***Documentation of SUML***

***Predicting the price for AirBNB***

***Acknowledgments***

***This Project is done by:  
-Konrad Maliszewski S23174***

***-Anastasiya Sukach S23566***

***-Abdelrahman Mobarak S22878***

**1.Project Overview**

***The Airbnb Price Prediction project centers on predicting rental prices using machine learning models based on diverse datasets from major cities. The primary goal is to provide accurate price estimates for potential hosts and travelers. The project encompasses various stages, including data merging, cleaning, model selection, and the development of a user-friendly interface using Streamlit.***

**2.Data Sources**

***After exploring and searching for suitable datasets, we have found these datasets on Kaggle platform, (Airbnb Prices in European Cities).***

**3.Data Description**

***This dataset conducts a detailed examination of Airbnb prices in popular European cities, assessing various attributes such as room type, cleanliness, satisfaction ratings, bedrooms, and distance from the city center. Using spatial econometric methods, it identifies determinants of pricing on both weekdays and weekends, shedding light on global market dynamics and pricing strategies.***

**4.Data Merging**

***The data merging phase involved the consolidation of datasets from key cities, such as Amsterdam, Athens, Barcelona, and others. By merging both weekday and weekend datasets, a comprehensive dataset was created. The addition of a 'city' column allows for geographical distinctions, contributing to a holistic understanding of Airbnb pricing trends across different locations.***

**5.Data Cleaning**

***In the data cleaning process, meticulous steps were taken to ensure dataset quality. Handling missing values, encoding categorical variables, and defining room types facilitated effective data preparation. The removal of unnecessary columns and outliers resulted in a refined dataset, laying a solid foundation for accurate model training.***

**6.Model Selection  
  
*The project involved the evaluation of various machine learning models, including Linear Regression, Random Forest, Decision Tree, SVR, and Gradient Boosting. After careful consideration, Random Forest emerged as the chosen model due to its superior performance, ensemble learning capabilities, and potential for hyperparameter tuning.***

**7.Model Evaluation**

***A comprehensive analysis of model performance metrics was conducted, revealing that Random Forest outperformed other models. Its ability to provide accurate predictions and handle complex datasets solidified its position as the primary model for Airbnb price prediction.***

**8.Hyperparameter Tuning**

***Utilizing RandomizedSearchCV, hyperparameter tuning was performed to optimize the Random Forest model. The best parameters, including the number of estimators, minimum samples split, and maximum depth, were identified. This fine-tuning contributed to improved predictive capabilities.***

**9.User Interface with Streamlit**

***The user interface was developed using Streamlit, a Python tool for web application development. This interface enhances user interaction, providing a seamless experience for individuals seeking Airbnb price estimates. Its intuitive design accommodates a diverse user base.***

**10.Team Collaboration and Development**

***Effective communication and collaboration were paramount throughout the project. The team utilized Discord for communication and Trello for task assignment and tracking. Roles were distributed, with Abdelrahman focusing on data merging and documentation, Konrad on data cleaning and UI development, and Anastasiya on model training and selection.***

***11.Conclusion***

***This documentation provides a detailed explanation of the codebase for Airbnb price prediction. It covers data merging, cleaning, application of machine learning models, and optional tuning for improved performance. Further analysis and insights can be derived by exploring the visualizations and evaluation metrics presented in the code.***

***12. References***

***Data Sources***

[**https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities**](https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities)

***Our Data***

[**https://github.com/Tykher/AutoML\_group\_3\_AirBnB**](https://github.com/Tykher/AutoML_group_3_AirBnB)

***Full project***[**https://colab.research.google.com/drive/1jkgcfAeSexOjqgvP4tFncB63HRWMpvDQ?usp=sharing#scrollTo=EPL9raZ8usR1**](https://colab.research.google.com/drive/1jkgcfAeSexOjqgvP4tFncB63HRWMpvDQ?usp=sharing#scrollTo=EPL9raZ8usR1)

***13.Codebase***

***pandas: A powerful data manipulation and analysis library for Python. Link to pandas***

***scikit-learn: A machine learning library that provides simple and efficient tools for data analysis and modeling. Link to scikit-learn***

***matplotlib: A comprehensive library for creating static, animated, and interactive visualizations in Python. Link to matplotlib***

***numpy: A fundamental package for scientific computing with Python. Link to numpy***