# Breast Cancer Classification Project Documentation

## 1. Introduction

This documentation presents an analysis of breast cancer data for classification using Support Vector Machine (SVM) model. The project involves data collection, preprocessing, exploratory data analysis (EDA), feature extraction, model training, and evaluation.
Dataset: Breast Cancer Dataset (kaggle.com)

## 2. Importing Libraries

Pandas, NumPy, matplotlib, and seaborn libraries are imported for data manipulation and visualization.

Scikit-learn libraries are imported for machine learning tasks.

## 3. Data Collection and Preprocessing

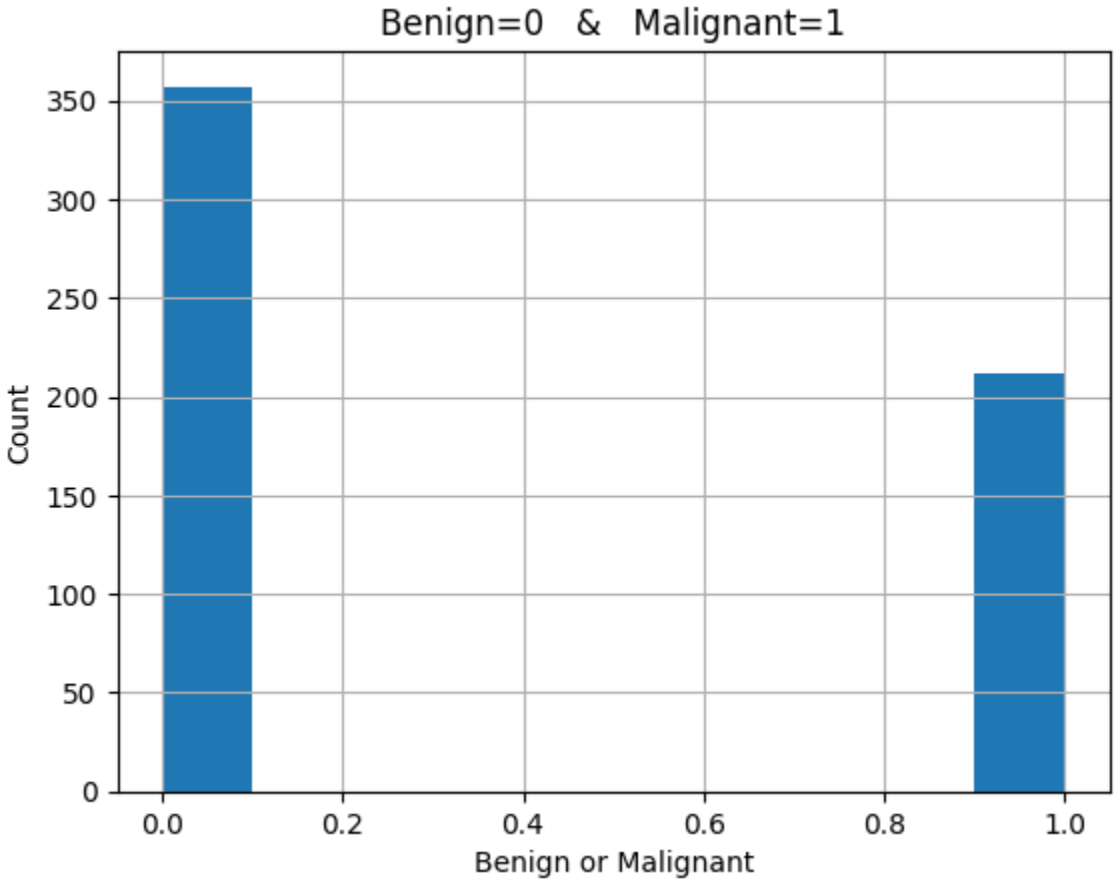
- Breast cancer dataset is loaded from a CSV file into a pandas DataFrame.
- Data shape and information are analyzed to understand the dataset"s structure.
- Missing values are checked and handled if present.
- Label encoding is applied to convert categorical target variable ("diagnosis") into numerical format.
- Functions used:
  - → pd.read_csv()
  - → df.drop()
  - → df.head()
  - → df.info()df.isnull().sum()
  - → df.describe()
  - → from sklearn.preprocessing import LabelEncoder
    label_encoder = LabelEncoder()
    df["diagnosis"] = label_encoder.fit_transform(df["diagnosis"])

## 4. Exploratory Data Analysis (EDA)

### 4.1 Histogram Analysis for Class Imbalance

A histogram is a graphical representation of the distribution of numerical data. It consists of a series of vertical bars, where each bar represents the frequency or count of data values falling within specific intervals or bins. Histograms provide insights into the shape, center, and spread of the data distribution, allowing for visual analysis of data patterns and identifying potential outliers or skewness.

