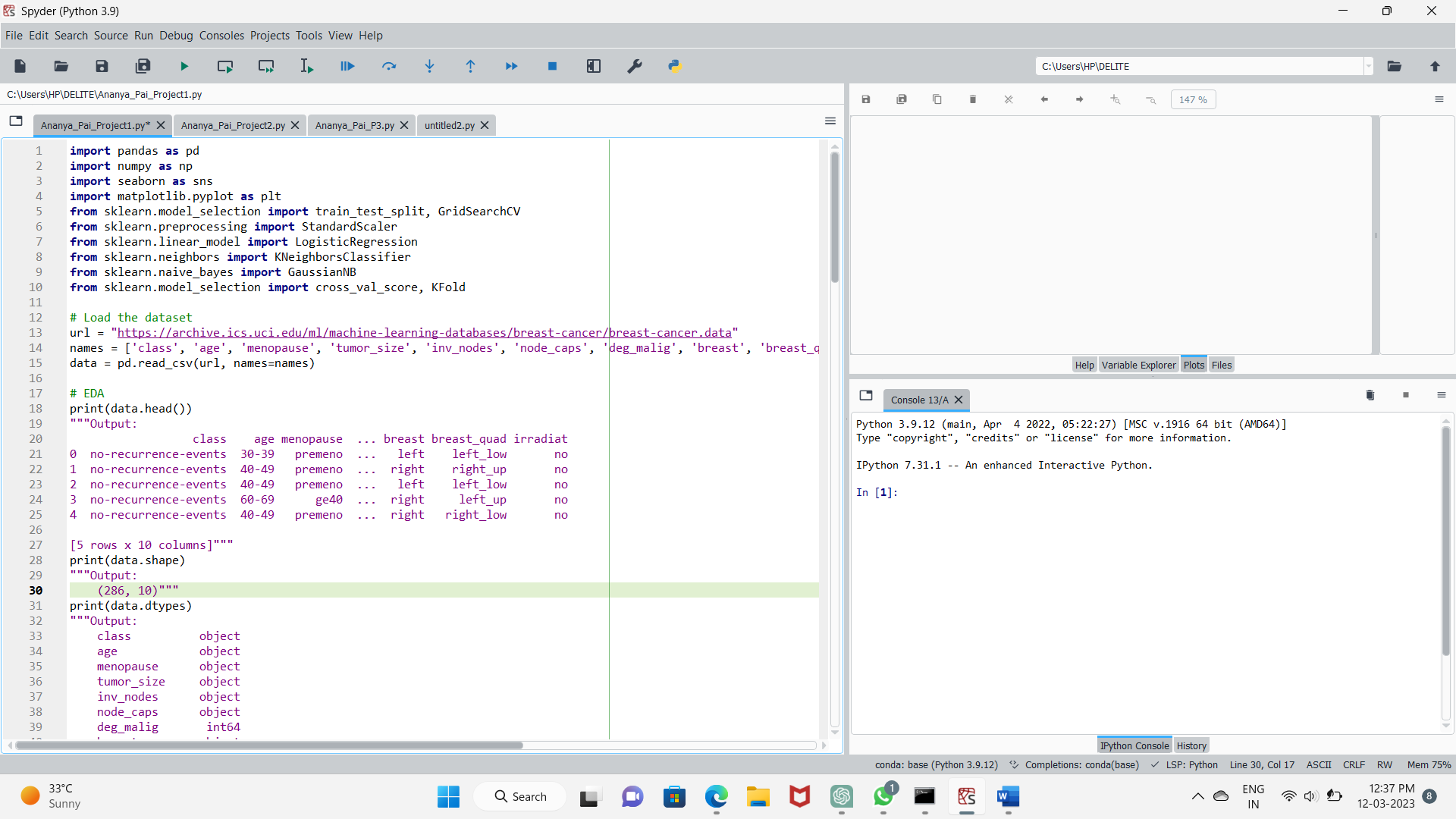
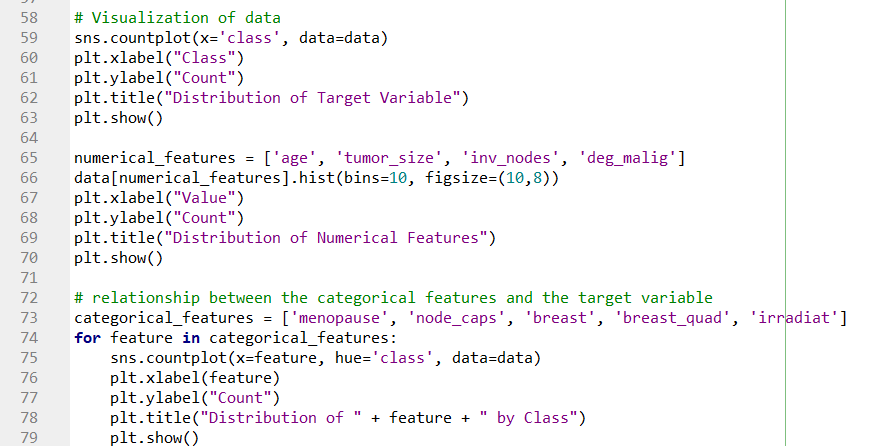
**DOCUMENTATION OF REGRESSION MODELS**

