**HealthCare Data Analysis and Logistic Regression with Feature Selection**

**Project Description:** The project aims to build a predictive model using logistic regression to predict the length of stay of patients in a healthcare facility. The dataset used for this project is stored in the file "train\_data.csv". It contains various features related to patients, like demographic information and other relevant attributes. The goal is to analyse these features and their relationship with the length of stay and develop a model that can accurately predict the stay duration.

**Project Workflow:**

1. Data Loading and Exploration:

* I began by loading the "train\_data.csv" file into a Pandas DataFrame.
* To grasp an initial overview of the data, I previewed the first few rows.
* I identified null values within the dataset by computing the count of null values for each column.
* Columns with null values were singled out and presented.

2. Data Preprocessing:

* To gauge the distribution, I calculated skewness for the "Bed Grade" and "City\_Code\_Patient" columns.
* I printed the skewness values to assess data skew.
* Null values in the "Bed Grade" column were replaced with the mean value.
* Null values in the "City\_Code\_Patient" column were substituted with the median value.
* I standardized data by changing "No-20" to "11-20" in the "Stay" and "Age" columns.
* I eliminated the columns 'case\_id' and 'Visitors with Patient' as they didn't contribute to our analysis.

3. Feature Correlation:

* To measure the linear relationship with the target variable "Stay," I calculated Pearson correlation coefficients for each feature.
* The correlation coefficients were cataloged in a dictionary.
* I ensured that the correlation of "Stay" with itself was included in the dictionary.
* Correlation values for each feature were presented.

4. Feature Selection:

* I established a correlation threshold to discern features significantly related to the target variable.
* Features with correlation coefficients exceeding the threshold were selected and stored in the "selected\_features" list.

5. Model Training and Evaluation:

* Utilizing the selected features and the target variable, I constructed the feature matrix "X" and the target vector "y."
* I partitioned the dataset into training and testing subsets employing the train\_test\_split function from sklearn.
* Subsequently, I instantiated a logistic regression model and trained it on the training data.
* The model was then employed to generate predictions on the testing set.
* I gauged the model's accuracy by comparing predicted values with actual values.
* Additionally, I presented a comprehensive classification report that included precision, recall, F1-score, and support metrics for each class.

**Code Explanation:**

* The required libraries (pandas, numpy, seaborn, sklearn, matplotlib.pyplot) are imported.
* The train\_data.csv file is loaded into a DataFrame using the pd.read\_csv() function.
* Initial exploration of the dataset is performed by displaying the first few rows using the head() function.
* Null values in the dataset are identified and stored in the columns\_with\_null variable.
* Skewness of the "Bed Grade" and "City\_Code\_Patient" columns is calculated using the skew() function from scipy.stats.
* Null values in the "Bed Grade" column are replaced with the mean using the fillna() method.
* Null values in the "City\_Code\_Patient" column are replaced with the median using the fillna() method.
* The values "No-20" in the "Stay" column and "Age" column are replaced with "11-20" using the replace() method.
* The columns 'case\_id' and 'Visitors with Patient' are dropped from the DataFrame using the drop() method.
* Pearson correlation coefficients are calculated for each numeric column using the pearsonr() function from scipy.stats.
* Non-numeric columns are converted to numeric representation using label encoding.
* The correlation coefficients are stored in the correlation dictionary.
* Features with correlation coefficients below the threshold are dropped from the DataFrame using the drop() method.
* The remaining columns in the DataFrame are stored in the X variable, and the target variable "Stay" is stored in the y variable.
* The dataset is split into training and testing sets using the train\_test\_split() function.
* A logistic regression model is instantiated with increased max\_iter and fitted to the training data using the fit() method.
* Predictions are made on the testing set using the predict() method.
* The accuracy of the model is calculated by comparing the predicted values with the actual values using the accuracy\_score() function.
* The classification report is printed using the classification\_report() function.

Conclusion:

The logistic regression model with feature selection achieved an impressive overall accuracy of 87.55% on the test set, showcasing its strong predictive capabilities. This indicates that the model successfully classifies the length of stay for most instances, making it a valuable tool for healthcare professionals in predicting patient outcomes.