CANCER RISK CLASSIFICATION

Darius Dylan Govender

Table of Contents

Introduction

The goal of the project is to develop a machine learning classification model which is capable of predicting cancer risk based on both medical and lifestyle factors. This would allow for medical professionals to easily identify and address high risk patients. This project will make use of a Kaggle dataset (https://www.kaggle.com/datasets/rabieelkharoua/cancer-prediction-dataset/data) in order to build a well performing model. This model will make use of features such as 'Age', 'Gender', 'BMI', 'Smoking', 'GeneticRisk', 'PhysicalActivity', 'AlcoholIntake', and 'CancerHistory' to determine a patients Diagnosis (Cancer/No Cancer). This analysis is structured and will follows a process of data cleaning, exploratory analysis, modelling, evaluation and business interpretation.

Data Overview and Cleaning

This dataset is already pre-processed, and these steps will still be followed as a precaution. Data cleaning involves finding and removing missing, duplicate and/or irrelevant data (GeekForGeeks, 2025). Data cleaning aims to maintain data accuracy, consistence and is noise free (GeekForGeeks, 2025). This step is critical for ensuring data quality as well as well performing model. The dataset used was already pre-processed however this doesn't mean these steps were skipped.

To ensure data quality and readiness for modelling, the following cleaning steps were taken. First, all necessary libraries (numpy, pandas, matplotlib, seaborn, sklearn, scipy, statsmodels and imblearn) were imported. The second step was to preform an initial inspection of the data using multiple checks such as df.info (), df.nunique() and df.describe this showed the dataset consisted of 1500 rows and 8 columns. No missing values, no duplicate values and no outliers were found within the dataset. The dataset is already pre-processed, so there was no need to address this for feature encoding as it was already done.

Step	Action Performed	Outcome
Import	Loaded libraries: numpy, pandas, matplotlib, seaborn, sklearn, scipy, statsmodels and imblearn.	Data analysis environment setup.
Inspect	Ran df.info(), df.nunique() and df.decribe.	1500 rows and 8 columns
Missing values	df.isna().sum() and df.isnull().sum()	No missing values
Duplicate values	df.duplicated().sum()	No duplicate values
Outliers	Made use of IQR rule	No outliers
Encode	df.head() – to check if feature encoding needed	No feature encoding needed
Target Balance	The dataset is balanced since the split is within the range of the imbalance threshold of 70%/30%.	No action taken but attempted both SMOTE and class_weight='balanced' to see if there is any value offered below in modelling.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an important step in the data analysis process, in this process the aim is to understand patterns and find relationships by using statistical and visual tools (GeekForGeeks, 2025).

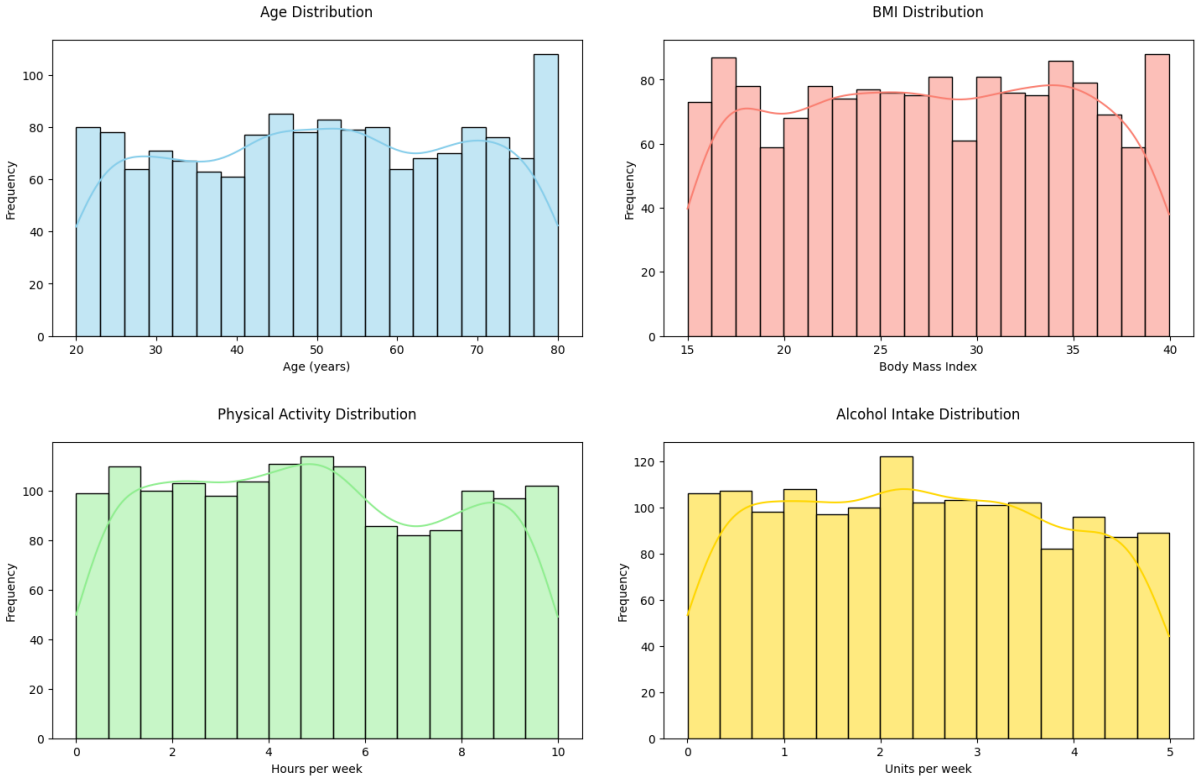
In this section it was broken down into the Univariate analysis, Bivariate analysis and the Feature selection.

