerous.

**Limitations:**

* The model output suggests possible quasi-complete separation (noted in the summary). This means a portion of the data can be predicted perfectly, which might indicate overfitting or high-class separability. Some coefficients may be unstable or have inflated standard errors.
* Stepwise methods can be sensitive to data variations and may exclude variables that could be important in other contexts or samples.
* The model is based solely on statistical fit and does not incorporate domain knowledge unless manually guided.

**20.0 Model Sensitivity to Key Drivers**

The most influential predictors based on coefficient size and p-values include:

* **concave\_points\_worst** (coef = 92.24, p < 0.001)
* **radius\_se** (coef = 13.64, p < 0.001)
* **texture\_mean** (coef = 0.41, p < 0.001)
* **concave\_points\_se** (negative coef, -340.18, p < 0.001)
* **radius\_mean** (coef = 0.66, p = 0.001)

These variables capture key aspects of cell shape irregularity, size, and texture, which are biologically relevant in identifying malignant cells. The strong statistical significance of all selected features suggests that the model is highly sensitive to these drivers, and any change in their distributions may affect the model’s predictions.

**21.0 Additional Models to Address Business Objectives**

While the stepwise logistic regression provides strong performance and interpretability, other models could support different business goals:

* **Random Forest or XGBoost**: These models are more robust to outliers and can handle non-linear relationships, which is useful in production settings or when model accuracy outweighs interpretability.
* **Decision Tree**: Useful for visualizing logic and explaining decisions to non-technical stakeholders (e.g., in a clinical or patient-facing context).

**22.0 Recommendations -**

The stepwise model plays a pivotal role in the overall development of this project because it directly supports the business objectives at hand. The stepwise logistic regression model achieves a high recall score, which is a critical priority in this project. It is also simple and highly interpretable, using a minimal yet effective set of variables that contribute meaningfully without being redundant or confusing. Additionally, the model is efficient in terms of resource usage it does not require intensive computing power and can be deployed in low-resource environments, such as small clinics or telehealth platforms.

Future development of the model should begin with comprehensive validation using completely unseen and minimally processed (“organic”) datasets to ensure generalizability and consistent alignment with clinical requirements. Performance should be benchmarked against other high-performing models, such as Random Forest and XGBoost, to confirm reliability and identify opportunities for integration or ensemble approaches. Following validation, the model can be piloted in small-scale, controlled environments such as research centers to be tested by business analysts or software developers/engineers. Insights from these pilots should guide iterative improvements which can refine both technical performance and clinical relevance. Once optimized, the model can be developed as a product such as Mobile App, Web/Cloud API, and Hospital System Integration. For deployment as a final product, the focus should be on low-cost, scalable implementation in accessible screening programs within small clinics, community health centers, institutions focused on women’s health, particularly breast cancer and research institutions to improve early detection and reduce diagnostic delays. Continuous user feedback from healthcare providers and patients should be used to monitor performance and identify areas for enhancement. Feedback from professionals such as oncologists and radiologists would then help refine and validate the model. After the product has been finalized and used in small scale and given enough time to go through ample number of screenings and generates consistently accurate results it can be open for adaptation. Collaborative partnerships with medical institutions, public health agencies, and research organizations can drive scalability, expand access to advanced medical technology for further refinements, and ensure the product aligns with public health priorities.

The more accurately the model performs, the more cases will be detected early, leading to better health outcomes for women facing breast cancer. Throughout this journey, the model can be open-sourced through a web-based diagnostic simulation tool, allowing public interaction for educational and awareness purposes. It can also be used as a research and development tool, encouraging collaboration among tech companies, non-profits, and government health departments. Ultimately, the project has the potential to bridge the gap between technology and preventive care, bringing impactful AI solutions into the hands of those who need them most.

**References**

**The Guardian (2025, January 7).** More breast cancer cases found when AI used in screenings, study finds. *The Guardian*. Retrieved from <https://www.theguardian.com/society/2025/jan/07/more-breast-cancer-cases-found-when-ai-used-in-screenings-study-finds>

**Radiological Society of North America (RSNA). (2024, March 19).** Deep learning for predicting breast cancer [Press release]. RSNA. Retrieved from <https://www.rsna.org/news/2024/march/deep-learning-for-predicting-breast-cancer>

Yadav, N. (2021). *Breast Cancer Classification* [Kaggle notebook]. Kaggle. Retrieved from <https://www.kaggle.com/code/niteshyadav3103/breast-cancer-classification>