# Analysis report of Lab assignment : “SVM”

**Name : Sabra Elhajjaji**

**Student Number : 301488628**

# Assignment 3– exercise1:

### Load & check the data:

* The dataset contains 699 rows and 11 columns, representing different features related to breast cancer diagnosis.
* The 'ID' column is an integer identifier and does not contribute to classification, so it should be dropped.
* The 'thickness', 'size', 'shape', 'Marg', 'Epith', 'b1', 'nucleoli', and 'Mitoses' columns are integer features that appear to represent different cell characteristics.
* The 'bare' column is stored as an object type, which suggests it contains non-numeric values (possibly missing values encoded as '?'). This column needs preprocessing.
* The 'class' column represents the target variable, where values likely indicate different classes of diagnosis (e.g., benign or malignant).
* There are no missing values detected in the dataset, but the 'bare' column contains values that may need cleaning.
* The summary statistics show that most of the features range between 1 and 10, indicating categorical-like ordinal values.
* The mean and standard deviation suggest some features might require normalization before applying machine learning models.

### Pre-process and visualize the data

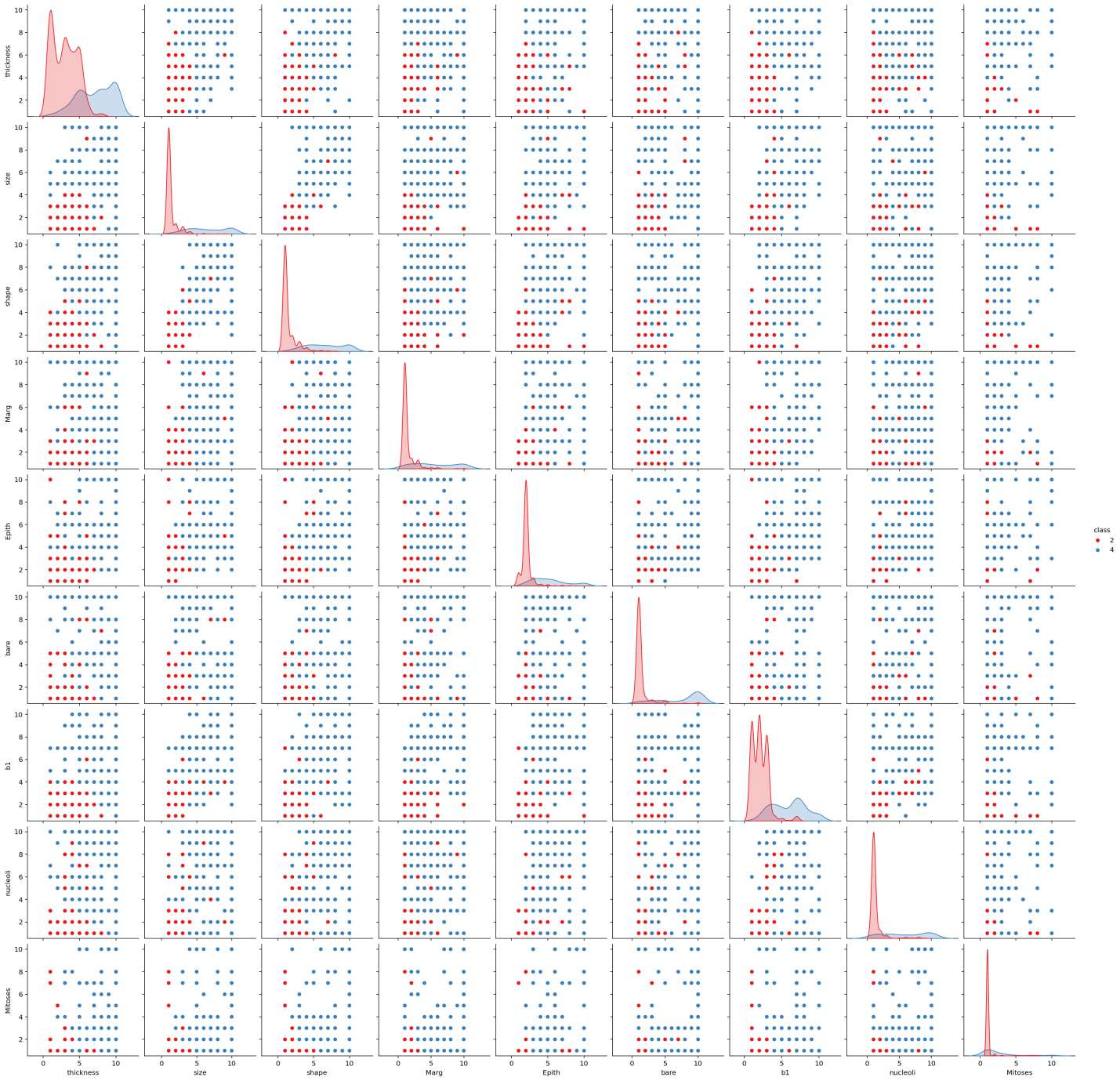


Figure . Pairplot to Visualize Relationships Between Features

* The pairplot shows the relationships between different features, colored by the class (benign or malignant).
* There is a clear separation between benign and malignant classes in several feature pairs, such as thickness vs. size and shape vs. Marg.

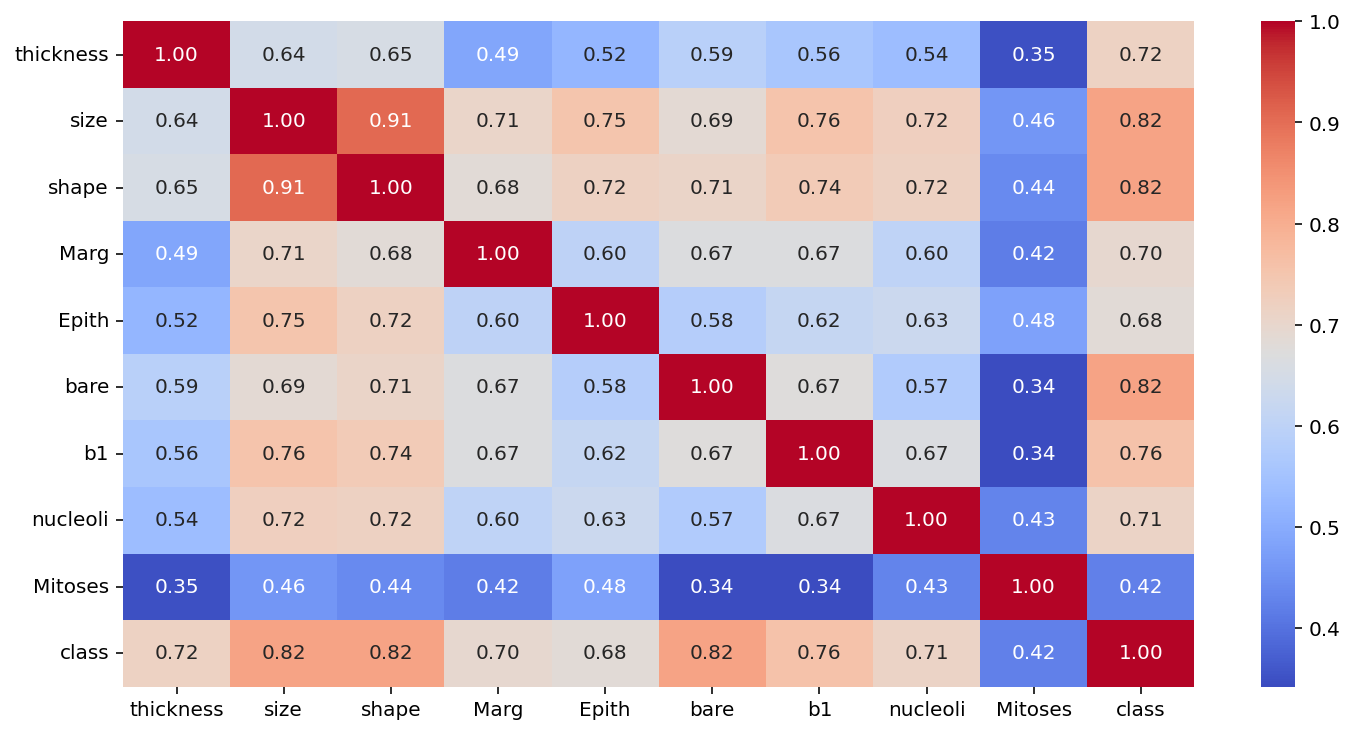


Figure 2:Heatmap to Show Correlation Between Features

* The heatmap shows the correlation between different features.
* Features like thickness, size, and shape have high positive correlations with each other.
* The class column has a strong positive correlation with thickness, size, and shape, indicating that these features are important in distinguishing between benign and malignant tumors.