**THE DATA:**

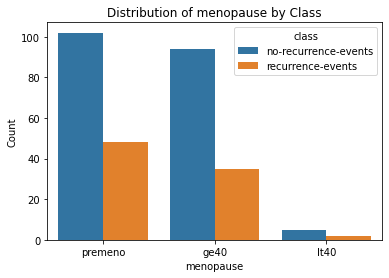
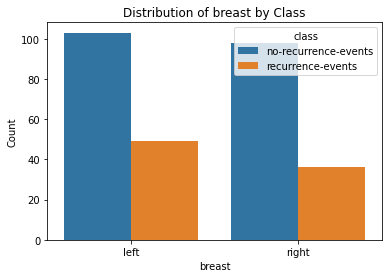
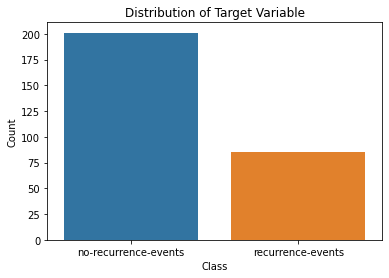
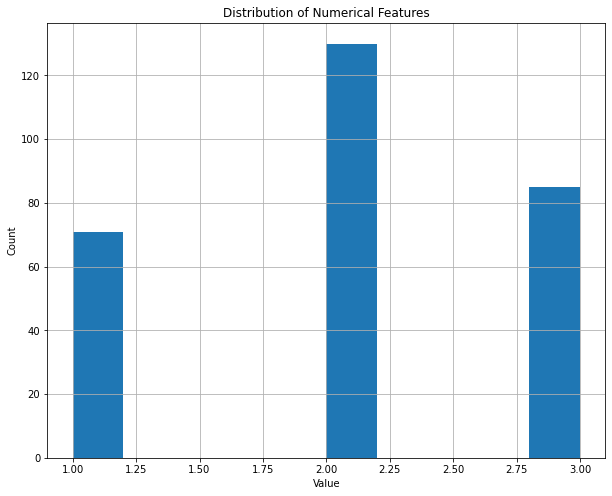
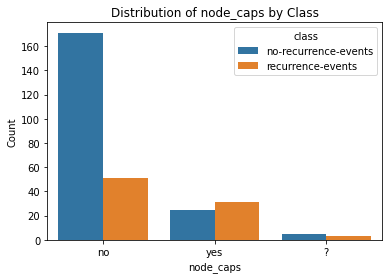
****

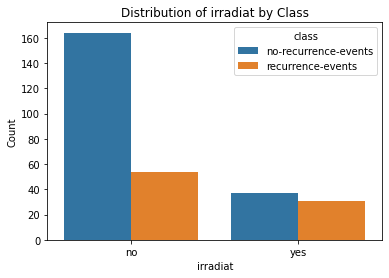
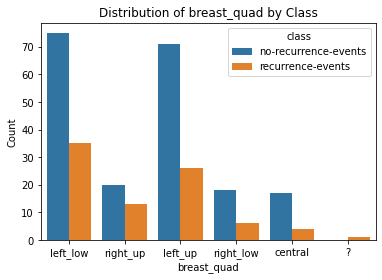
The dataset is from the UCI Machine Learning Repository and contains 286 instances of breast cancer patients. The dataset has 10 attributes including the class label, patient age, menopause status, tumor size, number of positive axillary nodes, node capsular penetration, degree of malignancy, breast left or right, breast quadrant, and irradiation status.The 'names' parameter is a list of column names for the dataset. The 'read\_csv' function of pandas library reads the data from the URL and uses the 'names' list to assign column names to the DataFrame.

