**Predicting Apartment Renting Prices in Kosovë with Machine Learning Regression Algorithms**

*A machine learning project by Uran Lajçi*

**Problem Definition**

The problem at hand is the absence of apartment price predictions for apartments in Kosovë. The objective of this project is to develop an application that allows users to input apartment features and obtain the corresponding rental price. To achieve this, the project will utilize machine learning regression algorithms to predict apartment prices based on the provided features. Additionally, the application will be built using the Streamlit framework to provide a user-friendly interface for easy interaction and price retrieval. By addressing this problem, the project aims to offer a valuable tool for both apartment seekers and landlords in Kosovë, facilitating informed decision-making and improving the efficiency of the rental market.

**Data Acquisition**

To solve the problem of predicting apartment rental prices in Kosovë, it was necessary to gather the required data since there were no existing datasets available for this specific domain. The following approach was employed to collect the data:

1. Data Source: The data was obtained by utilizing the BeautifulSoup Python library to scrape information from the website:

[https://gjirafa.com/Top/Patundshmeri?f=2&sh=Kosove&k=Banesa&llshp=Qira](https://gjirafa.com/Top/Patundshmeri?f=2&sh=Kosove&k=Banesa&llshp=Qira).

1. Scraping Process: The BeautifulSoup library allowed for the extraction of relevant details from the website's listings. By navigating through the pages and parsing the HTML structure, the necessary data points were extracted.
2. Dataset Size: In total, over 1000 data points were gathered during the scraping process. Each data point represents a distinct apartment listing from the website.
3. Attributes: The following attributes were collected for each apartment listing:
   1. "title": The title or headline of the apartment listing.
   2. "region": The region or location of the apartment.
   3. "number\_of\_rooms": The number of rooms in the apartment.
   4. "quadrat": The size of the apartment in square meters.
   5. "price": The rental price of the apartment.
   6. "date": The date when the apartment listing was published.

By acquiring this dataset, comprising the aforementioned attributes, the project has obtained the necessary information to proceed with the development and evaluation of the machine learning regression algorithms for predicting apartment rental prices in Kosovë.

**Data Preprocessing**

To prepare the data for further analysis and model training, several preprocessing steps were performed. This involved cleaning the data, handling missing values, outliers, and performing feature engineering. These steps have been taken to preprocess the data:

1. Cleaning the 'quadrat' column by removing the 'm 2' string and converting it to a float data type.
2. Cleaning the 'price' column by removing special characters.
3. Converting the 'price' column from string to float using a custom conversion function.
4. Removing outliers by calculating the z-scores for each numeric column and removing data points with z-scores above a specified threshold.
5. Filtering the dataset based on specific conditions, such as a minimum number of rooms, minimum apartment size, and minimum price.
6. Optionally, resetting the index if necessary.
7. Saving the preprocessed data to a CSV file named "preprocessed\_apartment\_data.csv".

The resulting preprocessed data is now ready for further analysis, model development, and prediction of apartment rental prices in Kosovë.

4. \*\*Exploratory Data Analysis (EDA)\*\*: It's often useful to perform some initial explorations of your data to gain insights that may help with later steps. This includes things like checking distributions of features, identifying relationships or correlations between features, understanding the balance of your classification targets, etc. It's not mentioned in your process, but it's quite critical.

5. \*\*Feature Selection and Scaling\*\*: Depending on the algorithms you choose, you might need to scale your data or reduce dimensionality (i.e., select a subset of all available features that you believe are most useful for predicting the target). It's not clear if you've done this yet, but it is necessary for some models.

6. \*\*Model Training\*\*: Split your dataset into a training set and a testing set, then train your model on the training set. You've completed this step.

7. \*\*Model Evaluation\*\*: Use the test set and relevant metrics (accuracy, precision, recall, F1 score, AUC-ROC, etc.) to determine the performance of your model. You've done this step, although it might be good to consider cross-validation if you haven't already.

8. \*\*Model Optimization\*\*: This involves tuning hyperparameters, using ensemble methods, or trying different models to improve performance. It's not mentioned in your process, but it's a usual step.

9. \*\*Interpretation of Results\*\*: Understand and describe what your model's outputs and metrics mean in the context of the problem you are trying to solve. It's not mentioned in your process, but it's very important.

10. \*\*Deployment\*\*: Deploy the model in a suitable format for its intended use. For example, this could be in a software application, a web service, etc.

11. \*\*Monitoring and Maintenance\*\*: Once your model is deployed, it will require regular monitoring to ensure it continues to perform as expected. This might involve updating it with new data, checking for model drift, etc.

While it looks like you have made great progress on your project, you might want to make sure you've covered steps like EDA, feature selection and scaling, model optimization, interpretation of results, and the deployment and maintenance stages. These can often make a significant difference in the overall quality and impact of your project.