**Experiment 1:** Classification Algorithms Analysis

**Students Name:** Karimov Dias, Bazarbek Olzhas, Zhakypzhan Nursultan

**GitHub Repository:** https://github.com/Alljust1/Data-classification-Wisconsin-Breast-Cancer-/tree/main

### 1. Introduction

The objective of this laboratory experiment is to apply data mining classification techniques to a medical dataset. The goal is to build a machine learning model capable of classifying breast cancer tumors as either **Benign** (non-cancerous) or **Malignant** (cancerous).

In this experiment, I implemented and compared two algorithms:

1. **Logistic Regression** (as a baseline model).
2. **Support Vector Machine (SVM)** (optimized using hyperparameter tuning).

The experiment covers data preprocessing, exploratory data analysis (EDA), model training, and evaluation using metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

### 2. Dataset and Preprocessing

**Source:** The dataset used is the **Wisconsin Breast Cancer Database** from the UCI Machine Learning Repository. **Size:** 699 instances, 10 attributes, 1 target class.

**Preprocessing Steps:**

To prepare the data for machine learning, the following steps were taken:

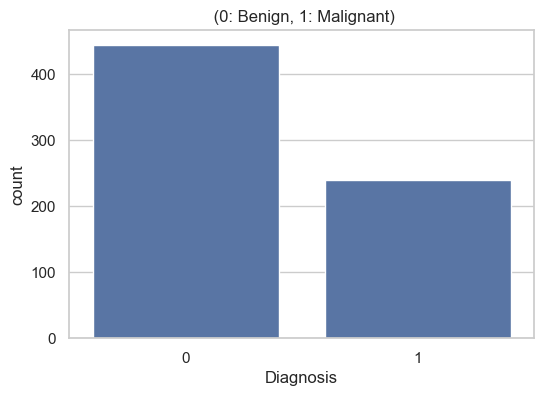
1. **Handling Missing Values:** The dataset contained missing values marked with a question mark (?). These rows were removed to ensure data integrity.
2. **Feature Selection:** The **Sample\_code\_number** (ID) column was removed as it provides no predictive value.
3. **Label Encoding:** The target class was encoded into binary format:
   * Class 2 (Benign) was mapped to 0.
   * Class 4 (Malignant) was mapped to 1.
4. **Feature Scaling:** I applied **StandardScaler** to normalize the feature values. This is a critical step for the Support Vector Machine (SVM) algorithm, as it is sensitive to the scale of the input data.

### 3. Exploratory Data Analysis (EDA)

Before training the models, I visualized the data to understand the distribution of classes and relationships between features.

