**Executive Summary**

I would say, this was one of the fun labs to work on. By far the best that I have ever worked on throughout my master’s journey! It taught me patience, smart googling, and how to run code using cmd and allowed me to explore other alternative ways like Jupyter Notebook.

In this homework, firstly I saw the dataset and understood what each column means. Then checked if there were any values missing with the help of feature statistics and understood that this data was clean only since there was no further cleaning processing required, I ended up implementing the models around this dataset – since the requirement was to split the price data in the interval of $1,000,000. I was able to do this with the help of discretize node. After which, the data was loaded in a data table to see if everything is still intact and the changes for pricing range were seen in the column. Then comes the sampling part where to sample the data I used a Fixed proportion of data as the sampling type with 70% data to be used for training purposes and then applied that stream to multiple predictive model nodes like Tree, Logistic Regression, Neural networks, Random Forest, Naïve Bayes, SVM, etc. I think this was the moment wherein I could easily use and name so many predictive models in a single go! Later the outcomes of these models were connected to test and score nodes to evaluate and finalize the model which is performing best among them. Since the accuracy with 70% training data was not nice so I ended up increasing the sample data load to 80% and then reran the models where the accuracy was captured better than the previous attempt. I also tried to change the options available on the left pane of test and score node like changing the number of folds, made sure I try the random sampling radio button by playing around with the repeat test and training set size values and then capturing what changes are happening in the accuracy values. After all these trail and errors – I finalized to go with the decision tree model which gave me the highest accuracy of 0.421, once the model was finalized I had to save this model in a pickle file – then as mentioned in the writeup I used this save model feature for the first time in orange tool.

Now that the pickled file was stored in the folder, I moved on to the webapp part. In this, I had to first install a couple of libraries. There were a couple of errors that I encountered as part of the installation and tried to resolve them as per the guidelines given in the writeup. But still, the error was occurring, later I understood it was a very basic mistake that I missed the part from the writeup to make sure the Python version was 3.9 and not 3.11 so then I went on and uninstalled the version I had on my machine and installed 3.9.11 then tried to install the libraries gradio and orange3 (I know professor mentioned to visit him in case I had an issue running two python versions, I tried creating a conda environment as well on my machine with 3.9.16 but it did not work and since I wanted to finish this up over the weekend, my apologies!) – later I ran the webapp script and went on to the URL to see how the UI looks, put in the values provided as part of the exercise and took a screenshot.

Then I went ahead and made changes in the orange workflow – to make sure that prefera column is not skipped during the modeling and reran the flow with other models – check the accuracy as well. But it did not give any significant changes in the value so I stuck to the tree model only and saved the changes in the pickled file. Later as part of the change made in the workflow, I had to change the script as well, by adding a new variable for the preferred area column, once the script was ready I ended up running the script and then on the URL with the required inputs took a screenshot for the report.

This project will help the business to understand the pricing of a certain requirement from the client and it can also help the clients to set up a budget for their new house. This would end up saving a lot of time for both parties. It will lead to increased customer engagement and if we are able to store the hits made by users will be able to also see some patterns to understand which configurations are in demand. This will also give the business an advantage to capture large market with advanced technology.

**Screenshots from Web Application:** (Without Preferred Area)

a. Bedrooms: 3

b. Bathrooms: 2

c. Main road: Yes

d. Guest room: No

e. Basement: No

f. Hot water: Yes

g. Air conditioning: No

h. Parking spots: 2

i. Furnished: Semi-furnished

Graphical user interface, application

Description automatically generated

(With Preferred Area)

a. Bedrooms: 2

b. Bathrooms: 2

c. Main road: No

d. Guest room: Yes

e. Basement: No

f. Hot water: Yes

g. Air conditioning: Yes

h. Parking spots: 1

i. Prefarea: No

j. Furnished: Unfurnished

Graphical user interface, application

Description automatically generated

**Updated Orange and Python code:**

 