**Breast Cancer Diagnosis Prediction**

**Introduction**

This report details the process of building and evaluating machine learning models to predict breast cancer diagnosis based on clinical and cell images features. The dataset used contains 569 instances with 30 features and 1 target variable.

**Data Exploration and Preprocessing**

**Data Overview**

The dataset includes:

* **569 instances** with no missing values
* **30 numerical features** and **1 categorical target variable (Diagnosis)**

**Exploratory Data Analysis (EDA)**

**Histograms of Features**

* A group of blue bars

  Description automatically generatedVisualized distributions of each feature to understand their distributions and identify potential outliers.

**Class Imbalance Check**

**Class Distribution**:

* B (Benign): 357 instances
* M (Malignant): 212 instances
*  **Imbalance Ratio**: 0.59
*  Recommendation: Since the dataset is slightly imbalanced (ratio < 0.5), oversampling techniques could be considered.

**Data Preprocessing**

* **Drop ID Column**: Removed 'ID' column as it does not contribute to prediction.
* **Encoding**: Converted categorical target variable (Diagnosis) into binary labels

(B: 0, M: 1)

* **Train-Test Split**: Split the data into training (80%) and testing (20%) sets for model evaluation.

**Model Building and Evaluation**

**Logistic Regression**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **B** | 0.9740 | 1.0000 | 0.9868 | 75 |
| **M** | 1.0000 | 0.9487 | 0.9737 | 39 |
| **accuracy** |  |  | 0.9825 | 114 |
| **macro avg** | 0.987 | 0.9744 | 0.9803 | 114 |
| **weighted avg** | 0.9829 | 0.9825 | 0.9823 | 114 |

**A graph of a function

Description automatically generated with medium confidence**

**Logistic Regression Classifier with Hyperparameter Tuning**

Optimized Model using GridSearchCV.

* **Best Parameters**:

C: 50, Penalty: L1, Solver: 'liblinear'

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **B** | 0.9868 | 1.0000 | 0.9934 | 75 |
| **M** | 1.0000 | 0.9744 | 0.9870 | 39 |
| **accuracy** |  |  | 0.9912 | 114 |
| **macro avg** | 0.9934 | 0.9872 | 0.9902 | 114 |
| **weighted avg** | 0.9913 | 0.9912 | 0.9912 | 114 |

**Random Forest Classifier**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **B** | 0.9733 | 0.9733 | 0.9733 | 75 |
| **M** | 0.9487 | 0.9487 | 0.9487 | 39 |
| **accuracy** |  |  | 0.9649 | 114 |
| **macro avg** | 0.9610 | 0.9610 | 0.9610 | 114 |
| **weighted avg** | 0.9649 | 0.9649 | 0.9649 | 114 |

**A graph of a function

Description automatically generated with medium confidence**

**Support Vector Machine (SVM)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **B** | 0.9605 | 0.9733 | 0.9669 | 75 |
| **M** | 0.9474 | 0.9231 | 0.9351 | 39 |
| **accuracy** |  |  | 0.9561 | 114 |
| **macro avg** | 0.9539 | 0.9482 | 0.9510 | 114 |
| **weighted avg** | 0.9560 | 0.9561 | 0.9560 | 114 |

**A colorful squares with numbers

Description automatically generated**

Figure 1 | confusion Matrix of SVM

**SVM Classifier with Hyperparameter Tuning**

GridSearchCV Optimized SVM

* **Best Parameters**:

C: 100, Gamma: 0.01, Kernel: 'rbf' (Radial Basis Function (RBF) kernel, also known as the Gaussian kernel)

A graph of a function

Description automatically generated

**Performance Metrics After Tunning Hyper Parameters**:

* **Accuracy**: 98.25%
* **Precision (Malignant)**: 1.0
* **Recall (Malignant)**: 0.97

**K-Nearest Neighbors Classifier**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **B** | 0.9733 | 0.9733 | 0.9733 | 75 |
| **M** | 0.9487 | 0.9487 | 0.9487 | 39 |
| **accuracy** |  |  | 0.9649 | 114 |
| **macro avg** | 0.9610 | 0.9610 | 0.9610 | 114 |
| **weighted avg** | 0.9649 | 0.9649 | 0.9649 | 114 |

A graph with orange lines

Description automatically generated

**Decision Tree Classifier**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **B** | 0.9730 | 0.9600 | 0.9664 | 75 |
| **M** | 0.9250 | 0.9487 | 0.9367 | 39 |
| **accuracy** |  |  | 0.9561 | 114 |
| **macro avg** | 0.9490 | 0.9544 | 0.9516 | 114 |
| **weighted avg** | 0.9566 | 0.9561 | 0.9563 | 114 |

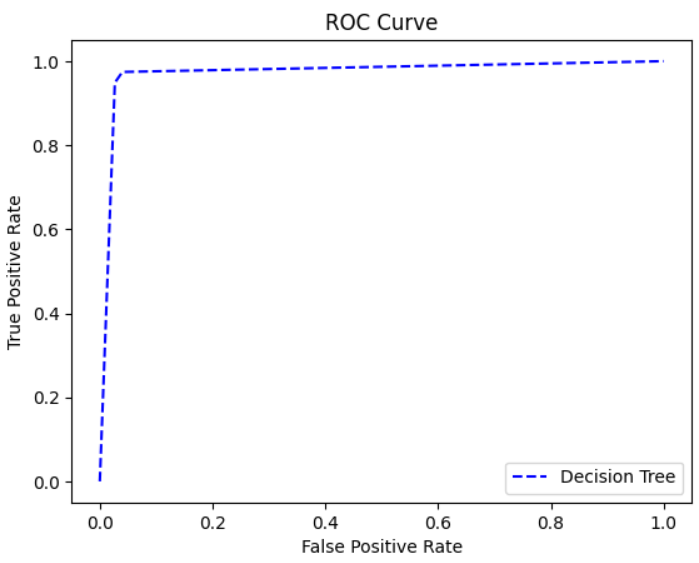
**Hyperparameters tuning with Decision Tree**

GridSearchCV Optimized SVM

* **Best Parameters**:

Criterion (gini, entropy) , max\_depth , min\_samples\_split , min\_samples\_leaf

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **B** | 0.9733 | 0.9733 | 0.9733 | 75 |
| **M** | 0.9487 | 0.9487 | 0.9487 | 39 |
| **accuracy** |  |  | 0.9649 | 114 |
| **macro avg** | 0.9610 | 0.9610 | 0.9610 | 114 |
| **weighted avg** | 0.9649 | 0.9649 | 0.9649 | 114 |

****

**Conclusion**

* **Best Model**: SVM with GridSearchCV (Accuracy: 98.25%, AUC: 0.99)
* **Recommendation**: Deploy the optimized SVM model for predicting breast cancer diagnosis due to its high accuracy and robust performance.