- Univariate Analysis focuses on individual variables (GeekForGeeks, 2025).
- Bivariate Analysis focuses on feature relationships (GeekForGeeks, 2025).
- Feature Selection looks to determine which features are the most valuable for training the models.

Univariate Analysis

For Univariate Analysis of the features, histograms (for numerical features) and count plots (for categorical features) were made use of to understand feature distribution and identify trends.

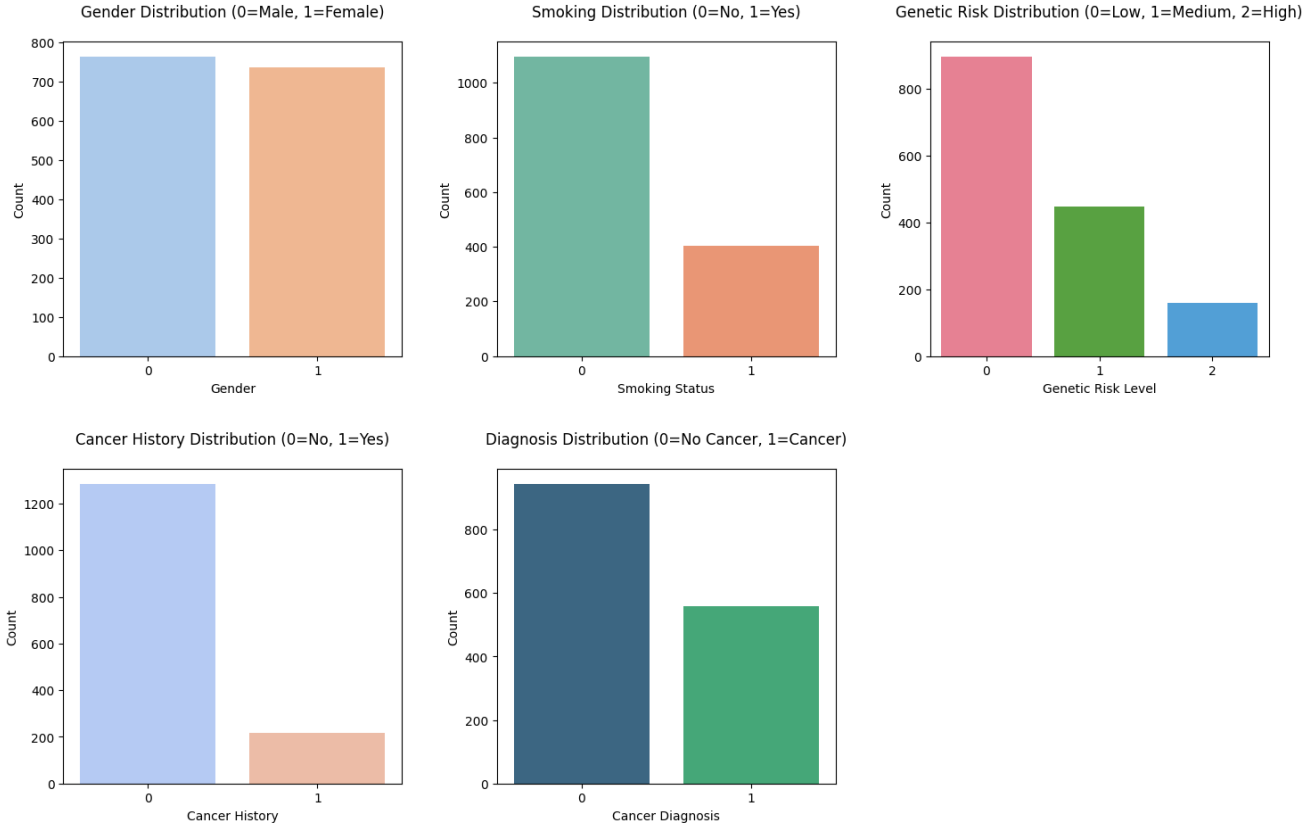
Numerical Feature



In terms of the univariate analysis in regard to numerical values the dataset age feature displays a normally distributed, with a mean of 50.3 years and ranging from 20 to 80 years. Body Mass Index (BMI) of the patients displayed a slight right shewed distribution with a

mean of 27.5, meaning that most patients are overweight (25-30). However, there were outliers in the 15 and 40 range but all within reasonable ranges. Physical activity displays a bimodal distribution with a mean of 4.9 (hours per week). This suggests that there are two separate groups present (active