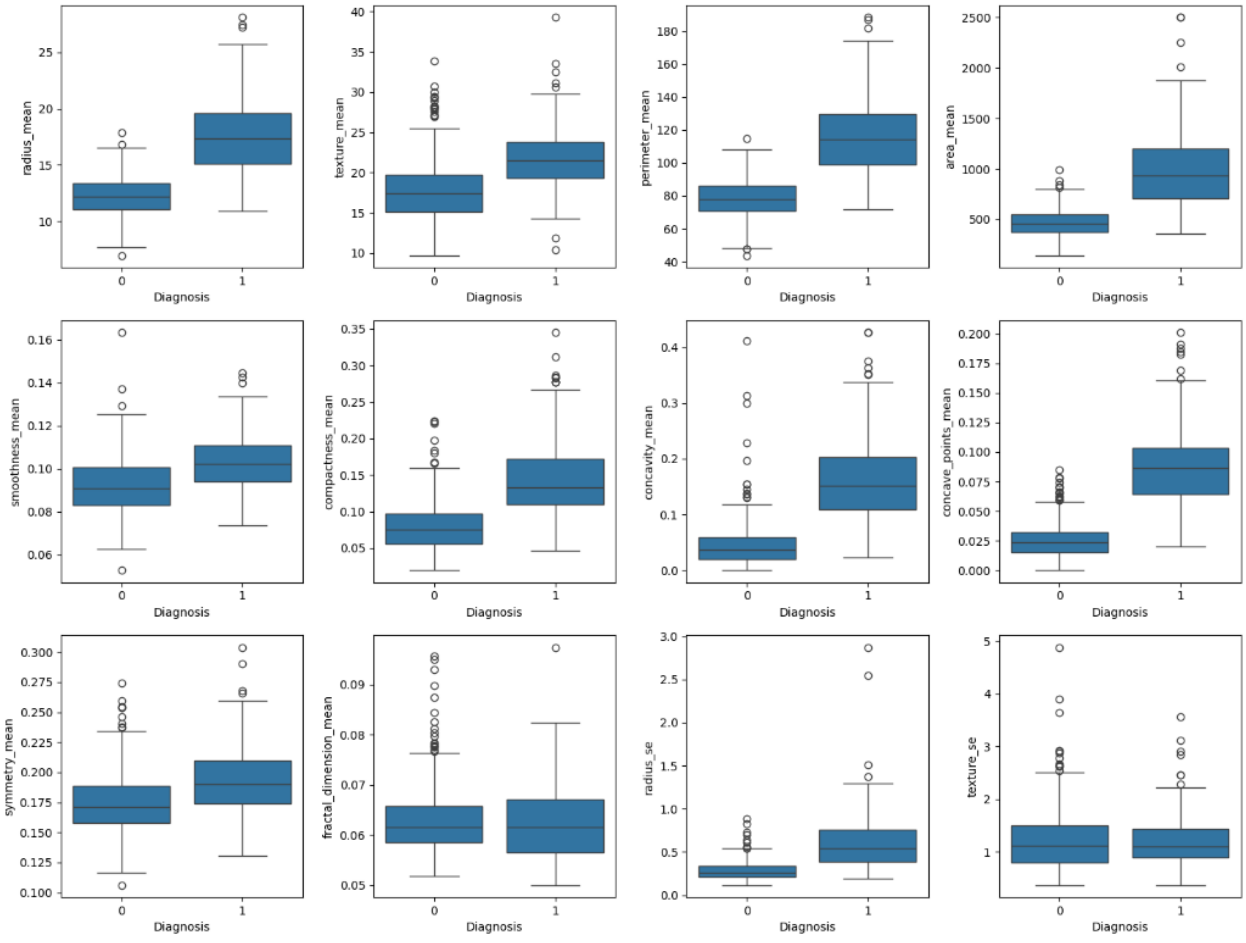
★ Histogram is plotted to visualize the distribution of benign and malignant diagnoses.