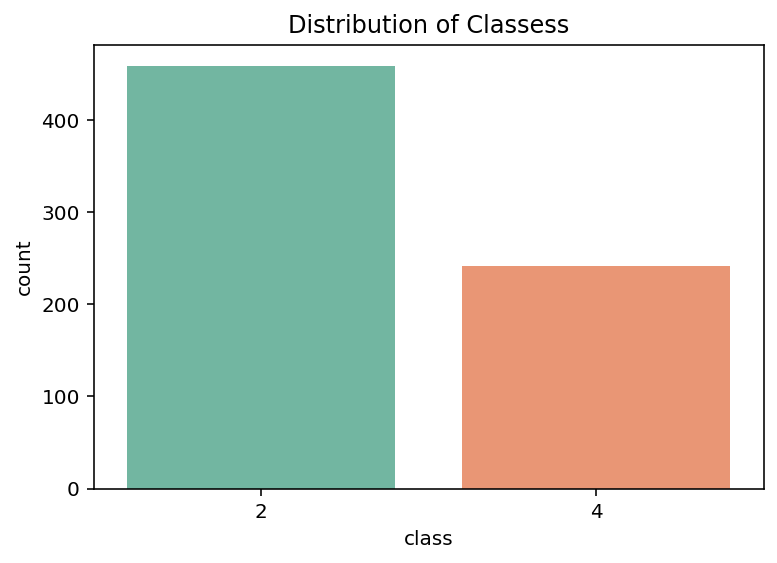


Figure 3: Countplot to Show the Distribution of Classes

* The countplot shows the distribution of benign (class 2) and malignant (class 4) tumors.
* There are more benign cases than malignant cases in the dataset, indicating a class imbalance.