and non-active). Finally, alcohol consumption displayed a right shewed distribution with a mean of 2.4 (units per week), this showed that most patients were within recommended limits of alcohol consumption.

Categorical/Binary Features



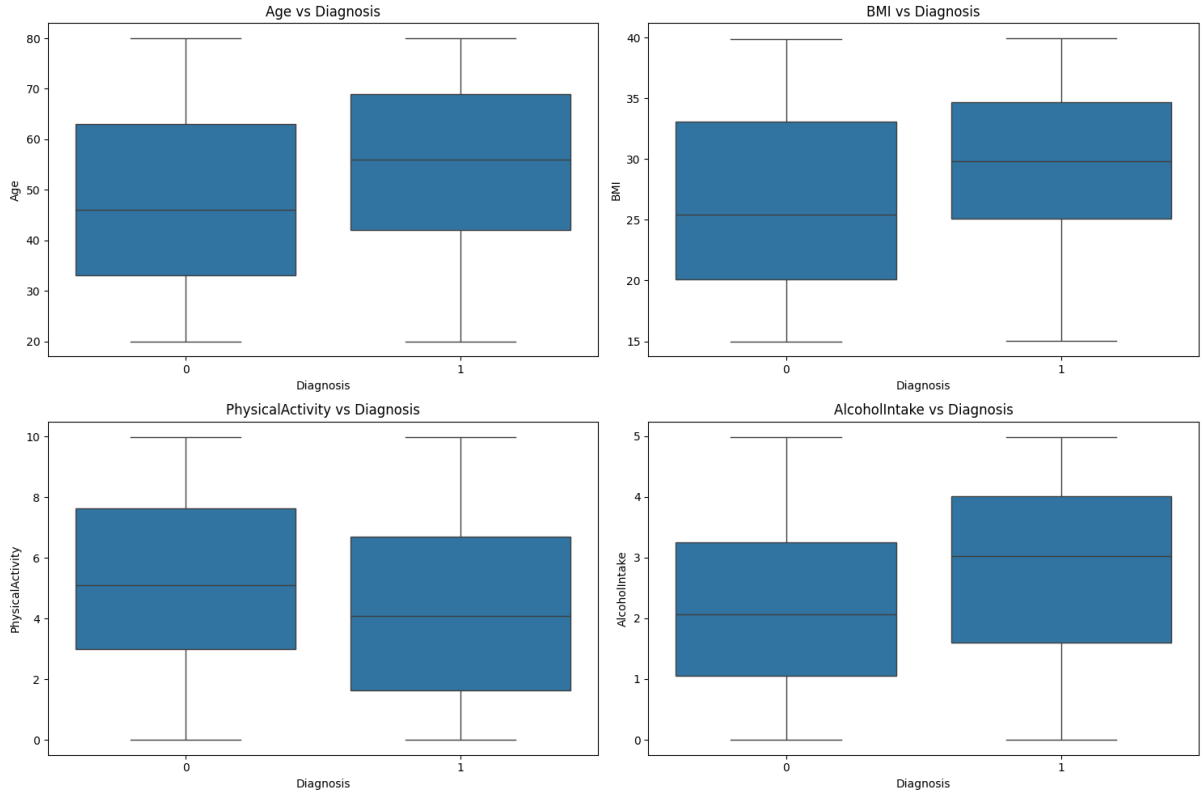
In terms of the univariate analysis in regard to categorical features the dataset displays a nearly balanced gender distribution, with 764 male and 736 female patients. This ensures equal and fair representation of both gender groups. The smoking status shows that most of the patients (1096) are non-smokers compared to only 404 being smokers. This suggests that in this dataset smoking status may be an uncommon risk factor. Genetic risk displays a declining trend, with most patients (895) being low risk, second being medium risk (447) and finally the smallest group being high risk (158). Another finding is that in terms of cancer history most patients (1284) do not have prior cancer history compared to a smaller group of patients (216) who do have cancer history, this means that most cases are new diagnoses.

