**Project Description**

**Dataset**

The data is downloaded from Kaggle. Its name is “Customer Experience Dataset”. It is a simulated dataset, specially designed for training machine-learning models to optimize customer experience. The data has 1000 rows (excluding the column title row), corresponding to data from 1000 customers. There are 14 columns, and according to Kaggle, they can be divided into five categories:

1. Demographics: Customer\_ID (numerical, continuous), Age (numerical, continuous), Gender(text, categorical), Location (text, categorical)
2. Interaction Data: Num\_Interactions (numerical, continuous), Feedback\_Score (numerical, continuous), Products\_Purchased (numerical, continuous)
3. Behavioral Data: Products\_Viewed (numerical, continuous), Time\_Spend\_on\_Site (numerical, continuous)
4. Satisfaction and Retention: Satisfaction\_Score (numerical, continuous), Retention\_Status (text, categorical)
5. Encoding: Gener\_Encoded (numerical, catagorical), Location\_Encoded, (numerical, catagorical), Retention\_Status\_Encoded (numerical, catagorical)

**Independent and Dependent Variable**

I chose variables that are both numerical and continuous as my interested variables (excluding customer id and age). I would like to use (1) Num\_Interactions, (2) Feedback\_Score, (3) Products\_Purchased, (4) Products\_Viewed, and (5) Time\_Spend\_on\_Site (5 independent variables) to predict Satisfaction\_Score (dependent variable).

**Research question:**

1. Which model best predicts the satisfaction score?
2. Which independent variable is most important regarding the prediction?

**Data transformation**

I fill the potential NA values in the data using the column means. And then I applied Z-score normalization to the models used in the first research question (both test and training data uses the mean and standard deviation of the training data).

**First Research Question**

I chose three models to implement: Multivariant Linear Regression, K-Nearest Neighbors for Regression (KNN), and Multivariant Polynomial Regression. I built the first two model from scratch, and I used the sklearn library to implement the last one (I chose the power of two in the polynomial transformation).

The test performance is assessed by Root Mean Square Error (RMSE) score, with lower scores indicating better performance. The results are: Linear Regression: RMSE = 2.8658289792600367; KNN for Regression: RMSE = 3.3302207690137258; Polynomial Regression RMSE = 2.942011299575351. Based on these results, the Linear Regression model performed the best, and the Polynomial Regression model is not much worse than the Linear Regression model.

**Second Research Question**

In order to decide which independent variable is most important in terms of prediction, I built a Decision Tree algorithm. In a decision tree, for each branch, it is split based on which choice of the independent variable and split at where could best reduce the variance. The variable that reduced the most variant is the most important variable in predicting the satisfaction score.

The results were as follows: 0: 0.103729187548726, 3: 0.2061139724114159, 4:0.21552791102752097, 2: 0.22849904912264762, 1: 0.24612987988968943. Therefore, the second variable is the most important variable in terms of predicting satisfaction score, which is the “Num\_Interactions”. Also, I have tested the performance of this model as well using RMSE, and the results show that it is good, indicating that this model is trustworthy (training RMSE: 2.457866745968065, testing RMSE = 3.316548895830722).

**Limitation**

I only use a single method to access the performance of the models. Also, the data points are two little to use more complex models like neural network or polynomial network with power of higher than 2.