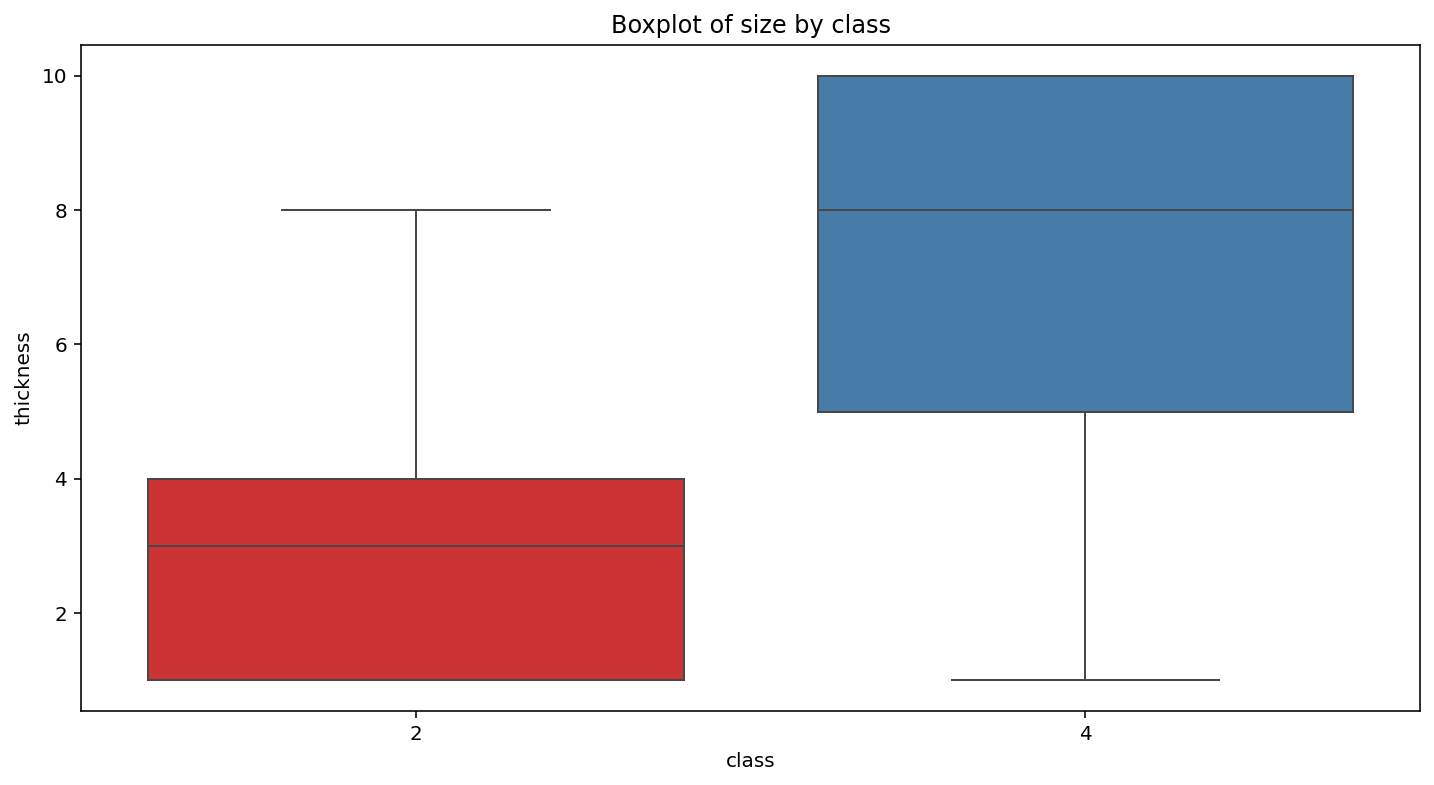


Figure 4:Boxplot to Compare Feature Distributions by Class

* The boxplot shows that malignant tumors (class 4) tend to have higher thickness values compared to benign tumors (class 2).
* This suggests that thickness is a significant feature in differentiating between benign and malignant tumors.

### Examination of Results and Recommendation

The results for the four SVM classifiers with different kernels are as follows:

1. **SVM with Linear Kernel**:
   * **Training Accuracy**: 0.9696
   * **Testing Accuracy**: 0.9714
   * **Confusion Matrix**:[[86,3],[1,50]][[86,3],[1,50]]
   * **Classification Report**: High precision, recall, and F1-scores for both classes (2 and 4).
2. **SVM with RBF Kernel**:
   * **Training Accuracy**: 0.9732
   * **Testing Accuracy**: 0.9714
   * **Confusion Matrix**:[[86,3],[1,50]][[86,3],[1,50]]
   * **Classification Report**: Similar performance to the linear kernel, with high precision, recall, and F1-scores.
3. **SVM with Polynomial Kernel**:
   * **Training Accuracy**: 0.9785
   * **Testing Accuracy**: 0.9643
   * **Confusion Matrix**:[[88,1],[4,47]][[88,1],[4,47]]
   * **Classification Report**: Slightly lower testing accuracy compared to linear and RBF kernels, but still strong performance.
4. **SVM with Sigmoid Kernel**:
   * **Training Accuracy**: 0.4633
   * **Testing Accuracy**: 0.4000
   * **Confusion Matrix**:[[56,33],[51,0]][[56,33],[51,0]]
   * **Classification Report**: Poor performance, with very low precision, recall, and F1-scores, especially for class 4.

### Recommendation

Based on the results, I would recommend using the **SVM with the RBF kernel** or the **SVM with the linear kernel**. Both models achieve the highest testing accuracy of **97.14%**, with excellent precision, recall, and F1-scores for both classes. The RBF kernel has a slightly higher training accuracy (97.32%) compared to the linear kernel (96.96%), but both perform equally well on the testing set.

The **polynomial kernel** also performs well but has a slightly lower testing accuracy (96.43%) compared to the linear and RBF kernels. The **sigmoid kernel** performs poorly and is not suitable for this dataset.