Bivariate Analysis

For Bivariate Analysis of the features, boxplots (for numerical features) and count plots (for categorical features) were made use of to understand feature and target relationship. A correlation matrix in the form of a heatmap was used to identify feature

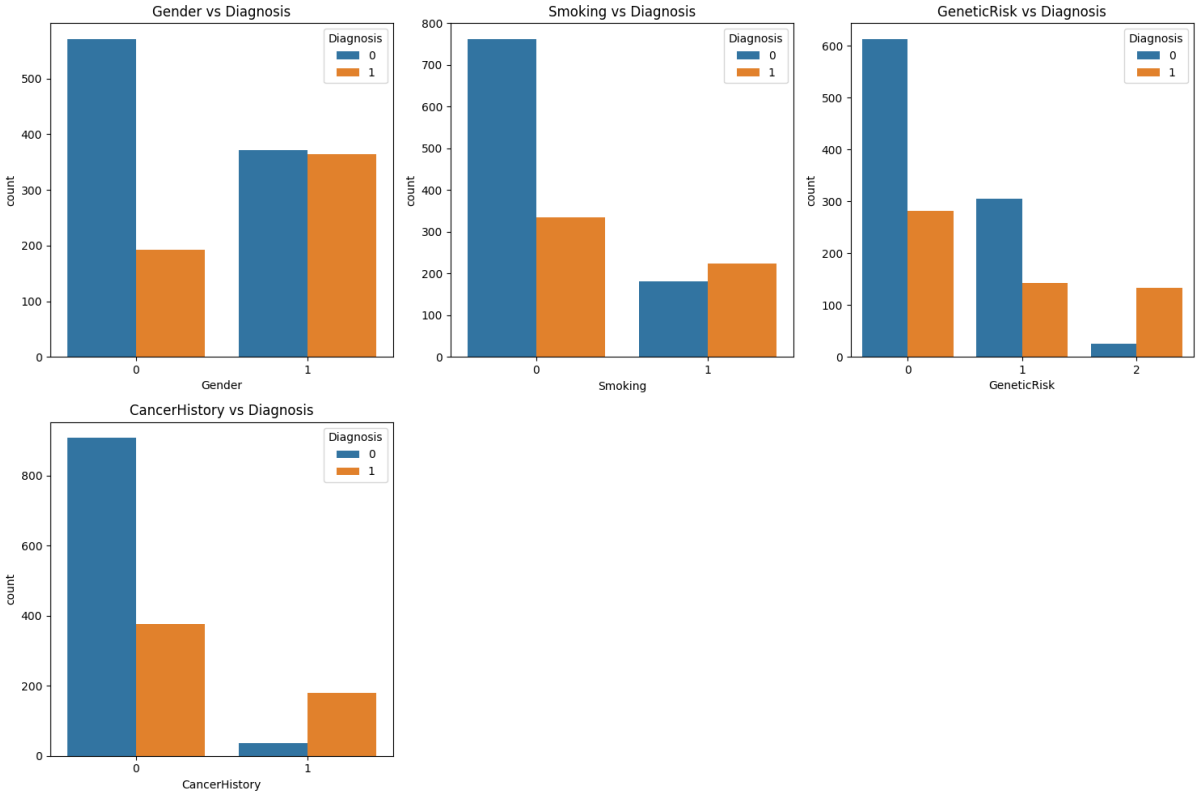
and target correlations. Additionally, a multicollinearity check was performed with the use of Variance Inflation Factor (VIF).

