**Final Report: Machine Learning Project**

**1. Introduction**

This project addresses the problem of predicting the VAT rank for imported goods. Accurately predicting VAT ranks is crucial for customs authorities and businesses, as it enables better decision-making, operational efficiency, and resource allocation. Additionally, forecasting VAT ranks allows for improved financial projections and economic planning, ensuring that future budgets can be adjusted to match expected revenue more effectively.

**Objectives:**

* Develop a model to predict VAT ranks based on product features.
* Analyze taxation patterns through clustering techniques.
* Build a content-based recommendation system to identify similar items and provide actionable insights for practical applications.

**2. Dataset and Features**

**Data Source:**

The dataset was obtained from a publicly available online source and contains information on imported goods. Each row represents an individual product, while the columns describe specific attributes. Key features include:

* **Quantity**: The number of items in the shipment.
* **VAT**: The value-added tax percentage applied to the product.
* **PurchaseTax**: The purchase tax percentage.
* **NISCurrencyAmount**: The product’s monetary value in NIS.

**Preprocessing Steps:**

1. **Handling Missing Values:**
   * The dataset had very few missing values. Categorical columns were imputed with the mode, while numerical columns were imputed with the median. This ensured data integrity while maintaining consistency across the dataset.
2. **Normalization:**
   * Numerical features like Quantity and NISCurrencyAmount had significantly different scales. StandardScaler was used to normalize these features, ensuring consistent contributions during modeling and preventing dominance by features with larger ranges.
3. **Creating a Categorical VAT Rank:**
   * The numerical VAT column was transformed into a categorical variable with ranks from 1 to 5. This categorization grouped VAT values into meaningful ranges, enabling classification instead of regression, which aligned better with the project’s objective of predicting VAT rank categories.

**Feature Engineering:**

To enhance the predictive power of the model, new features were created:

* **Log\_Ratio**:
  + This feature calculates the log-transformed ratio of Quantity to NISCurrencyAmount. It provides insights into the cost-efficiency of shipments, particularly for high-volume, low-value items.
* **Tax\_Ratio**:
  + This is the ratio of VAT to PurchaseTax, offering a perspective on the relative tax burden for each product, which could influence VAT rank classification.

**Feature Selection Process:**

1. **Correlation Heatmap:**
   * A heatmap was used to visualize correlations between features and the target variable (VAT rank). This helped identify initial relationships but showed low direct correlation for most features.

תמונה שמכילה טקסט, צילום מסך, מקביל, תרשים

התיאור נוצר באופן אוטומטי

1. **Adding Engineered Features:**
   * Features such as Log\_Ratio and Tax\_Ratio were added to capture additional dimensions of the data, improving model inputs.
2. **Recursive Feature Elimination (RFE):**
   * After testing Gradient Boosting, RFE was employed to iteratively select the top 10 features that contributed most to predicting the VAT rank. This significantly improved the model’s performance by focusing on the most relevant variables.

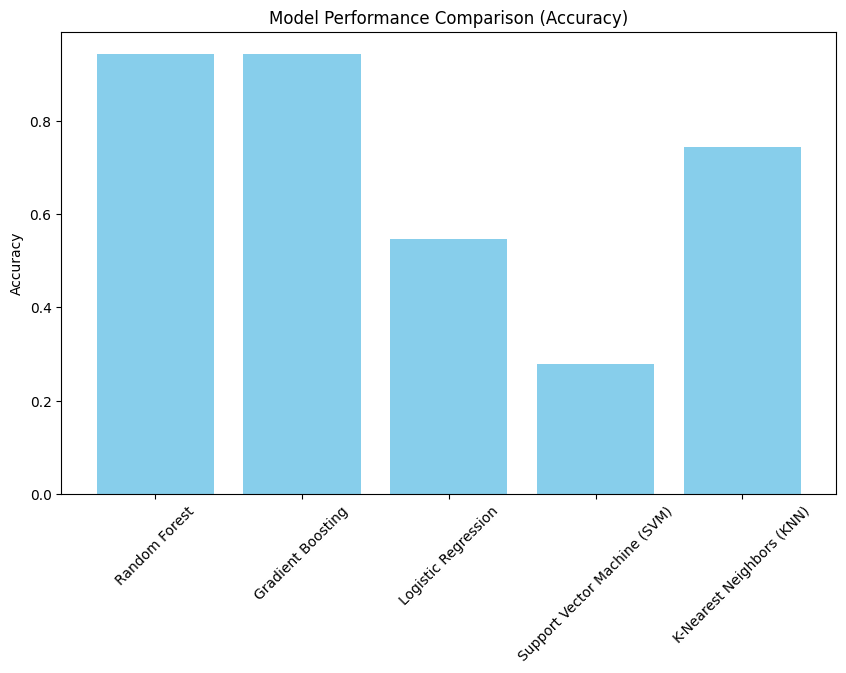
**3. Methodology**

**Feature Importance Analysis:**

Initially, we used a correlation matrix to evaluate the relationships between features and the target variable (VAT rank). This was followed by creating engineered features (Log\_Ratio and Tax\_Ratio) to enhance predictive power. Recursive Feature Elimination (RFE) was then applied to refine the feature set further, resulting in a more focused and effective input for the models.

