Assignment 3
Report

INF8245AE
Machine Learning

Name	Student ID
Ali Abbas	2078188

Polytechnique Montréal
10 November 2025

Data Exploration and Preprocessing

a) Results of data exploration

- Average Age: 38.64 years
- Percentage of Women: 33.15%
- Percentage earning more than $50K: 23.93%
- Percentage of Missing Values: 1.10%

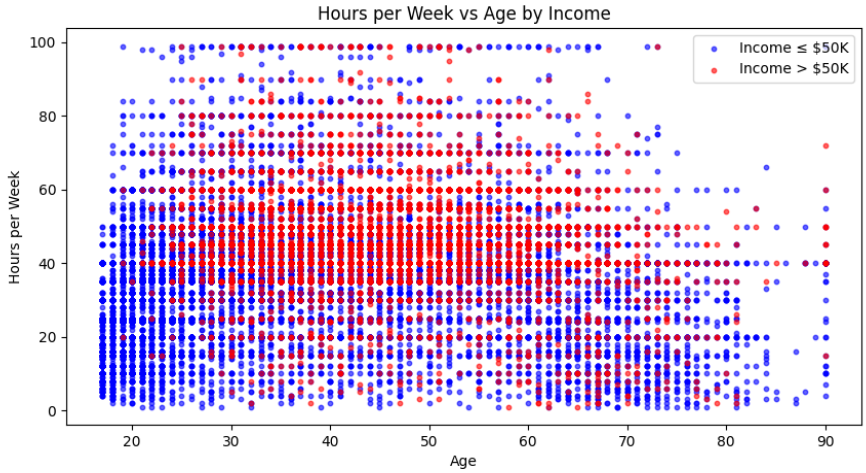


Figure 1 scatter plot with features "Hours per week" and "Age"

b) There are 479 features after preprocessing.

Model Training and Evaluation

4 b) Training and testing results

	Train Accuracy	Test Accuracy	Train F1-score	Test F1-score
Decision Tree	0.9739	0.8257	0.9736	0.8240
Random Forest	0.9739	0.8495	0.9738	0.8450
SVM	0.8789	0.8682	0.8725	0.8612

Table 1 Accuracy and F1-scores

The Decision Tree overfits, 97% train accuracy vs 83% test accuracy. The same trend is observed for Random Forest, but it performs better on the testing set compared to the decision tree (83% vs 85% test accuracy). The SVM avoids overfitting (88% training vs 87% test accuracy), it provides the strongest generalization, as well as the highest F1 test score.

4 d) Training and testing classification report

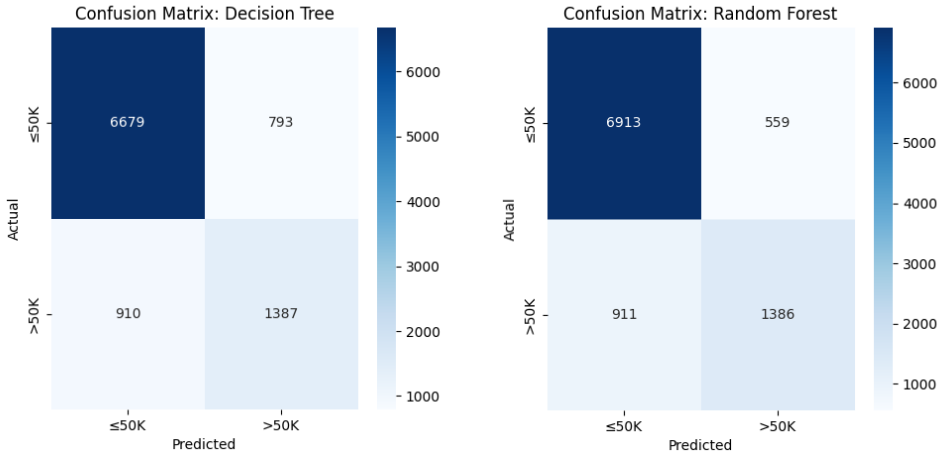
	Precision	Recall	F1-score	Support
Decision Tree				
0 (≤ 50K)	0.97	0.99	0.98	29683
1 (> 50K)	0.98	0.91	0.94	9390
Accuracy			0.97	39073
Macro average	0.97	0.95	0.96	39073
Weighted average	0.97	0.97	0.97	39073
Random Forest				
0 (≤ 50K)	0.98	0.99	0.98	29683
1 (> 50K)	0.96	0.93	0.95	9390
Accuracy			0.97	39073
Macro average	0.97	0.96	0.96	39073
Weighted average	0.97	0.97	0.97	39073
SVM				
0 (≤ 50K)	0.89	0.96	0.92	29683
1 (> 50K)	0.83	0.62	0.71	9390
Accuracy			0.88	39073
Macro average	0.86	0.79	0.82	39073
Weighted average	0.88	0.88	0.87	39073

