**Analysis of Support Vector Machines (SVM) and Decision Trees for Predicting Purchasing Behavior**

This essay and analysis assume prior reading of Homework #2 and the accompanying Jupyter notebook, where a comprehensive exploratory data analysis (EDA) of the dataset was conducted. In the previous work, detailed insights into the dataset, including the distribution and relationships between the features (age, gender, and income) and the target variable (purchased status), were explored. Much of this exploratory analysis is not reiterated here, as the focus of this essay is on comparing the performance of different machine learning algorithms, specifically Support Vector Machines (SVM) and Decision Trees, for predicting purchasing behavior. Readers are encouraged to refer to Homework #2 for the detailed EDA, which provides foundational context for understanding the dataset and its characteristics.

As a student exploring machine learning models in real-world applications, I recently analyzed a dataset to predict the likelihood of individuals making a purchase based on three features: age, gender, and income. This task was part of my ongoing research in data science, where I am learning to apply machine learning techniques to small datasets. In this context, I compared the performance of Support Vector Machines (SVM), Decision Trees, and Random Forests. While SVMs are often recommended for smaller datasets, I found that other models, particularly Decision Trees and Random Forests, outperformed the SVM in this case, which offers insights into the strengths and weaknesses of these algorithms. A comprehensive exploratory data analysis (EDA) of the dataset, including visualizations and statistical summaries, can be found in the essay and Jupyter notebook from Homework #2. This analysis provided a deeper understanding of the distribution of the features, the target variable, and potential relationships within the data, which informed the selection of algorithms for this task.

**Dataset Overview**

The dataset contained three features: age, gender, and income, and the target variable, purchased status, is a binary indicator. Specifically, 40.2% of individuals had made a purchase, while the remaining 60% had not, reflecting a class imbalance. This binary classification task aimed to predict purchasing behavior based on demographic information. The features in the dataset were categorical (gender) and continuous (age, income), posing unique challenges for model performance.

**Results from SVM**

For this analysis, I initially chose SVM as my primary model because of its strong performance in small datasets with well-defined margins between classes. As discussed in the literature, SVMs are effective at finding hyperplanes that separate different classes, especially when the data is not overly noisy or complex (Raschka, 2023). To optimize the SVM's performance, I conducted a randomized grid search to find the best hyperparameters. The grid search was performed on a subset of the data, and some kernel functions were omitted due to computational constraints. This was a significant limitation, as SVMs are computationally intensive and require considerable resources, especially when tuning hyperparameters over large datasets.

The grid search results showed that the best parameters were:

* **Kernel**: 'rbf'
* **Gamma**: 0.01
* **C**: 10

Despite this optimization, the SVM model's performance was still not as high as that of Decision Trees and Random Forests. The SVM model struggled with the class imbalance, and the accuracy and F1 score were modest. Additionally, it showed a tendency to misclassify the minority class, which in this case was the "purchase" category. This is a common issue for SVMs when dealing with imbalanced datasets, as the algorithm tends to favor the majority class, leading to less reliable predictions for the minority class (Alkheder et al., 2022).

A potential reason for the underperformance of the SVM model could be the computational constraints during hyperparameter tuning. By omitting some kernel functions and running the grid search on only a subset of the data, the model might not have explored the optimal set of hyperparameters fully. This highlights a limitation of SVMs: they require significant computational power and careful tuning, especially when working with large datasets or when computational resources are limited.

**Decision Trees and Random Forests**

In contrast, Decision Trees and Random Forests handled the dataset much better. The Decision Tree model, with its simple structure, was able to identify important splits based on the features, especially distinguishing between high-income and low-income groups for predicting purchasing behavior. The performance of Decision Trees was better than the SVM, as the model handled the imbalanced data more effectively, giving appropriate weight to the minority class.

However, the Random Forest model, which is an ensemble of Decision Trees, provided the best results. By aggregating the predictions from multiple trees, Random Forests reduced the overfitting that could occur with individual Decision Trees, leading to better generalization. The model showed higher accuracy, F1 score, and AUC compared to SVM and Decision Trees, especially in predicting purchases (the minority class). Random Forests also performed well in dealing with the non-linearity of the dataset, where the relationships between features and the target were not purely linear (Gatera et al., 2023).

**Comparison with Literature**

The performance observed in my analysis aligns with findings from previous studies comparing SVMs with other algorithms like Random Forests. Alkheder et al. (2022) found that Random Forest outperformed SVM in predicting traffic accidents, primarily because it can better handle complex and imbalanced datasets. Similarly, Gatera et al. (2023) showed that Random Forest models had a lower error rate compared to SVM for predicting road accidents in Rwanda, which suggests that for tasks like classification in small datasets, Random Forests often outperform SVMs. Moreover, Raschka (2023) discusses that SVMs are well-suited for small datasets but often require careful tuning of hyperparameters, especially when handling imbalanced classes. In my analysis, I found that SVM's performance could be improved with parameter tuning, but even after tuning, it did not surpass the performance of Decision Trees and Random Forests. This is because, unlike SVM, Decision Trees and Random Forests inherently handle class imbalances and non-linear data relationships better without extensive hyperparameter optimization.

**Conclusion**

Based on my analysis of the dataset with age, gender, and income as features, the Random Forest model emerged as the best-performing algorithm for predicting purchasing behavior. It outperformed SVM in terms of accuracy, F1 score, and AUC, especially when dealing with the imbalanced target variable. While SVM can be effective for small datasets, its performance was hindered by the class imbalance, which is a critical consideration in real-world datasets. Additionally, the computational cost of tuning the SVM model with a randomized grid search on a subset of the data may have limited its full potential.

For this classification task, Random Forests are recommended due to their ability to handle imbalanced data and complex relationships without requiring extensive tuning. This result supports the idea that, while SVM may be suitable for certain small datasets, other models like Decision Trees and Random Forests are more robust in typical classification tasks with class imbalances, like the one used in this analysis.

**References:**

* Alkheder, S., AlRukaibi, F., & Aiash, A. (2022). Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN), and Bayesian Network for Prediction and Analysis of GCC Traffic Accidents. *Journal of Ambient Intelligence and Humanized Computing*.
* Gatera, A., Kuradusenge, M., Bajpai, G., Mikeka, C., & Shrivastava, S. (2023). Comparison of Random Forest and Support Vector Machine Regression Models for Forecasting Road Accidents. *Science Progress*.
* Raschka, S. (2023). Machine Learning FAQ: Random Forests vs. SVMs. Retrieved from <https://sebastianraschka.com/faq/docs/deeplearn-vs-svm-randomforest.html>