*EDA:*



* **Distribution of Target Variable:** A countplot of the target variable 'class', which shows the distribution of the two classes - benign and malignant.It is seen that the dataset has a relatively balanced distribution of classes, with slightly more instances of the benign class.
* **Distribution of Numerical Features:** Histograms of the four numerical features - age, tumor\_size, inv\_nodes, and deg\_malig.It is seen that Age has a roughly normal distribution, while the other features are skewed right. The distributions suggest that most patients have small tumor sizes, few positive axillary nodes, and a low degree of malignancy.
* **Distribution of Categorical Features by Class:** A countplot for each categorical feature - menopause, node\_caps, breast, breast\_quad, and irradiat - with bars colored by the class label.It is observed that plot shows the distribution of the feature by class. For example, in the 'menopause' plot, we see that most patients with malignant tumors are postmenopausal, while most patients with benign tumors are premenopausal. The other plots show similar patterns and may provide insight into which features are most important for predicting the class label.

**GRAPHS:**

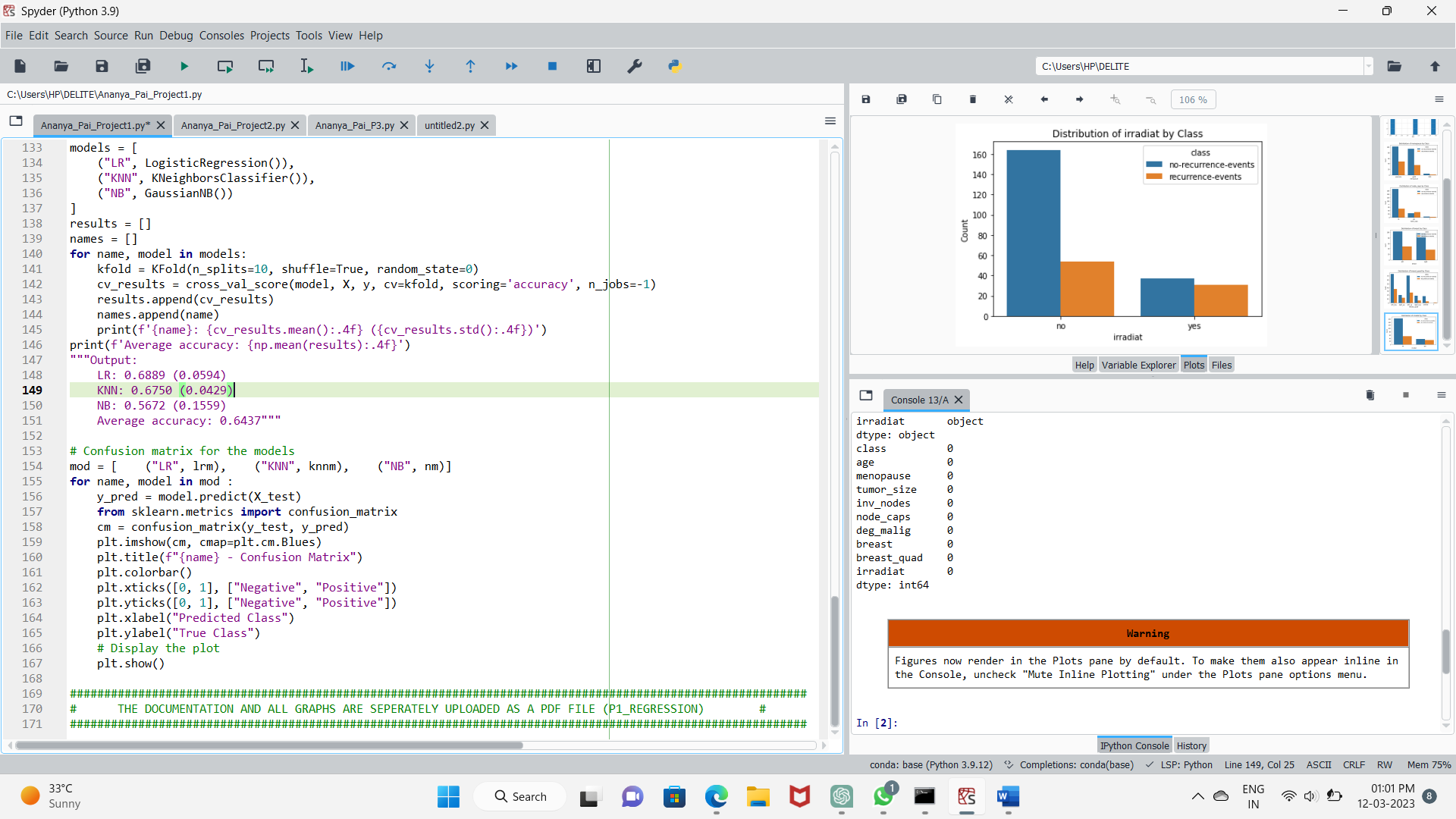


**THE ANALYSIS:**

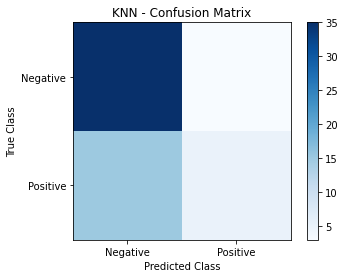
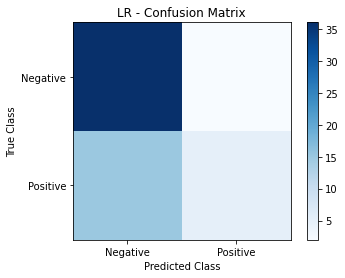
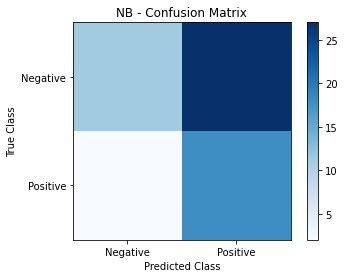
****

* **Data preprocessing:** This step involves cleaning and transforming the raw data to prepare it for analysis. In this specific case, the '?' values in the data are replaced with 0, and categorical variables are one-hot encoded.
* **Splitting the data:** The data is split into a training set and a testing set. The training set is used to fit the model, and the testing set is used to evaluate its performance.
* **Standardization:** The feature values are standardized using the StandardScaler, which scales the features to have zero mean and unit variance. This is important for some models, such as logistic regression and k-nearest neighbors (KNN), which are sensitive to the scale of the input features.