# Assignment 3– exercise2:

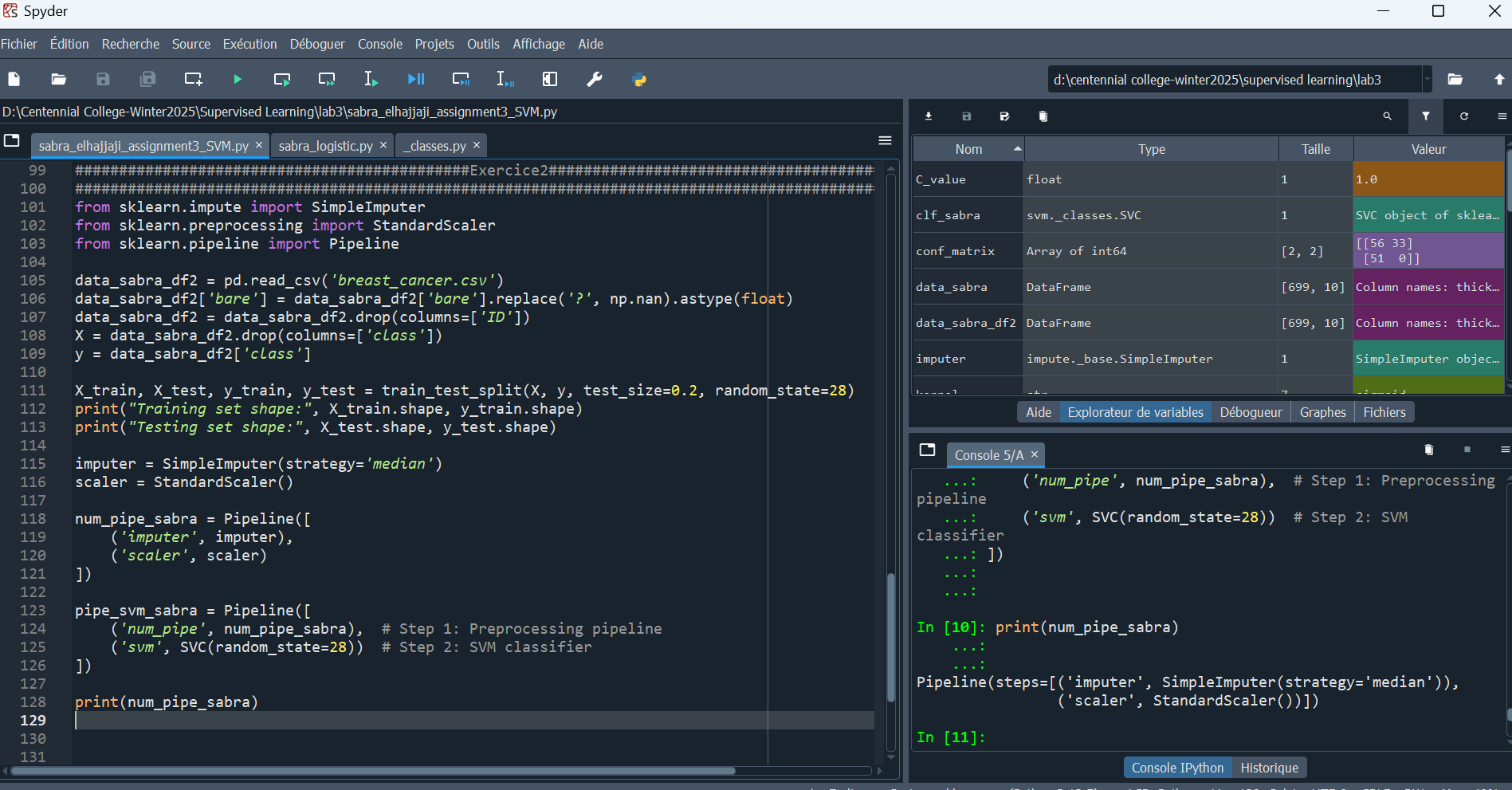


Figure 5: screenshot showing my num\_pipe\_sabra

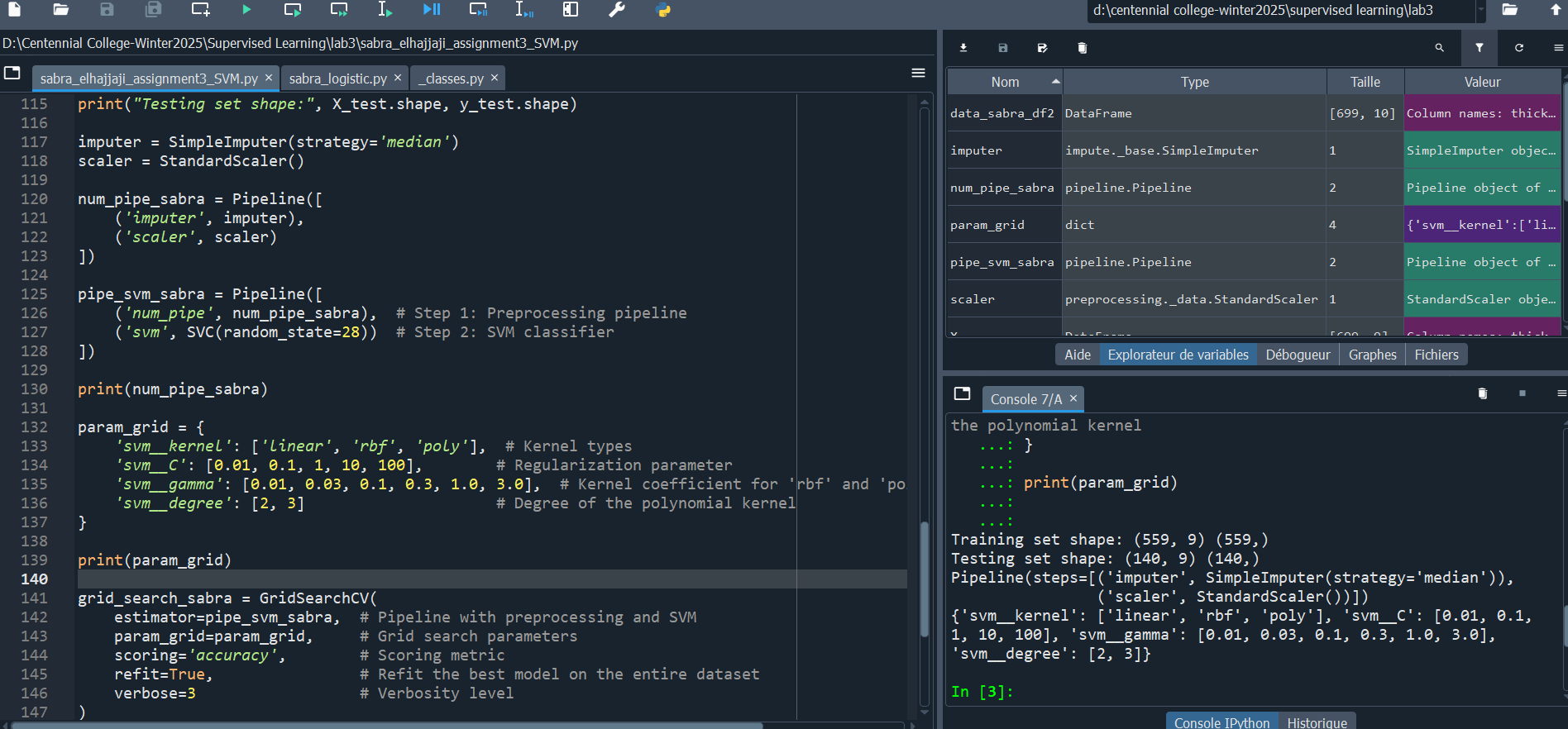


Figure 6: Screenshot showing my param\_grid

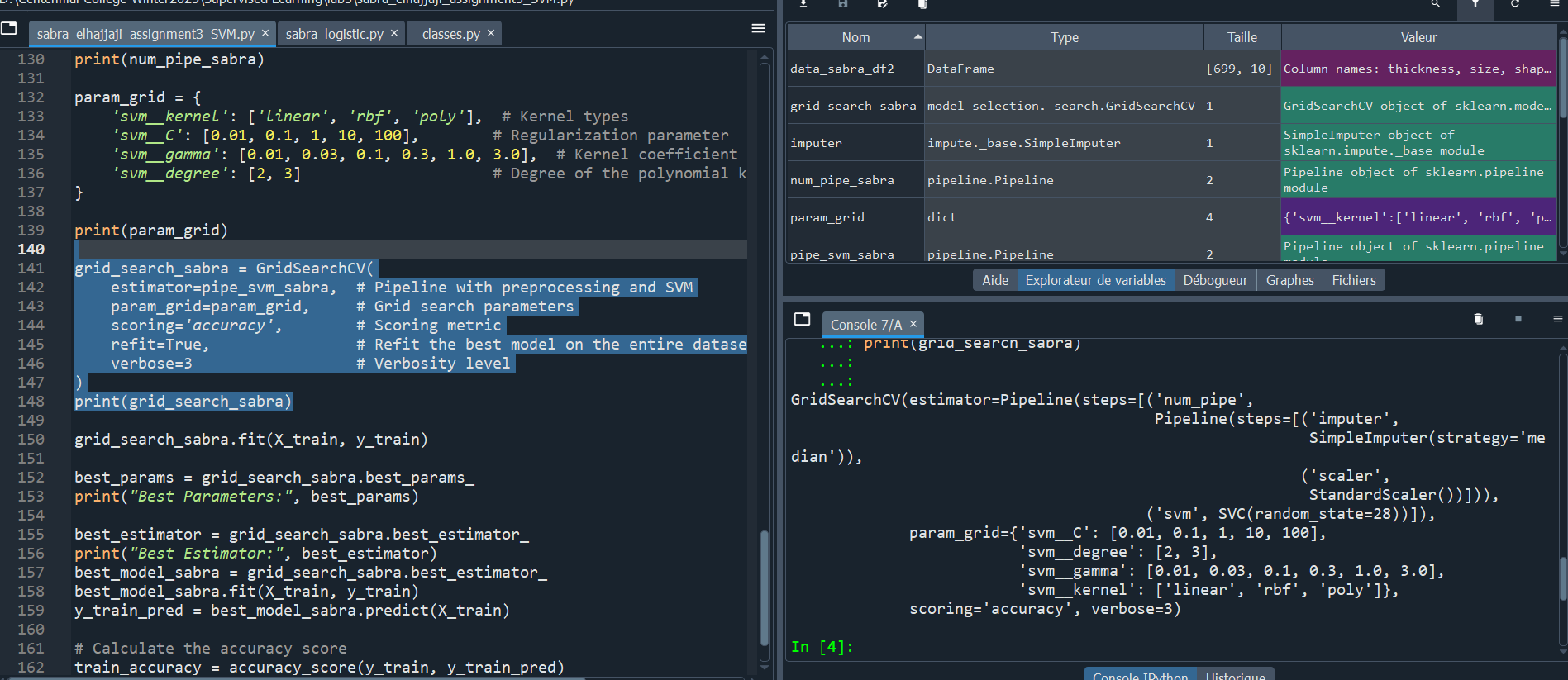


Figure 7:Screenshot showing grid search

### Comparison of Results and Conclusions

##### Results from Exercise #2:

* **Best Parameters**: {'svm\_\_C': 1, 'svm\_\_degree': 2, 'svm\_\_gamma': 0.03, 'svm\_\_kernel': 'rbf'}
* **Best Estimator**: A pipeline combining preprocessing (SimpleImputer and StandardScaler) and an SVM model with the above parameters.
* **Training Accuracy**: **96.96%**

##### Comparison with Exercise #1:

* In **Exercise #1**, we manually trained SVM models with different kernels and evaluated their performance. The best model achieved a testing accuracy of **97.14%** with the RBF kernel.