Numerical Features



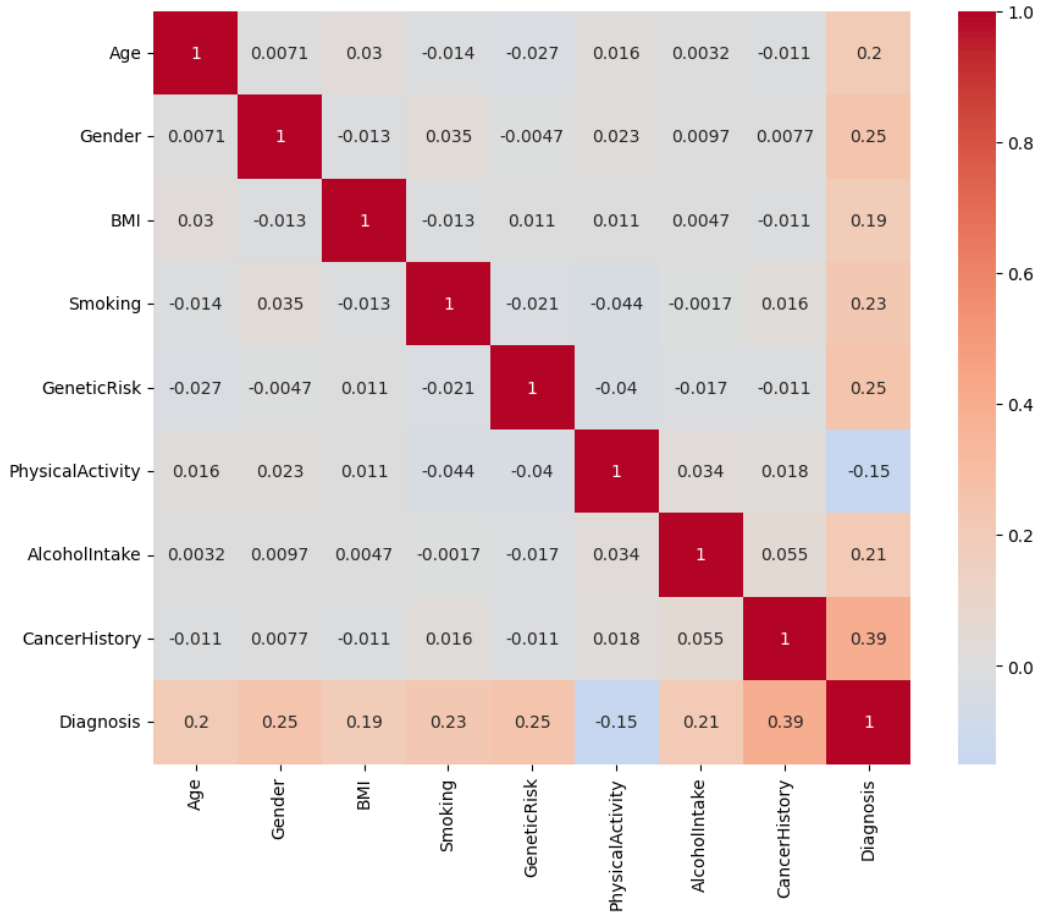
In terms of the bivariate analysis in regard to numerical features and target variable relationships the following was determined. In regard to age, it was found the younger a patient is the less likely they are to have cancer. It also displays that as age increases so does the risk of cancer, with greater risk increases after 40. BMI shows that patients with higher BMIs (25+) have a higher risk of cancer, which suggest that obesity may contribute to cancer risk. This analysis showed in regard to physical activity, patients with lower weekly activity (0-5 hours a week) have a higher risk of cancer. Finally, when looking at alcohol consumption it was determined that patients who consume 3 or more units are more likely to have cancer. These findings demonstrate that age, BMI, physical activity and alcohol consumption are valuable predictors for cancer diagnosis (target).

Categorical Features



In terms of the bivariate analysis in regard to categorical features multiple demographic and risk factor in cancer diagnosis. In terms of gender females display a risk of cancer compared to males. Smoking status shows a strong influence, with smokers being at higher risk of cancer as compared to non-smokers. In terms of genetic risk feature, patients with higher genetic risk show a much higher likelihood of cancer diagnosis compared to patients with lower genetic risk. However, the most interesting finding is in regard to cancer history (patients who have previously had cancer), the analysis showed that patients with prior cancer history are at higher risk of cancer diagnosis. These findings demonstrate that gender, lifestyle choices (smoking), genetic risk and cancer history are valuable predictors for cancer diagnosis (target).