## 4.2 Box Plot for Visualizing Relationship

A box plot, also known as a box-and-whisker plot, is a graphical representation of the distribution of numerical data through quartiles. It consists of a rectangular box, which spans the interquartile range (IQR) of the data, with a line inside representing the median. Whiskers extend from the box to show the range of the data excluding outliers. Box plots provide insights into the central tendency, variability, and skewness of the data distribution, facilitating comparison between different groups or datasets.

★ Box plots are created to visualize the relationship between the dependent variable ("diagnosis") and independent variables.



## 4.3 Correlation Analysis and Heatmap

A correlation matrix is a table that shows the correlation coefficients between variables in a dataset. Each cell in the matrix represents the correlation between two variables, ranging from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. Correlation matrices help identify relationships between variables and understand how they change together.

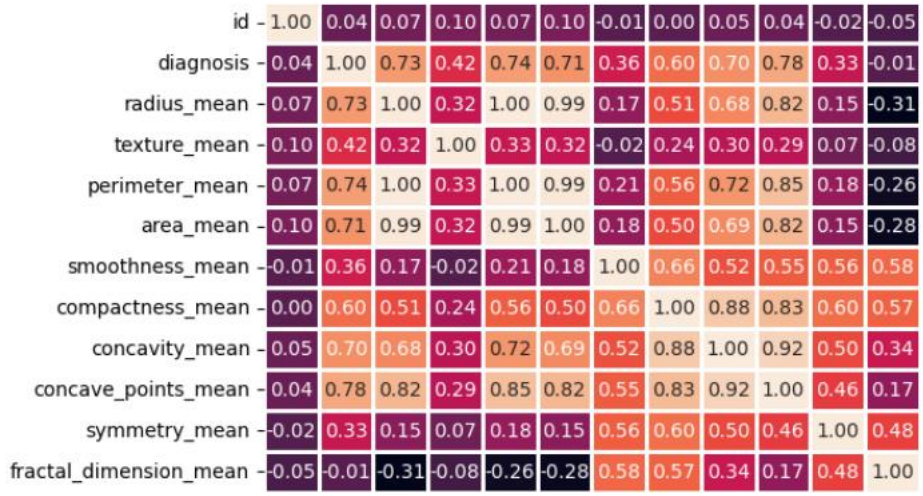
A heatmap is a graphical representation of data where values in a matrix are represented as colors. In the context of a correlation matrix, a heatmap visualizes the correlation coefficients using a color scale, with different colors indicating the strength and direction of correlations. Heatmaps provide a visual summary of correlations in a dataset, making it easier to identify patterns and relationships between variables.

★ Correlation matrix is computed to analyze the correlation between features.

```
df.corr()
```

| | id | diagnosis | radius_mean | texture_mean | perimeter_mean | area_mean | smoothness_mean | compactness_mean | concavity_mean |
|---|---|---|---|---|---|---|---|---|---|
| **id** | 1.000000 | 0.039769 | 0.074626 | 0.099770 | 0.073159 | 0.096893 | -0.012968 | 0.000096 | 0.050080 |
| **diagnosis** | 0.039769 | 1.000000 | 0.730029 | 0.415185 | 0.742636 | 0.708984 | 0.358560 | 0.596534 | 0.696360 |
| **radius_mean** | 0.074626 | 0.730029 | 1.000000 | 0.323782 | 0.997855 | 0.987357 | 0.170581 | 0.506124 | 0.676764 |
| **texture_mean** | 0.099770 | 0.415185 | 0.323782 | 1.000000 | 0.329533 | 0.321086 | -0.023389 | 0.236702 | 0.302418 |
| **perimeter_mean** | 0.073159 | 0.742636 | 0.997855 | 0.329533 | 1.000000 | 0.986507 | 0.207278 | 0.556936 | 0.716136 |
| **area_mean** | 0.096893 | 0.708984 | 0.987357 | 0.321086 | 0.986507 | 1.000000 | 0.177028 | 0.498502 | 0.685983 |
| **smoothness_mean** | -0.012968 | 0.358560 | 0.170581 | -0.023389 | 0.207278 | 0.177028 | 1.000000 | 0.659123 | 0.521984 |
| **compactness_mean** | 0.000096 | 0.596534 | 0.506124 | 0.236702 | 0.556936 | 0.498502 | 0.659123 | 1.000000 | 0.883121 |
| **concavity_mean** | 0.050080 | 0.696360 | 0.676764 | 0.302418 | 0.716136 | 0.685983 | 0.521984 | 0.883121 | 1.000000 |

★ Heatmap is plotted to visualize the correlation matrix.



