Project Title:

Cancer Cell Classification using Support Vector Machines.

Project Description:

The goal of this project is to develop a machine learning model using Support Vector Machines (SVM) to predict whether a cell is benign or malignant based on the cell's features. We will be using the Breast Cancer Wisconsin (Diagnostic) dataset, which contains features such as the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension of the cell nuclei.

Steps Involved:

1. **Data Preparation:**

In this step, we will first download the dataset and then perform some preprocessing steps such as data cleaning, feature selection, and normalization to prepare the dataset for SVM classification.

2. **Feature Extraction:**

We will extract the relevant features from the dataset and split the data into training and testing sets.

3. **SVM Model Creation:**

We will create a SVM model using scikit-learn library, and then train the model on the training dataset.

4. **Model Evaluation:**

We will evaluate the performance of the model using various metrics such as accuracy, precision, recall, and F1-score. We will also visualize the results using confusion matrix and ROC curves.

5. **Model Tuning:**

In this step, we will tune the hyperparameters of the SVM model to improve its performance and increase the accuracy of the predictions.

6. **Prediction:**

Finally, we will use the trained model to predict whether a new cell is benign or malignant based on its features.

Tools and Technologies:

1. Python Programming Language:
We will be using Python to implement the machine learning model and perform data preprocessing.

2. Scikit-Learn Library:
We will be using scikit-learn library to create and train the SVM model.

3. Jupyter Notebook:
We will use Jupyter Notebook for coding and to document our progress.

Expected Outcome:

The expected outcome of this project is a highly accurate SVM model that can predict whether a given cell is benign or malignant with high accuracy. By using this model, we can aid in the diagnosis of cancer patients and help in the early detection of cancer.

Support Vector Machines (SVMs):

Support Vector Machines (SVMs) are a class of supervised machine learning algorithms that can be used for classification and regression analysis. SVMs are particularly effective in high-dimensional data spaces and when there is a clear margin of separation between different classes.

The basic idea behind SVMs is to find a hyperplane that best separates the different classes in the data space. The hyperplane is chosen such that it maximizes the margin between the closest points from each class, known as support vectors. This results in a decision boundary that is robust to noise and can generalize well to new data.

In SVM, there are different kernel functions that can be used to transform the data into a higher-dimensional space where a hyperplane can better separate the different classes. The choice of kernel function can significantly affect the accuracy and generalization performance of the model. Some of the commonly used kernel functions are:

1. Linear Kernel: The linear kernel simply calculates the dot product between two feature vectors. It is computationally efficient and works well when the data is linearly separable.

2. Polynomial Kernel: The polynomial kernel maps the data into a higher-dimensional space using a polynomial function. It is useful when the data is not linearly separable and can capture complex non-linear relationships.

3. Radial Basis Function (RBF) Kernel: The RBF kernel maps the data into an infinite-dimensional space using a Gaussian function. It is one of the most widely used kernel functions and can capture complex non-linear relationships.

4. Sigmoid Kernel: The sigmoid kernel maps the data into a higher-dimensional space using a sigmoid function. It is useful for neural network applications but is generally not recommended for SVM.

In the project, different kernel functions were used to create SVM models for the cancer cell dataset. The accuracies of the models were evaluated using the test data. The accuracies of the models using different kernel functions are as follows:

RBF Kernel: 0.956

Linear Kernel: 0.965

Polynomial Kernel: 0.964

Sigmoid Kernel: 0.610

Based on the results, the linear kernel performed slightly better than the other kernel functions, with an accuracy of 0.965. However, all of the kernel functions produced high accuracies, except for the sigmoid kernel.

About the Model and Dataset:

In the code, We will be using the concept of Support Vector Machines (SVM) to create a svm model to predict weather the cell is benign or malignant. The dataset used in this model have collection of cells of cancer pateints and on the basis of this we will be making our prediction.

Data Preprocessing:

In the dataset, we have some rows in BareNuc which is empty so we will be dropping those rows from the column BareNuc from the dataset. And by using the concept of list comprehension checking weather still do we have any null values or not. As per our code empty list [] means there is no null values in that column as you can see below in the code.

SVM Model Creation:

Created a SVM model for prediction and each time SVM model we have created by using different kernel functions just to check weather which function is better for this dataset. After creating the model in next cell storing the prediction into the variables and then in next cell printing the accuracy score of all the models created by using different kernel functions. We can observe we got almost 95% accuracy in all the function except Sigmoid function, so we can use any of the three functions for prediction by using this dataset.

Conclusion:

In this project, I have used the concept of Support Vector Machines (SVM) to create an SVM model to predict whether a cell is benign or malignant. Firstly, visualized the data using a pair plot and then preprocessed the data by dropping rows with null values in the BareNuc column and converting it to an integer type. Then created an SVM model using the RBF kernel function and evaluated its performance using the accuracy score, confusion matrix, and classification report. The model achieved an accuracy of 96%, which is pretty good.