Correlation Analysis



The correlation analysis revealed important insights into the risk factors affecting cancer diagnosis. CancerHistory was found to be the strongest predictor with 0.39 (positive correlation). Gender (0.25), GeneticRisk (0.25) and Smoking (0.23) demonstrate moderate but still meaningful positive correlations with diagnosis, emphasising their importance in cancer detection. Though Age (0.20), BMI (0.19), and AlcoholIntake (0.21) have weaker positive correlations, they still have significant value in terms of cancer detection. An interesting finding is that Physical activity displays a weak negative correlation (-0.15), this suggests that a higher physical activity level would cause a decrease in cancer risk. This shows an inverse relationship between physical activity and cancer outcomes (diagnosis).

Multicollinearity Check

The next step was to identify and address any multicollinearity within the dataset. To do this Variance Inflation Factor (VIF) was used to identify which features needed to be addressed. After running the test, it was determined that since all VIF values were below 10 it meant that there was no extreme multicollinearity present that needed to be addressed.

Feature Selection

In the feature selection statistical tests like t-tests, f-scores and chi-squared tests were used and found that:

T-tests and Chi-squared tests showed that all features had significance in terms of cancer diagnosis as all p-values were below 0.05. Further tests with the use of F-scores revealed feature importance, with cancer history being the strongest, followed by genetic risk and then gender. Based all univariate analysis, bivariate analysis and statistical tests performed it is determined that all features have value and should not be removed.

Model

Normalisation and Scaling

Before training the models the numerical features of the dataset were standardised in order to ensure accuracy. Both Gradient Decent and Elastic Net are sensitive to unscaled data, which can cause biases in regularisation. All categorial data have been encoded.

```
scaler = StandardScaler() ## initialise scaler

[ ] X_train[['Age','BMI','PhysicalActivity','AlcoholIntake']] = scaler.fit_transform(X_train[['Age','BMI','PhysicalActivity','AlcoholIntake']])
    X_test[['Age','BMI','PhysicalActivity','AlcoholIntake']] = scaler.transform(X_test[['Age','BMI','PhysicalActivity','AlcoholIntake']])
```

Smote implementation

SMOTE implementation

Note: That the use of SMOTE was attempted however, it negatively impacted the models performance. As stated above that the class imbalance is not extreme and falls within the 70/30 range. For this reason SMOTE was not kept in the modelling process

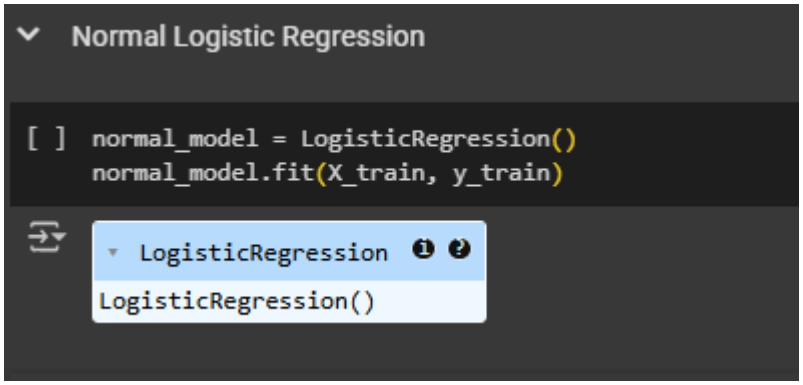
```
smote = SMOTE()
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)
```

Due to the findings after applying SMOTE, it was not implemented in the model training process.

Training models

In order to process the best possible model, multiple models were trained using the same data. The models a trained are a normal logistic regression, gradient decent logistic regression and elasticnet logistic regression.

Normal Logistic Regression



A baseline logistic regression model was trained to serve as a benchmark for the other techniques applied to the model.

Gradient Decent Logistic Regression

```
[ ]  ## GeekForGeeks;
     ## 24 April 2025;
     ## Comparing various online solvers in Scikit Learn;
     ## 1;
     ## source code;
     ## Available at: https://www.geeksforgeeks.org/comparing-various-online-solvers-in-scikit-learn/;
     ## [Accessed on: 19 May 2025].

     gradient_descent_model = SGDClassifier(
         loss='log_loss',
         penalty=None,
         max_iter=1000,
         learning_rate='constant',
         eta0=0.01,
         class_weight='balanced',
         random_state=42
     )

     gradient_descent_model.fit(X_train, y_train)
```

A logistic regression model with gradient decent was trained as one of the techniques applied to better the model performance. In this model multiple attempts with different learning rates were tried, though setting the learning rate to ‘constant’ produced the best results for this model. Another advantage that this class had was the ‘balanced’ class_weight, it did greatly improve model performance (discussed below).

