Introduction:

E-ComPredict: Customer Purchase Forecasting

Project Goal

The goal of this project is to predict whether a customer will purchase or not based on their session details (age, time spent, device type, discount applied, etc.).

- Type of Problem → Supervised Machine Learning (Binary Classification).
- Target Variable → purchase (0 = No, 1 = Yes).

Why important? → Businesses can target ads, optimize discounts, and improve sales by knowing which customers are most likely to buy.

Dataset Introduction: Realistic E-Commerce Dataset

- This dataset simulates customer interactions on an e-commerce website.
- Each row represents a customer session (one visit to the website).
- The aim is to analyze user behavior and predict purchase decisions.

Shape:

- Rows (records) → 1,200 sessions
- Columns (features) → 25 attributes

Key Columns

Customer Info

- user_id → unique customer ID
- gender → Male / Female
- age → customer age
- location → country of customer
- membership_status → Guest / Registered
- returning_customer → 1 if returning, 0 if first time

Traffic & Marketing

- traffic_source → Organic / Social / Referral / Paid
- ad_campaign → Campaign_A / Campaign_B / Campaign_C
- coupon_used → 1 if coupon used, 0 otherwise
- discount_applied → 1 if discount applied, 0 otherwise

Device & Browsing Info

- device_type → Desktop / Mobile / Tablet
- browser → Chrome, Firefox, Safari, Edge
- time_of_day → Morning / Afternoon / Evening / Night
- time_spent_minutes → total time spent on site
- pages_viewed → number of pages browsed
- scroll_depth → how far they scrolled (percentage)
- clicks → number of clicks

Key Columns

Shopping Behavior

- product_category → Clothing, Electronics, Books, Sports, etc.
- wishlist_items → number of items in wishlist
- cart_items → number of items added to cart
- avg_session_value → average monetary value of session ($)
- payment_method → UPI / Debit Card / Credit Card / NetBanking / COD

Target Variable

- purchase → 1 = Purchased
- 0 = Did Not Purchase

Exploratory Data Analysis (EDA)

Before building the model, we explored the dataset to understand patterns.

- Demographics → Age distribution, gender split, customer locations.
- Behavioral patterns → Time spent, pages viewed, clicks, wishlist/cart items.
- Traffic sources → Organic, Social, Referral, Paid.
- Effect of discounts/coupons → Did discounts increase purchase likelihood?
- Purchase rate by device, membership, and campaign.

Data Preprocessing

- Removed irrelevant columns (user_id, session_id, date).
- Encoded categorical features (e.g., Gender → 0/1).
- Standardized numerical values (scaling).
- Split into training (80%) and testing (20%).
- Saved feature names & scaler for deployment consistency.

Model Building

- Algorithm Used → Random Forest Classifier.
- Why? → Handles mixed data well, robust, and interpretable.
- Trained model on preprocessed features.
- Saved trained model (.pkl file) for reuse in deployment.

Model Evaluation

- Metrics used: Accuracy, Precision, Recall, F1-score.
- Confusion Matrix → shows true/false positives/negatives.
- Business interpretation:
- High recall → Good at catching all buyers.
- High precision → Ensures predicted buyers are real buyers.

Random Forest Classifier

What is Random Forest?

- Random Forest is a machine learning algorithm that builds many decision trees and combines them.
- Each tree gives a prediction, and the forest takes a majority vote → this makes it more accurate and stable than a single tree.

Why Not a Single Decision Tree?

- A single decision tree is easy to interpret, but:
  - It can overfit (memorize training data).
  - It is unstable → small data changes can change the whole tree.

Why Random Forest is Better

- More accurate → because it averages multiple trees.
- Less overfitting → because it uses random samples and features.
- Handles mixed data → works well with both numbers and categories.
- Robust → performs well on most datasets without heavy tuning.

Deployment (Streamlit)

- Built an interactive web app using Streamlit.
- Sidebar inputs → Users fill session details (age, device, time spent, etc.).
- Model predicts:
- Prediction → Will the customer purchase? (Yes/No).
- Probability → Confidence (%) of purchase.
- Deployed to Streamlit Cloud for public access.

Final Outcome

By the end of this project, students see the full journey:

- From raw e-commerce data → insights (EDA) → ML model → deployed web app.
- They learn both technical ML concepts and real-world application in business.

Implementation

Exploratory Data Analysis

- Univariate Analysis → Examines the distribution of a single variable to understand its central tendency, spread, and patterns.
- Bivariate Analysis → Explores relationships between two variables to find associations, correlations, or dependencies.
- Multivariate Analysis → Investigates interactions among three or more variables simultaneously to uncover complex patterns.
- Descriptive Statistics → Summarizes data with measures like mean, median, variance, and standard deviation.
- Data Visualization → Uses plots (histogram, boxplot, scatter, heatmap) to make patterns and insights visually clear.

Model Developing