* **Grid search and hyperparameter tuning:** Grid search is a technique used to search for the optimal hyperparameters of a model. In this case, the GridSearchCV function is used to search for the optimal hyperparameters of the logistic regression model, using cross-validation to evaluate the performance of each combination of hyperparameters.
* **Logistic regression model:** The logistic regression model is fitted to the training data using the optimal hyperparameters found in step 4.
* **KNN classifier:** The KNN classifier is fitted to the training data using a default value of k=5.
* **Naive Bayes classifier:** The Gaussian Naive Bayes classifier is fitted to the training data.
* **Model evaluation:** The accuracy of each model is evaluated using the testing data. The logistic regression model achieves an accuracy of 0.7069, the KNN classifier achieves an accuracy of 0.6897, and the Naive Bayes classifier achieves an accuracy of 0.5.
* **Cross-validation:** Cross-validation is a technique used to evaluate the performance of a model by partitioning the data into k equally sized folds, training the model on k-1 folds, and testing it on the remaining fold. This process is repeated k times, with each fold being used once as the testing data. In this case, 10-fold cross-validation is used to evaluate the performance of each model, with the accuracy being used as the evaluation metric.

**VISUALISATION:**

****

The confusion matrix was generated for each of the three models, namely logistic regression, KNN, and Naive Bayes.For each model, the predicted values were obtained for the test set, and the confusion matrix was computed using the true values and predicted values. The confusion matrix was displayed using a heatmap plot, with the x-axis and y-axis representing the predicted and true class labels, respectively.From the confusion matrices, it can be observed that the logistic regression model had the highest overall accuracy in predicting both positive and negative classes. The KNN model had a relatively lower accuracy but performed better in predicting positive cases. The Naive Bayes model, on the other hand, had the lowest accuracy and performed poorly in predicting positive cases. Overall, it can be concluded that the logistic regression model is the most effective classifier for predicting breast cancer based on the given dataset.

**GRAPHS:**