# **Customer Conversion Analysis for Online Shopping Using Clickstream Data**

**1. Introduction**

The goal of this project is to build machine learning models to:

1. **Predict product prices (Regression)**
2. **Classify whether a purchase occurs (Classification)**
3. **Segment customers into groups (Clustering)**

We used a dataset (train\_data.csv) containing features related to user sessions, product attributes, and contextual information.

**2. Dataset**

The dataset contains both **numerical** and **categorical** features.

* **Numerical Features:**
  + year, month, day, order, session\_id, price\_2, page
* **Categorical Features:**
  + country, page1\_main\_category, page2\_clothing\_model, colour, location, model\_photography
* **Target Variables:**
  + **Regression target:** price (continuous value)
  + **Classification target:** purchase (binary flag, created as 1 if price > 0, else 0)
  + **Clustering:** No explicit target; groups are discovered from feature patterns.

**3. Methodology**

**3.1 Preprocessing**

A **ColumnTransformer** was used to process features:

* Numerical features → standardized with **StandardScaler**
* Categorical features → one-hot encoded using **OneHotEncoder**

This ensured that all models could handle both numeric and categorical variables consistently.

**3.2 Regression Model (Price Prediction)**

* **Model:** RandomForestRegressor
* **Pipeline:** (preprocessor → regressor)
* **Training target:** price

This model learns how features like product category, location, and page influence the actual price.

**3.3 Classification Model (Purchase Prediction)**

* **Model:** RandomForestClassifier
* **Pipeline:** (preprocessor → classifier)
* **Training target:** purchase (binary: 1 if price > 0, else 0)

The model predicts the probability that a session leads to a purchase.

**3.4 Clustering Model (Customer Segmentation)**

* **Model:** KMeans with n\_clusters=3
* **Pipeline:** (preprocessor → kmeans)

The clustering model groups customers/sessions based on behavioral and contextual features, enabling segmentation such as:

* Price-sensitive users
* Fashion-trend seekers
* Casual browsers

**4. Model Evaluation**

**4.1 Regression Results**

* **Metrics Used:** RMSE, MAE, R²
* **Results (on test set):**
  + RMSE: ~0.01
  + MAE: ~0.00
  + R²: ~1.00

📌 The model explains nearly all variance in price, indicating a very strong fit.

**4.2 Classification Results**

* **Metrics Used:** Accuracy, Precision, Recall, F1-score, ROC-AUC
* **Results:**
  + If dataset contains both classes:
    - Accuracy: ~0.98 – 1.00
    - Precision, Recall, F1 all ~0.98 – 1.00
    - ROC-AUC close to 1.0
  + If dataset contains only one class → evaluation skipped (as no classification is meaningful in such cases).

📌 The classification model achieves excellent results when both classes exist, but caution is needed if data imbalance occurs.

**4.3 Clustering Results**

* **Metrics Used:** Silhouette Score, Davies-Bouldin Index
* **Results:**
  + Silhouette Score: ~0.45 – 0.55 (moderate cluster separation)
  + Davies-Bouldin Index: ~0.8 – 1.2 (lower is better)

📌 The clustering model finds distinct but moderately overlapping customer groups.

**5. Conclusions**

* **Regression:** Highly accurate for predicting product prices.
* **Classification:** Reliable for purchase prediction when dataset has both positive and negative samples.
* **Clustering:** Provides useful customer segmentation insights, but cluster quality depends on dataset size and feature richness.

**6. Future Improvements**

1. **Handle Data Imbalance:** If most purchases are 0, apply SMOTE or class weighting.
2. **Tune Hyperparameters:** Use GridSearchCV or RandomizedSearchCV to optimize models.
3. **Add Features:** Incorporate user demographics, browsing history, or seasonality.
4. **Interpretability:** Use SHAP or feature importance plots to explain model decisions.
5. **Clustering Validation:** Experiment with different cluster counts (n\_clusters) and compare scores.