## 5. Feature Extraction and Engineering

- Less relevant features are dropped based on correlation analysis.
  Example: ["id", "fractal_dimension_mean", "texture_se", "smoothness_se", "symmetry_se", "fractal_dimension_se"]

- New features are engineered by combining highly related features like radius, perimeter, area, etc.

- The final dataset shape is ***(569, 20)*** after feature engineering.

## 6. Training the SVM Model for Tumor Classification

- Data is split into training and testing sets.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train ,y_test = train_test_split(X, y, test_size=0.2, random_state=2)
```

- Features are scaled using StandardScaler.

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

- Support Vector Machine (SVM) model with optimized hyperparameters is trained on the training data.

```
from sklearn.svm import SVC
model = SVC(C=15, gamma=0.01, probability=True)
model.fit(X_train, y_train)
```

```
SVC(C=15, gamma=0.01, probability=True)
```

- Model predictions are made on the testing data.

```
y_pred = model.predict(X_test)
y_pred
```

```
array([0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0,
       0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0,
       1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0,
       1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0,
       1, 1, 0, 1])
```

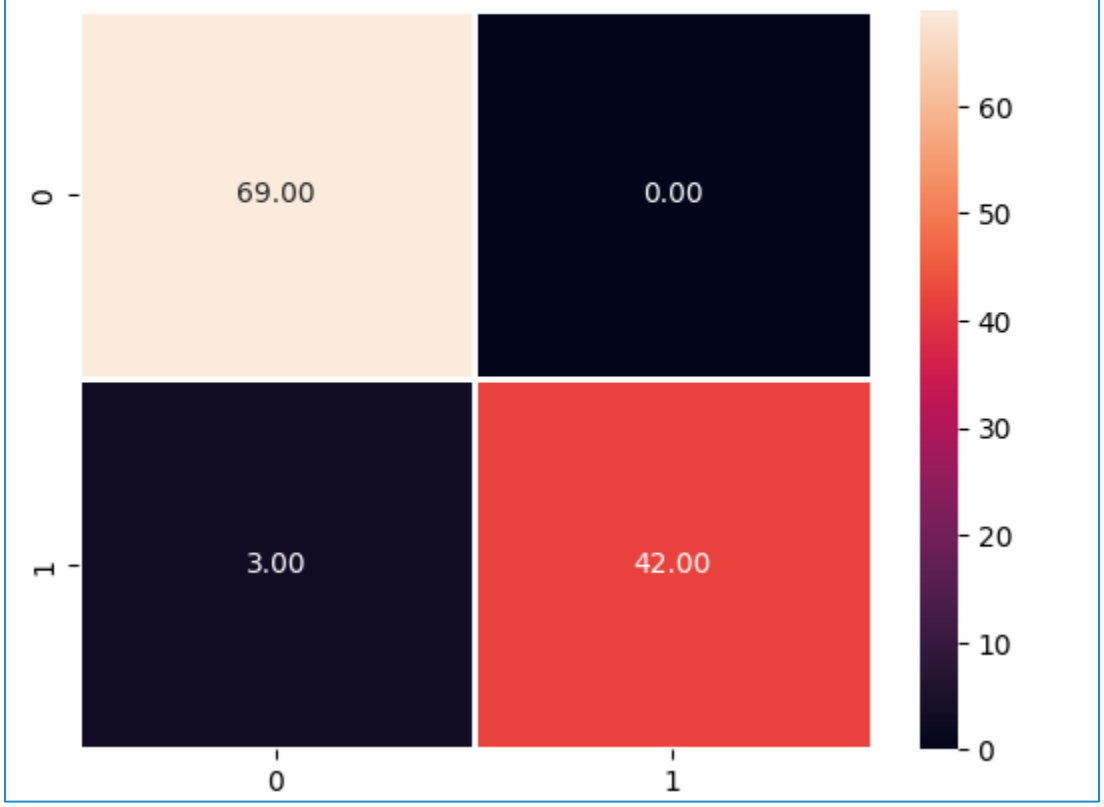
## 7. Evaluating the model's performance

### 7.1 Confusion Matrix and Heatmap

★ Accuracy score is computed and Heatmap is plotted to visualize the confusion matrix.

