Final Report

Introduction

This project focuses on analyzing and modeling coffee sales data from a real-world retail context. The dataset, titled **"DatasetForCoffeeSales2.csv"**, contains **730 rows and 11 columns**, covering transaction data from **January 1, 2023 to September 9, 2024**. It includes various attributes such as **city**, **product category**, **unit price**, **quantity sold**, **discounts**, and **final sales**.
The goal was to build predictive models that estimate the **final sales amount** for each transaction. Understanding what drives sales is crucial for better demand forecasting, inventory management, and pricing strategies.

Methodology

Data Preprocessing

- Parsed the Date column and created new time-based features (like Month).
- Ensured no missing values were present.
- Converted categorical features (e.g., Product, City) using one-hot encoding.
- Feature-target separation:
    - **Features (X)**: Product attributes, location, pricing, quantity, and discount info.
    - **Target (y)**: Final Sales.

Model Training and Evaluation
Six machine learning models were trained:

1. **Linear Regression**
2. **K-Nearest Neighbors (KNN)**
3. **Decision Tree**
4. **Random Forest**
5. **XGBoost**
6. **Support Vector Regressor (SVR)**

All models were evaluated using the following metrics:

- **MAE (Mean Absolute Error)**
- **MSE (Mean Squared Error)**
- **RMSE (Root Mean Squared Error)**
- **R^2 (Coefficient of Determination)**

Results

Model	**MAE**	**RMSE**	**R^2**
Linear Regression	12.91	14.99	-0.05
KNN	13.39	15.81	-0.17
Decision Tree	16.36	20.02	-0.87
Random Forest	13.47	15.92	-0.18
SVR	8.51	10.71	0.25
XGBoost	**7.37**	**8.52**	**0.53**

🔍 **Comparison**

- **XGBoost** clearly outperformed all other models, achieving the **lowest error values** and the **highest R^2 score (0.53)**, meaning it explains 53% of the variance in the target variable.
- **SVR** was also competitive, showing better generalization than traditional models like KNN or Decision Tree.
- Models like **Decision Tree** and **Random Forest** underperformed, potentially due to overfitting or limited feature complexity.
- **Linear Regression** produced a negative R^2, indicating poor performance for the given nonlinear data structure.

Discussion & Future Work

This project demonstrates how various machine learning models can differ in effectiveness based on the nature of the dataset. The results show that **ensemble methods** like **XGBoost** are highly effective for sales prediction in structured datasets.

Limitations

- The dataset is relatively small (730 rows), which may limit model generalizability.
- Some useful external variables (e.g., seasonal trends, holidays, marketing spend) were not included.

Future Improvements

- Enrich the dataset with **external features** like public holidays, weather conditions, or regional events.
- Apply **hyperparameter tuning** (e.g., GridSearchCV) for all models to further optimize their performance.
- Use **time series models** like ARIMA or Prophet if sequential trends are to be captured.
- Deploy the best model (XGBoost) in a **real-time dashboard** using tools like **Streamlit** or **Power BI** for business decision-makers.