Defne Yaz Kılıç-22102167
Boran Kılıç-22103444

EEE485 Statistical Learning and Data Analytics
Project Proposal

Introduction
In this project, we are going to predict the prices of Airbnb houses based on the properties available in the chosen dataset. We will use python to implement our algorithms. Linear regression, decision tree and neural network are the methods that we wish to use.

Project Details
Throughout the project, several machine learning methods will be used and compared with each other in order to choose the best algorithm for our purposes. It was decided to use linear regression as the first method since it is a simple yet effective and widely used method for prediction and classification. If needed, this method can be expanded by adding penalty terms and implementing ridge regression. Feature expansion can also be added during the process for capturing more information beyond a linear relation. Decision tree is chosen as the second algorithm due to its popularity and effectiveness of laying out all the possible outcomes. As for the last method, neural network will be implemented.

Dataset Information
We intend to utilize a dataset called the "Airbnb Price Dataset," sourced from Kaggle.com, a platform offering freely accessible datasets. It contains 75000+ data instances with 29 features listed as follows: id, log_price, property_type, room_type, amenities, accommodates, bathrooms, bed_type, cancellation_policy, cleaning_fee, city, description, first_review, host_has_profile_pic, host_identity_verified, host_response_rate, host_since, instant_bookable, last_review, latitude, longitude, name, neighbourhood, number_of_reviews, review_scores_rating, thumbnail_url, zipcode, bedrooms, beds. It is available on the following link:

https://www.kaggle.com/datasets/rupindersinghrana/airbnb-price-dataset

Expected Challenges
The hardest part of this project would probably be the restrictions about the library usage, since we will create our own algorithms the process will be hard but instructive at the same time. Also since the dataset has many different features choosing the appropriate ones for accurate predictions will be a time consuming task.