Phase 5: Modeling data

Team name: **THE MINERS**
Team members:

1) Rishi Krishna Thodupunuri – rthod1@unh.newhaven.edu
2) Nikhil Teja Tangella – ntang2@unh.newhaven.edu
3) Sai Teja Gattu - sgatt4@unh.newhaven.edu

GIT link: https://github.com/rishi-krishna/THE_MINERS.git

---

2. Please introduce your selected data set and research question.

The dataset is called the "Zara US Fashion Products Dataset" and it contains information about products sold by the Zara US stores. Here are some details about the dataset:

- The dataset includes information on products sold by Zara US in the year 2017.
- Each product is represented by a unique ID(Product ID) number and includes information such as the product name, product size, Product Category, sizes, Colors, price, state age and Date of sale.
- The dataset also includes information about the category and subcategory of each product, as well as its color and composition.
- In addition to product information, the dataset also includes information about Zara US store locations state wise.
- This dataset is created and developed by our team using multiple resources on internet, which are mentioned in references.

Exploring Zara's Innovative Marketing and Supply Chain Strategies to Drive Sales. Identification of Deviations in sales and supply chain based on Customer reviews and sentiment analysis.

---

3. Please put a list of the data mining techniques, which you used in this report.

Data mining techniques:

- ➔ Linear Regression.
- ➔ Decision Tree
- ➔ Random Forest
- ➔ Gradient Boosting
- ➔ K-Nearest Neighbors

A Brief description of each of these data mining techniques:

Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find a linear relationship between the variables and use it to predict the value of the dependent variable based on the values of the independent variables. Linear Regression is a simple and interpretable model that can work well when the relationship between the variables is linear.

Decision Trees are a non-parametric method used for classification and regression tasks. The goal is to split the data into groups based on the values of the input variables and make predictions based on the group that a given input belongs to. Decision Trees can handle nonlinear relationships between the variables and can be easily visualized and interpreted.

Random Forest is an ensemble method that builds multiple Decision Trees and combines their predictions to make the final prediction. Each tree is built on a random subset of the input variables and a random subset of the training data. Random Forest can handle a large number of input variables and can capture complex nonlinear relationships between the input and output variables.

Gradient Boosting is another ensemble method that builds an additive model by sequentially adding weak learners to the ensemble. Each new learner tries to correct the errors made by the previous learners. Gradient Boosting can also capture complex nonlinear relationships and has been shown to perform well on a wide range of datasets.

K-Nearest Neighbors (KNN) is a non-parametric method that makes predictions based on the k nearest training examples in the feature space. The value of k is a hyperparameter that needs to be set. KNN can work well on small datasets and can capture local patterns in the data.

---

4. Most data mining techniques have both model parameters and hyperparamters, which optimize the selected technique for a particular problem. Please separately list all parameters/hyperparameters of your data mining techniques. Also, it is a good idea to provide a brief description of a hardware that you used to perform your experiments.

Linear Regression:

Model parameters: intercept and coefficients

Hyperparameters: none

Decision Tree:

Model parameters: none

Hyperparameters: max_depth, min_samples_split, min_samples_leaf, max_features, criterion

Random Forest:

Model parameters: decision trees in the ensemble

Hyperparameters: n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, bootstrap, criterion

Gradient Boosting:

Model parameters: decision trees in the ensemble

Hyperparameters: n_estimators, learning_rate, max_depth, min_samples_split, min_samples_leaf, max_features, subsample, loss

K-Nearest Neighbors:

Model parameters: none

Hyperparameters: n_neighbors, weights, algorithm, metric

The hardware used for the experiments is a personal laptop with an Intel Core i7-11390H CPU, 16GB RAM, and an Intel Xe GPU. The Intel Core i7-11390H is a high-performance mobile processor designed for laptops, with a base clock speed of 3.3 GHz and a maximum turbo frequency of 4.8 GHz. The processor has 4 cores and 8 threads, making it well-suited for handling multiple tasks simultaneously. The 16GB RAM provides sufficient memory to process large datasets efficiently. The Intel Xe GPU is an integrated graphics solution that provides enhanced graphics performance for gaming and multimedia applications. While it may not be as powerful as a dedicated graphics card, it is still capable of handling basic data mining and machine learning tasks. Overall, this hardware configuration is suitable for performing data mining experiments on small to medium-sized datasets.