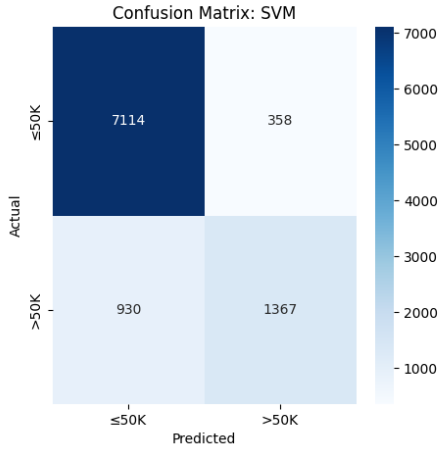
Table 2 Training set classification report

	Precision	Recall	F1-score	Support
Decision Tree				
0 (≤ 50K)	0.88	0.89	0.89	7472
1 (> 50K)	0.64	0.60	0.62	2297
Accuracy			0.83	9769
Macro average	0.76	0.75	0.75	9769
Weighted average	0.82	0.83	0.82	9769
Random Forest				
0 (≤ 50K)	0.88	0.93	0.9	7472
1 (> 50K)	0.71	0.60	0.65	2297
Accuracy			0.85	9769
Macro average	0.8	0.76	0.78	9769
Weighted average	0.84	0.85	0.85	9769

SVM				
0 (≤ 50K)	0.88	0.95	0.92	7472
1 (> 50K)	0.79	0.60	0.68	2297
Accuracy			0.87	9769
Macro average	0.84	0.77	0.8	9769
Weighted average	0.86	0.87	0.86	9769

Table 3 Testing set classification report





Similar to the analysis from question c), we can observe that the decision tree and the random forest tend to overfit the training set, especially in the case of class 1 (class 1 F1-score of 0.94 and 0.95 respectively) and fail to generalize as can be seen from the results of the testing set (F1 score of 0.62 and 0.65 respectively). Again, we can see that SVM results provide the best generalization, as well as the highest precision scores on average between the two classes. We can conclude that the SVM offers the best performance it is important to note it has the highest occurrences of False Negatives (930) but it is not too far off the decision tree and Random forest and is acceptable in comparison.

Model Tuning

Question 1: Hyperparameters

Random Forest

- Number of trees: More trees reduce variance and help generalization but are longer to compute.
- Maximum depth: Low depth trees are less complex (leads to low variance and high bias) and vice versa.
- Number of features: The number of features to consider when finding the best split, a bigger number of features leads to more accurate trees (and likely overfitting), a smaller number of features reduces computational cost as well as variance.
- Minimum samples in a leaf: Controls how deep a tree can be, in other words controls overfitting by specifying if a split is allowed or not.
- Bootstrap: Bootstrapping creates multiple training points (simulated datasets) by resampling.
  Note: Bootstrapping itself is a technique and not a hyperparameter, but in the context of *Scikit-learn* it is a tunable parameter.