Accuracy: 0.9736842105263158

Heatmap:



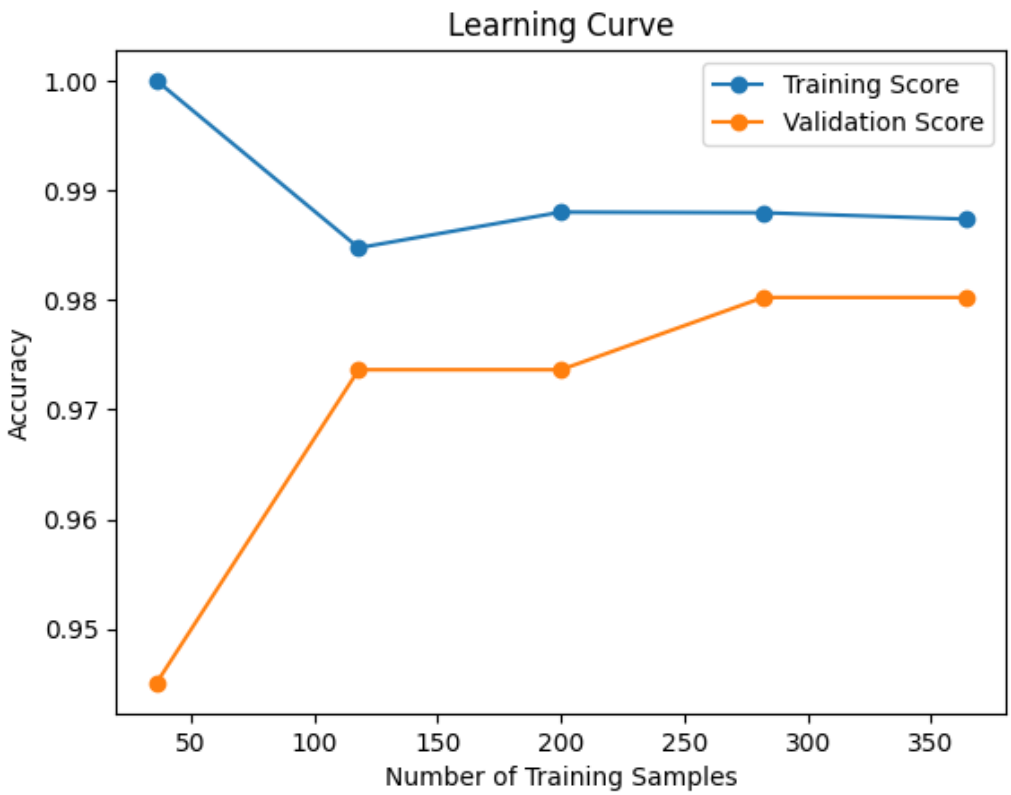

★ Classification report is computed to evaluate the performance of the SVM model.

Classification report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.96 | 1.00 | 0.98 | 69 |
| 1 | 1.00 | 0.93 | 0.97 | 45 |
| accuracy | | | 0.97 | 114 |
| macro avg | 0.98 | 0.97 | 0.97 | 114 |
| weighted avg | 0.97 | 0.97 | 0.97 | 114 |

## 7.2 Learning Curve

★ Learning curves are plotted to visualize the model"s performance on training and validation sets with varying training sample sizes.




# 8. Conclusion

*The SVM model demonstrates 97% accuracy in classifying breast cancer tumors.*

Additional Insights:

- Features like "shape", "texture", "compactness", "concavity", "concave points" and "symmetry" were found to be most important in distinguishing between benign and malignant tumors.
- Feature selection enhanced the model by identifying the most relevant factors for prediction, improving model accuracy, reducing overfitting, and increasing interpretability, ultimately leading to more effective risk assessment and treatment decisions.
- Future research in breast cancer detection models could focus on integrating multi-omics data for enhanced accuracy, improving interpretability to understand underlying biological mechanisms, developing real-time monitoring capabilities, considering environmental and lifestyle factors, addressing health disparities, and streamlining validation and clinical implementation processes for broader adoption and impact.