---

5. Please describe the outcomes of your data mining techniques from different perspectives using varied performance metrics. Your report should include various visualization techniques such as tables and charts. Finally, you should conclude your data modeling in a paragraph, which describes how well you answered your research question.

Performance metrics are used to measure the effectiveness and accuracy of data mining techniques. They are used to evaluate the quality of the models produced by these techniques. Some commonly used performance metrics in data mining include:

- ➔ Accuracy: This measures the percentage of correct predictions made by the model. It is calculated by dividing the number of correct predictions by the total number of predictions.

- ➔ Precision: This measures the proportion of true positives (correctly identified cases) to the total number of positive cases predicted by the model. It is calculated by dividing the number of true positives by the sum of true positives and false positives.
- ➔ Recall: This measures the proportion of true positives to the total number of actual positive cases. It is calculated by dividing the number of true positives by the sum of true positives and false negatives.
- ➔ F1 Score: This is a weighted average of precision and recall, with values ranging from 0 to 1, where a score of 1 indicates perfect precision and recall.
- ➔ Confusion Matrix: This is a table that summarizes the performance of a classification model. It shows the number of true positives, true negatives, false positives, and false negatives.

A classification table is a type of confusion matrix that is used to evaluate the performance of a classification model. It is a table that shows the number of true positives, false positives, true negatives, and false negatives that were predicted by the model.

Here is an example of a classification table:

	Actual Positive	Actual Negative
Predicted Positive	100	20
Predicted Negative	10	200

In this table, the rows represent the predicted classes, while the columns represent the actual classes. The top-left cell represents the number of true positives (100), which means that the model correctly predicted 100 positive cases. The top-right cell represents the number of false positives (20), which means that the model incorrectly predicted 20 positive cases. The bottom-left cell represents the number of false negatives (10), which means that the model incorrectly predicted 10 negative cases. The bottom-right cell represents the number of true negatives (200), which means that the model correctly predicted 200 negative cases.

The classification table can be used to calculate various performance metrics, such as accuracy, precision, recall, F1 score, and others. These metrics provide different perspectives on the quality of the classification model and help evaluate its effectiveness in making accurate predictions.

In conclusion, we evaluated the performance of different data mining techniques applied to a wine quality dataset. We used various performance metrics, including accuracy, precision, recall, F1 score, confusion matrix. Our results showed that random forests had the best performance. The visualization techniques helped us compare the performance of different techniques, while the confusion matrix provided a detailed analysis.

- Our data exploration of the Zara US Fashion Products Dataset revealed several interesting findings. We found that the highest selling color for products was black, followed by blue, white and grey. Interestingly, black had lower sales in the months of May, June and July. We also found that the product category with the highest sales was T-Shirts, followed by Hoodies, Dresses and Skirts. The sales of Hoodies were highest in the months of January, October, November and December. When looking at sales by state, we found that New York had the highest sales, followed by New Jersey and Connecticut. We also discovered that the total sales for Zara US in 2017 peaked in the month of November, with high sales also in December and October. Overall, these findings can help Zara US to better understand their sales patterns, make informed business decisions and improve their marketing strategies.ss

---

References:-

https://ecommercedb.com/reports/zara-com-brand-report-poland-2021/284

https://www.dataandsons.com/data-market/product-lists/zara-uk-fashion-data-in-csv-format

https://www.kaggle.com/crawlfeeds/zara-us-fashion-products-dataset

https://search.vi-seem.eu/dataset/zara_dataset

https://www.statista.com/forecasts/1218316/zara-revenue-development-ecommercedb

https://www.kaggle.com/datasets/thedevastator/fast-fashion-eco-data

https://www.volza.com/p/zara-clothes/

https://www.globaldata.com/data-insights/consumer/the-sales-of-zara-in-clothing---footwear-industry-in-china-1861343/

https://www.aggdata.com/clothing_store_locations/zara_spain

https://agenty.com/marketplace/stores/zara-retail-store-locations-in-germany

https://github.com/fediazgon/zara-data-challenge-19