* In **Exercise #2**, we used a pipeline and grid search to automate preprocessing and hyperparameter tuning. The best model achieved a training accuracy of **96.96%**, which is slightly lower than the testing accuracy in Exercise #1. However, this difference is negligible and likely due to the randomness in the data split or the grid search process.

##### Main Differences Between Exercise #1 and Exercise #2:

1. **Automation**:
   * **Exercise #1**: Manual training and evaluation of SVM models with different kernels.
   * **Exercise #2**: Automated preprocessing and hyperparameter tuning using a pipeline and grid search.
2. **Hyperparameter Tuning**:
   * **Exercise #1**: Hyperparameters were manually selected (e.g., C=0.1 for the linear kernel).
   * **Exercise #2**: Grid search was used to find the optimal hyperparameters (C=1, gamma=0.03, kernel='rbf').
3. **Scalability**:
   * **Exercise #1**: Suitable for small datasets and simple models but not scalable for larger datasets or more complex workflows.
   * **Exercise #2**: The pipeline and grid search approach is scalable and can handle larger datasets and more complex models efficiently.

#### Conclusions:

* The **pipeline and grid search approach** in Exercise #2 is more efficient and robust compared to the manual approach in Exercise #1.
* The best model in Exercise #2 achieved a training accuracy of **96.96%**, which is comparable to the testing accuracy of **97.14%** in Exercise #1.
* The grid search helped identify the optimal hyperparameters (C=1, gamma=0.03, kernel='rbf'), which improved model performance.
* The pipeline approach is highly recommended for real-world machine learning tasks due to its scalability, reproducibility, and efficiency.