ElasticNet Logistic Regression

```
[ ]  ## GeekForGeeks;
     ## 24 April 2025;
     ## Comparing various online solvers in Scikit Learn;
     ## 1;
     ## source code;
     ## Available at: https://www.geeksforgeeks.org/comparing-various-online-solvers-in-scikit-learn/;
     ## [Accessed on: 19 May 2025].

     elastic_net_model = LogisticRegressionCV(
         penalty='elasticnet',
         solver='saga',
         l1_ratios=[0.1, 0.5, 0.7, 0.9, 1],
         cv=5,
         max_iter=1000,
         class_weight='balanced',
         random_state=42
     )

     elastic_net_model.fit(X_train, y_train)
```

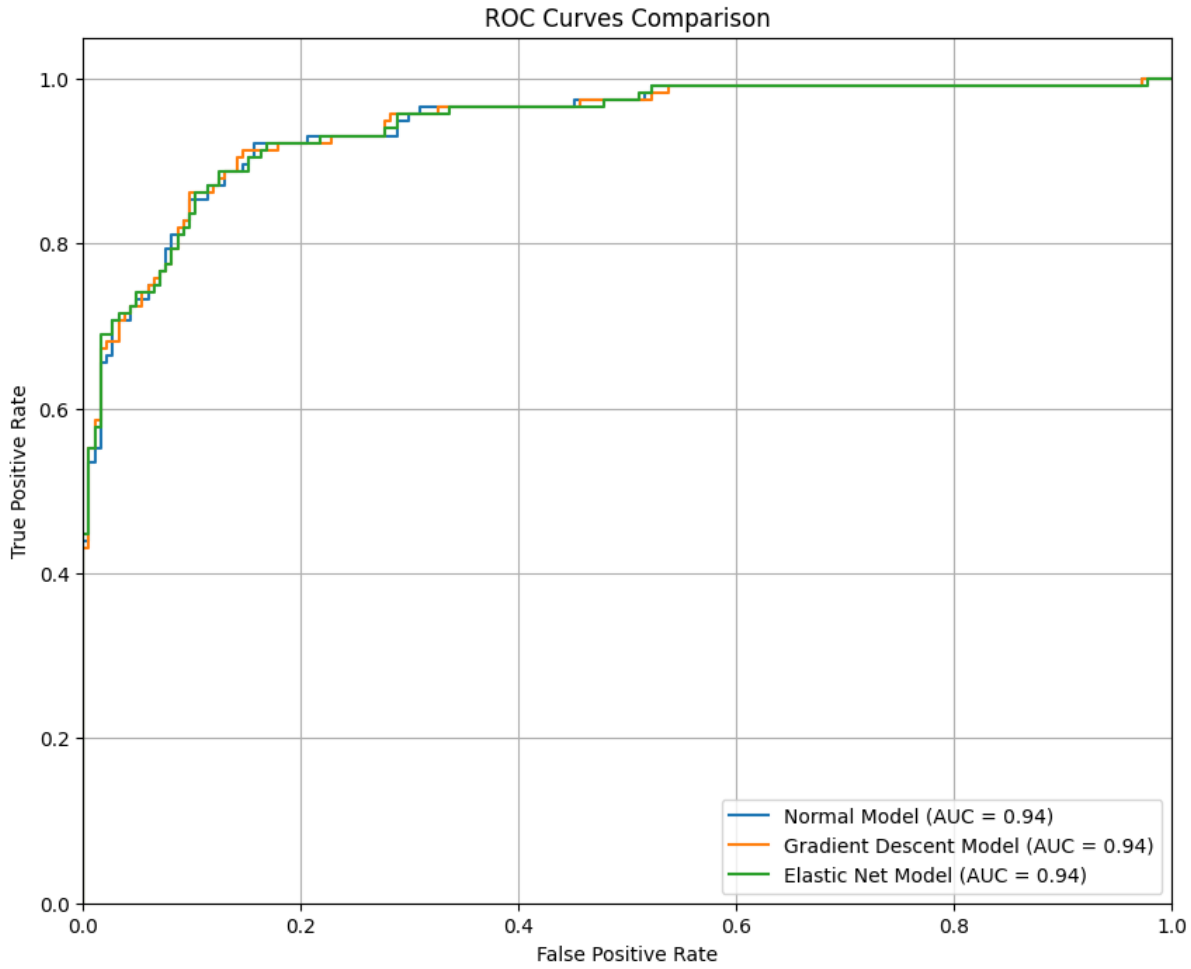
A logistic regression model with ElasticNet, Hyperparameter Tuning and Cross-Validation was trained as one of the techniques applied to better the model performance. In this model LogisticRegressionCV was used to automate the hyperparameter tuning and cross-validation of the model. Again, in this model applying the 'balanced' to class_weight greatly improved model performance.