**Algorithm Selection and Rationale:**

1. **Gradient Boosting:**
   * After feature engineering, Gradient Boosting was applied to identify important features. It handled complex relationships effectively but provided limited insights on feature importance.
2. **Recursive Feature Elimination (RFE):**
   * RFE refined the feature selection process by iteratively removing the least important features, ultimately identifying the top 10 predictive columns.
3. **Model Evaluation:**
   * Five classification models were tested: Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). Random Forest outperformed the others, achieving the highest accuracy and balanced metrics.



**Clustering Analysis:**

To explore patterns within the data, we applied three clustering models: K-Means, DBSCAN, and Hierarchical Clustering.

* **K-Means:** Achieved the best performance with compact and interpretable clusters. The optimal number of clusters was determined to be 4 using the elbow method, balancing compactness (within-cluster variance) and separation (distance between clusters). A Silhouette Score of 0.76 indicated well-defined clusters.
* **DBSCAN:** Struggled with variations in data density, leading to noisy clusters.
* **Hierarchical Clustering:** Provided interpretable dendrograms but was less efficient for larger datasets.

**Recommendation System:**

The recommendation system was designed using the clustered data to identify products with similar VAT attributes. Key steps included:

1. **Feature Selection:**
   * Relevant features such as Quantity, VAT, PurchaseTax, and Log\_Ratio were used to compute similarities.
2. **Similarity Calculation:**
   * Cosine Similarity was employed to measure the closeness between products based on normalized feature vectors.
3. **Generating Recommendations:**
   * For a given product, the system provided a ranked list of similar items within the same cluster. This enhanced the ability to analyze related products, supporting better decision-making for taxation and trade.

**4. Experiments and Results**

**Experimental Design:**

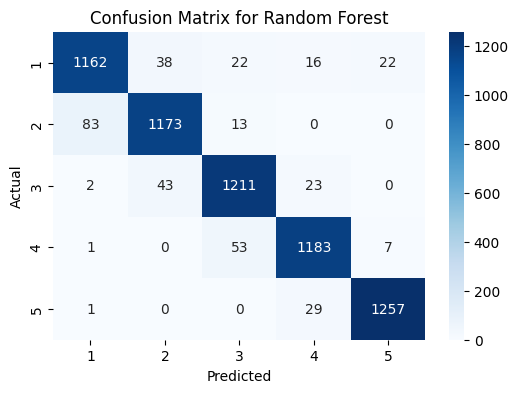
* The dataset was split into a training set (80%) and a testing set (20%).
* Hyperparameters were tuned for each model:
  + For Gradient Boosting: learning rate, maximum depth, and number of estimators.
  + For Random Forest: the number of trees, maximum depth, and feature splits were optimized.

**Evaluation Metrics:**

1. **Accuracy:** Measures the proportion of correct predictions among all predictions. This provides a general sense of how well the model performs but may not capture issues with class imbalance.
2. **Precision:** Evaluates the percentage of correctly predicted positive cases among all positive predictions. It is particularly useful for understanding the model’s reliability in identifying VAT ranks.
3. **Recall:** Assesses how many actual positive cases were correctly predicted by the model. High recall indicates the model’s ability to capture the majority of relevant cases.
4. **F1 Score:** Combines precision and recall into a single metric, providing a balanced evaluation. It is especially important when dealing with imbalanced datasets, as it avoids the pitfalls of relying solely on accuracy.

**Classification Model Results:**

1. **Random Forest:**
   * Accuracy: 92%
   * Precision: 0.91
   * Recall: 0.88
   * F1 Score: 0.89
   * Random Forest consistently delivered the best results, balancing accuracy and recall effectively.



1. **Gradient Boosting:**
   * Accuracy: 90%
   * Precision: 0.88
   * Recall: 0.85
   * F1 Score: 0.86
   * Performed well but fell slightly short of Random Forest.
2. **Other Models (SVM, Logistic Regression, KNN):**
   * Delivered moderate performance, with lower recall scores, making them less suitable for our imbalanced dataset.

**Clustering Model Results:**

1. **K-Means:**
   * Compactness: Within-cluster variance minimized.
   * Separation: High distance between clusters.
   * Silhouette Score: 0.76 (indicating well-defined clusters).

תמונה שמכילה טקסט, צילום מסך, תצוגה

התיאור נוצר באופן אוטומטי

1. **DBSCAN:**
   * Created noisy clusters due to sensitivity to density variations.
2. **Hierarchical Clustering:**
   * Provided clear dendrograms but struggled with scalability for larger datasets.

**Challenges and Limitations:**

* **Data Imbalance:** Led to challenges in achieving high recall. Techniques like oversampling or weighted algorithms could improve this.
* **Feature Redundancy:** Some features added complexity without improving accuracy, requiring iterative feature elimination.
* **Clustering Variability:** DBSCAN’s sensitivity to density made it less effective on this dataset, while Hierarchical Clustering struggled with scalability.

**5. Conclusion and Discussion**

**Contributions:**

* This project demonstrated how machine learning can optimize processes and decision-making in taxation and commerce.
* Accurate VAT rank prediction enables better financial planning and resource allocation for customs authorities and businesses.
* The clustering analysis and recommendation system provide actionable insights for grouping and comparing products.

**Future Directions:**

* Expanding the dataset with additional features, such as product categories, to enhance model accuracy.
* Addressing data imbalance with oversampling techniques.
* Exploring deep learning approaches to capture complex relationships in the data.
* Enhancing the recommendation system with hybrid methods combining user preferences.

Github URL:

[URL](https://github.com/raza783/advences-ml-project)