Essay: Algorithm selection, implementation, and results (Marketing prediction)

This assignment evaluated multiple supervised classification algorithms to predict client subscription outcomes in a marketing context using the bank marketing dataset. The dataset contains a mixture of numeric and categorical predictors and a binary target (y). Because the goal is to maximize correct campaign targeting decisions (finding customers likely to subscribe), our analysis focused on classification performance (accuracy, precision, recall, F1) as well as practical considerations such as interpretability and computational cost.

We began with careful preprocessing: removing 'duration' to avoid target leakage, flagging previously contacted clients via 'pdays_flag', converting categorical variables to factors, imputing numeric missing values with medians, and applying one-hot encoding followed by normalization. These preprocessing decisions are particularly important for SVMs, which require scaled numeric inputs, and for tree-based models, which handle categorical information differently. After an 80/20 stratified split, we trained and evaluated baseline and tuned models for Decision Tree, Random Forest, and Support Vector Machine (RBF and linear).

The baseline Decision Tree offered a straightforward, interpretable model with strong recall (i.e., it located most positive cases), achieving an accuracy close to 0.893. Constraining tree complexity (max depth = 3) produced similar accuracy while improving interpretability. Random Forest, an ensemble of trees, slightly improved raw accuracy (≈0.894) and kept a favorable balance between precision and recall, reflecting the ensemble's tendency to reduce variance and handle interactions robustly. For SVMs, we implemented both an RBF kernel (non-linear) and a linear kernel (fast baseline). The RBF SVM’s tuned performance matched the tree-based models (≈0.893 accuracy), while the linear SVM ran faster and produced comparable results after preprocessing and encoding—suggesting the transformed feature space may be approximately linearly separable for the classification task.

SVM tuning was initially computationally heavy because large kernel matrices and multiple hyperparameter combinations (cost and rbf_sigma) require extensive computation on ~36k training records. To make SVM tuning feasible, we tuned on a 40% subsample with a small 2x2 grid and 2-fold CV, then refitted the winning configuration on the full training set. This approach dramatically reduced runtime (from ~30+ minutes to about 3–5 minutes) while preserving predictive performance, a practical compromise aligned with literature guidance for large-data SVM tuning.

The literature supports our findings: SVMs often perform strongly on well-scaled datasets and in controlled feature spaces, but tree ensembles (Random Forest) are practical workhorses on tabular data because they require less tuning and provide variable importance measures useful for interpretation. The provided Hindawi and PMC articles emphasize ensemble robustness and SVM’s diagnostic strengths respectively, while additional comparative studies (Guia et al., Ozcan et al., Sheth et al.) show that SVMs and RFs often yield similar top-tier performance with trade-offs in tuning and interpretability.

Conclusions: For marketing prediction tasks with mixed feature types and large samples, Random Forest is recommended as the first-line algorithm due to a small advantage in accuracy, robustness to noisy predictors, and easier interpretability via feature importance. SVMs remain a viable alternative when computational resources and careful preprocessing are available and when margin-based classifiers are desired. Decision Trees are useful when clear rule-based explanations are required, but they typically underperform ensembles. Overall, model choice should weigh accuracy against interpretability and runtime constraints; in this assignment, differences in accuracy were marginal, so practical considerations should guide deployment.

Literature Review

Provided Articles:

- **Ahmad et al. (2021)** — *“Decision Tree Ensembles to Predict Coronavirus Disease”* (Hindawi, 2021). https://www.hindawi.com/journals/complexity/2021/5550344/
- **Guhathakurta et al. (2021)** — *“A Novel Approach to Predict COVID-19 Using Support Vector Machines”* (PMC, 2021). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8137961/

Additional Articles:

- **Guia et al. (2019)** — *“Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest.”* https://pdfs.semanticscholar.org/ded2/ab1b4add4061e49957ec279830a6bbaf0656.pdf
- **Ozcan et al. (2022)** — *“Comparison of Classification Success Rates of Different Machine Learning Algorithms.”* https://pmc.ncbi.nlm.nih.gov/articles/PMC9924317/
- **Sheth et al. (2022)** — *“Comparative Analysis of Machine Learning Algorithms.”* https://www.sciencedirect.com

The first two provided articles offered foundational insights into the comparative performance of Support Vector Machines (SVMs) and Decision Tree–based models. **Ahmad et al. (2021)** explored Decision Tree Ensembles for disease prediction and demonstrated that combining multiple trees through ensemble techniques such as Random Forests enhances model robustness and reduces overfitting. The study highlighted that Decision Tree–based models perform well in handling categorical and numerical data, much like the diverse features in our bank marketing dataset. On the other hand, **Guhathakurta et al. (2021)** focused on SVMs for medical data classification, emphasizing their high accuracy in complex, nonlinear data environments. However, they also noted that tuning SVMs can be computationally intensive, which mirrors our own experience when implementing the SVM model on a large dataset of over 45,000 records.

To supplement these findings, three additional scholarly articles were reviewed. **Guia et al. (2019)** compared SVM, Decision Tree, Naïve Bayes, and Random Forest algorithms, concluding that while SVMs excel with smaller and cleaner datasets, Random Forests tend to generalize better on large, diverse data. Similarly, **Ozcan et al. (2022)** found that both Random Forests and SVMs achieved comparable accuracy levels in biomedical prediction, but Random Forests required less parameter tuning and computational time. **Sheth et al. (2022)** expanded the comparison across multiple domains and concluded that model performance depends heavily on data characteristics and the trade-off between interpretability and accuracy. Together, these three studies provided a broader context for understanding how model selection should be data-driven and problem-specific.

In comparing all five studies, consistent themes emerged. SVMs are theoretically strong for achieving high precision and well-defined decision boundaries, but their computational complexity makes them less scalable for large datasets. Conversely, Random Forests and other ensemble tree-based models maintain competitive accuracy while being easier to train and interpret. These insights directly align with the outcomes of this assignment—where the Random Forest model achieved slightly higher accuracy and stability than SVMs, particularly in a large, real-world marketing prediction context. Collectively, the reviewed literature supports the conclusion that while SVMs are powerful, Random Forests are often the more practical and balanced choice for business and marketing analytics tasks.