Evaluation

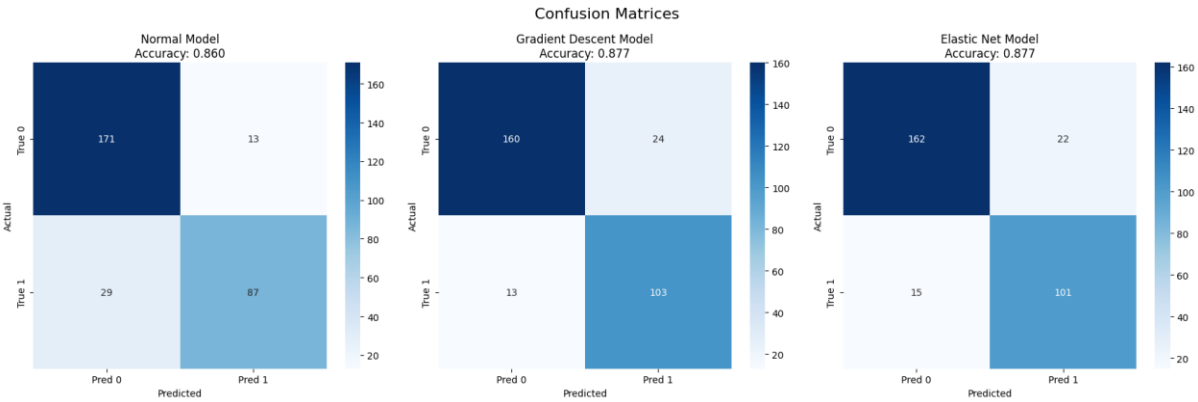
In the evaluation of the model multiple tests were run to measure model performance, identify potential issues as well as errors for improvement. Tools like classification reports and confusion matrixes as well as metrics like accuracies comparisons, ROC-AUC Curve were used to identify the best performing model.

Comparison of different model accuracy was used to understand two things, whether the model had falling victim to overfitting or underfitting and to measure the impact of 'balanced' class_weigth on the models. I found that there was no overfitting or underfitting present in the models and that when 'balanced' class_weigth was applied to the class it greatly improved the model accuracy and overall performance.

The classification report found that all ROC-AUC Scores are the same (0.943) for all three models, this means all models have the same level of performance in terms of distinguish between classes. In terms of accuracy the baseline Logistic Regression model have an accuracy of 86% whereas the Logistic Regression model with Gradient Decent and the Logistic Regression model with Hyperparameter (ElasticNet) Tuning and Cross-Validation have an accuracy of 88%.



Since all the ROC-AUC Scores were the same for all three models, confusion matrices were used to find underlying strengths and weaknesses of each model, to determine the best possible model.



The confusion matrix for the baseline Logistic regression model (normal model) found that the model underpredicts class 1 we can see this due to a high FN however, predictions for class 1 are reliable as seen from a high precision. For the gradient decent model, the confusion matrix showed that the model is better at identifying positives (lower FN) but at cost of more false alarms (FP). Finally, for the elastic net model the confusion matrix showed

that the model offers the best trade off as it has fewer FPs than Gradient Decent and offers a good recall of class 1.

Conclusion

So based on findings from the classification report, ROC-AUC Curve and the confusion matrix it is determined that the Gradient Decent model is the best model for this case. This is due to high recall for class 1 (Cancer is present) with 89%. It only missed 13 out of 116 cancer cases as compared to 29 cases from the normal model and 15 cases from the elastic net model. Though the precision may be at 81%, the lowest of three models, this is a fair trade-off.

Additionally, the Gradient Decent model is the most cost-effective model, though the elastic net model would decrease false alarms which would decrease costs (in terms of medical tests). This would come at a price of missing 2 cases for every 116, which is a big issue for a medical setting.