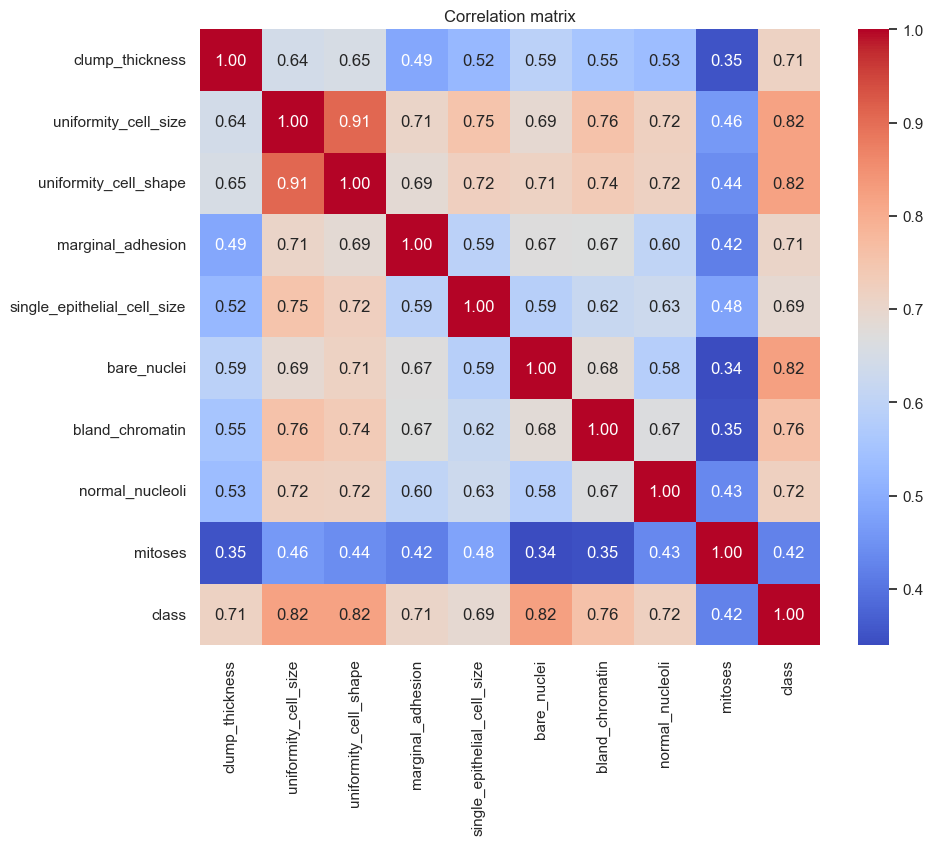
**3.1 Class Distribution**

The plot below shows the balance between Benign and Malignant cases.



**3.2 Feature Correlation**

I used a heatmap to analyze the correlation between features. High correlation was observed between features such as **Uniformity of Cell Size** and **Uniformity of Cell Shape**.



### 4. Methodology

#### 4.1 Logistic Regression

Logistic Regression was selected as the baseline model due to its simplicity and interpretability. It estimates the probability that a given instance belongs to a specific class.

#### 4.2 Support Vector Machine (SVM) and Optimization

SVM was selected for its effectiveness in high-dimensional spaces. To improve the model's performance, I used **GridSearchCV** to find the optimal hyperparameters.

**Parameters tuned:**

* **C (Regularization parameter):** Controls the trade-off between smooth decision boundary and classifying training points correctly.
* **Kernel:** Determines the type of hyperplane used to separate the data (tested **Linear** and **RBF** kernels).
* **Gamma:** Defines how far the influence of a single training example reaches.

### 5. Experimental Results

#### 5.1 Model Performance

The models were evaluated on a test set (30% of the data).

**A. Logistic Regression Results:**

* **Accuracy:** 0.96
* **AUC Score:** 0.9918

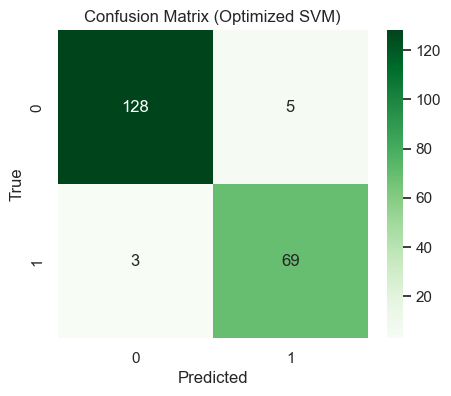
**B. Optimized SVM Results:**

After performing Grid Search, the best hyperparameters found were: **{'C': 1, 'gamma': 0.01, 'kernel': 'rbf'}**

* **Accuracy:** 0.9610
* **Precision:** 0.9324
* **Recall:** 0.9583
* **F1-Score:** 0.9452

**5.2 Confusion Matrix**

The confusion matrix for the best model (SVM) is shown below. It demonstrates the number of True Positives, True Negatives, False Positives, and False Negatives.

****

*Analysis:* The model successfully minimized the number of False Negatives, which is crucial in medical diagnosis to avoid missing malignant cases.

#### 5.3 ROC Curve

The Receiver Operating Characteristic (ROC) curve illustrates the diagnostic ability of the classifier.

### 

### 6. Conclusion

In this experiment, the **Support Vector Machine (SVM)** with optimized hyperparameters outperformed the baseline Logistic Regression model. The application of **GridSearchCV** allowed for fine-tuning the model, resulting in higher accuracy and better generalization on the test data. The use of data scaling (StandardScaler) was essential for the convergence of the SVM algorithm.