PREDICTIVE ANALYSIS
LAB PROJECT
REPORT

GROUP 2

NAME	SAPID
1. NATANSH SHARMA	500104474
2. PIYUSH SINGLA	500107861

Personalized Product Recommendation System with Sentiment Analysis

---

1. Introduction

We aim to create a personalized product recommendation system that incorporates **sentiment analysis** to enhance the accuracy of recommendations. Traditional recommendation systems rely primarily on product interactions and purchase history, but integrating customer reviews through sentiment analysis enables more refined suggestions, helping users make better choices.

2. Objective

The primary objective of this project is to develop a **personalized product recommendation system** that uses both **collaborative filtering** and **content-based filtering**, enhanced with **sentiment analysis** on customer reviews. By evaluating the sentiment behind user reviews, we will be able to assign ratings to products and recommend items based on both user behavior and review sentiments.

---

3. Data Collection

3.1 Defining Data Sources

We collected data from multiple online services, including e-commerce platforms and review websites. The data includes product details, customer reviews, and user interactions with products such as views, clicks, and purchases.

3.2 Data Acquisition

Data was acquired by:

- **APIs:** Extracting product and user data using API calls to the respective platforms.

- **Web Scraping:** For some sources, we utilized web scraping to collect customer reviews and product metadata.
- **Databases:** Collected from product databases provided by e-commerce services.

The collected datasets are stored in CSV format to ensure portability and ease of use for preprocessing.

3.3 Data Privacy

All data was handled in compliance with privacy policies and regulations. Sensitive user data has been anonymized, and the datasets are stored securely.

---

4. Data Preprocessing

4.1 Data Cleaning

- **Missing Values:** Handled by removing or imputing missing data to ensure a complete dataset.
- **Duplicates:** Duplicate entries were removed to avoid skewing results.
- **Data Consistency:** Checked for and corrected inconsistencies, such as inconsistent product IDs or mismatched reviews.

4.2 Data Transformation

- **Normalization and Scaling:** Applied to numerical data to improve the performance of machine learning models.
- **Categorical Encoding:** Converted categorical features, such as product categories, into numerical format using techniques like one-hot encoding.

4.3 Data Segmentation

Data was segmented into key components:

- **Users:** User-specific features like demographics and purchase history.
- **Products:** Product-related features such as category, price, and description.
- **Interactions:** User-product interactions, including purchases, clicks, and reviews.

---

5. Exploratory Data Analysis (EDA)

5.1 Statistical Analysis

We performed statistical analysis on various aspects of the data:

- **User Demographics:** Age, gender, and location distribution.
- **Product Categories:** Distribution of products across different categories.
- **Purchase Frequency:** The frequency of product purchases by users.

5.2 Visualization

We created charts and graphs to visualize:

- **User Behavior:** Heatmaps and bar charts to show purchase patterns.
- **Product Popularity:** Pie charts and bar plots to illustrate the most popular products.
- **Correlations:** Scatter plots and correlation matrices to highlight relationships between product features and user interactions.

---

6. Recommendation Algorithm

We employed a combination of algorithms for our recommendation system:

6.1 Collaborative Filtering

Used techniques like:

- **User-based collaborative filtering**: Recommends products based on the preferences of similar users.
- **Item-based collaborative filtering**: Recommends products that are similar to items the user has interacted with.

6.2 Content-Based Filtering

We recommend products that are similar to those the user has liked or purchased, based on:

- **Product Features** such as category, price range, and description.
- **Customer Reviews:** Sentiment analysis of the reviews, allowing for recommendations based on the tone of feedback.

6.3 Hybrid Model

We combined collaborative and content-based filtering to leverage the strengths of both methods. This improves the overall recommendation accuracy by incorporating both user preferences and product similarities.

6.4 Sentiment Analysis on Reviews

We analyzed the customer reviews using natural language processing (NLP) techniques, particularly sentiment analysis, to evaluate the emotions expressed in user reviews. Reviews were classified as positive, neutral, or negative, which further refined the product recommendation based on user feedback.

---

7. Model Development

We developed and trained the model using the following steps:

- **Cross-validation:** Ensured the model's robustness by splitting the data into training and validation sets.
- **Evaluation Metrics:** Used Root Mean Square Error (RMSE), precision, recall, and F1 score to evaluate model performance.

---

8. Model Evaluation

The model's performance was evaluated using:

- **Precision and Recall:** To measure how many recommended items were relevant.
- **F1 Score:** A balanced metric between precision and recall.
- **AUC-ROC Curve:** For classification tasks, the AUC-ROC curve was used to assess the model's discriminatory ability.
- **RMSE and MAE:** For regression tasks like rating prediction, RMSE and Mean Absolute Error (MAE) were calculated.

---

9. Deployment

9.1 Model Integration

We integrated the recommendation system into an application using:

- **Flask API:** A web-based API was developed to serve recommendations in real-time to users.

9.2 Scalability

We deployed the system on **AWS** and ensured scalability to handle large traffic volumes by using services such as **Amazon EC2** and **Load Balancers**.

---

10. Monitoring and Maintenance

We plan to continuously monitor the system post-deployment by:

- **Gathering Feedback:** Regularly collecting feedback from users on the relevance of the recommendations.
- **Performance Monitoring:** Tracking performance metrics like response time and user engagement to ensure the system functions smoothly.

---

11. Iteration and Improvement

11.1 Continuous Improvement

Based on feedback and newly acquired data, the model will be improved iteratively. Regular updates will include:

- Retraining the model with new data.
- Adding new features or adjusting existing ones to improve recommendation accuracy.

11.2 Experimentation

We plan to experiment with different algorithms, data sources, and feature engineering techniques to refine the model further. This iterative approach will ensure the system remains accurate and relevant as user preferences and product trends evolve.

---

12. Tools and Technologies Used

- **Programming Languages:** Python, R
- **Libraries:** Scikit-learn, TensorFlow, Keras, Pandas, NumPy, Matplotlib, Seaborn
- **Databases:** MySQL
- **APIs:** Flask (for API development)
- **Cloud Services:** AWS, Google Cloud, Microsoft Azure (for deployment and scalability)

Conclusion

In conclusion, our project successfully developed a personalized product recommendation system that integrates sentiment analysis. By leveraging both collaborative and content-based filtering along with sentiment from user reviews, the system can provide users with more accurate and relevant product recommendations. We look forward to iterating on the model based on user feedback and expanding its capabilities further.