SVM

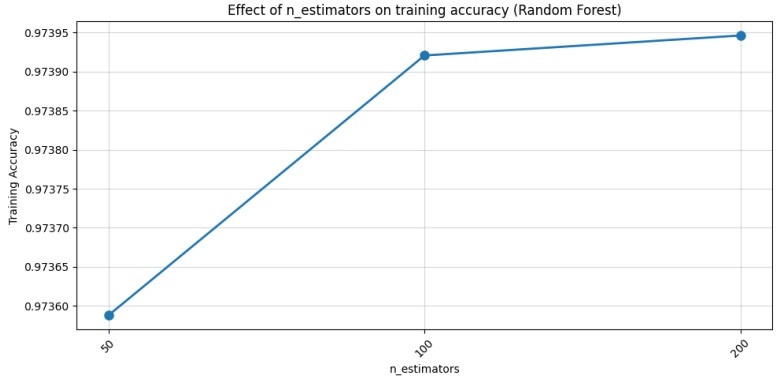
- Kernel: Determines shape of decision boundary (linear, poly and sigmoid for example).
- C: Margin strength/ Regularization parameter, a smaller C creates a soft margin classifier (widens the margin) and allows for more errors, whereas a larger C fits the data with a hard margin (higher variance and less bias).
- Gamma: (for RBF/poly/sigmoid) It is the kernel coefficient, a higher gamma can produce overfitting.
- Degree: For a poly kernel, it is the degree of the polynomial, a higher degree can increase variance similar to the gamma coefficient.
- Tolerance (tol): Determines the precision of the stopping criterion (in other words the optimality of the solution).

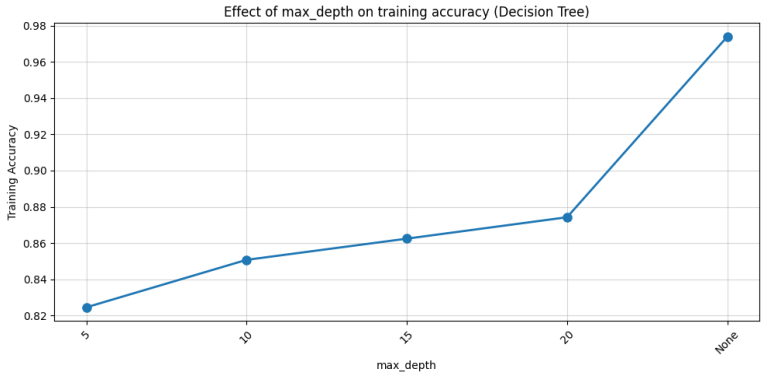
Question 3:

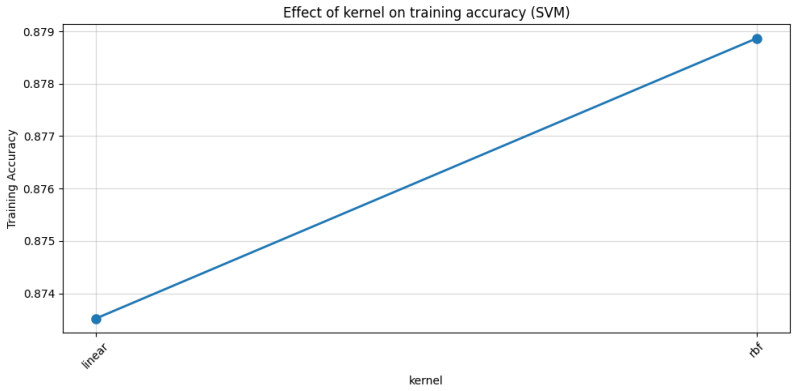
Given the limited data we have, cross validation is ideal to provide better estimates of the model's performance, it is well suited here because multiple hyperparameters are used and so they can be tested in different train-test splits. Stratified k-fold is used because it maintains the distribution of the classes in each split, in other words we obtain more consistent and less biased results, which of course leads to a more robust model performance.

Report

Question 1

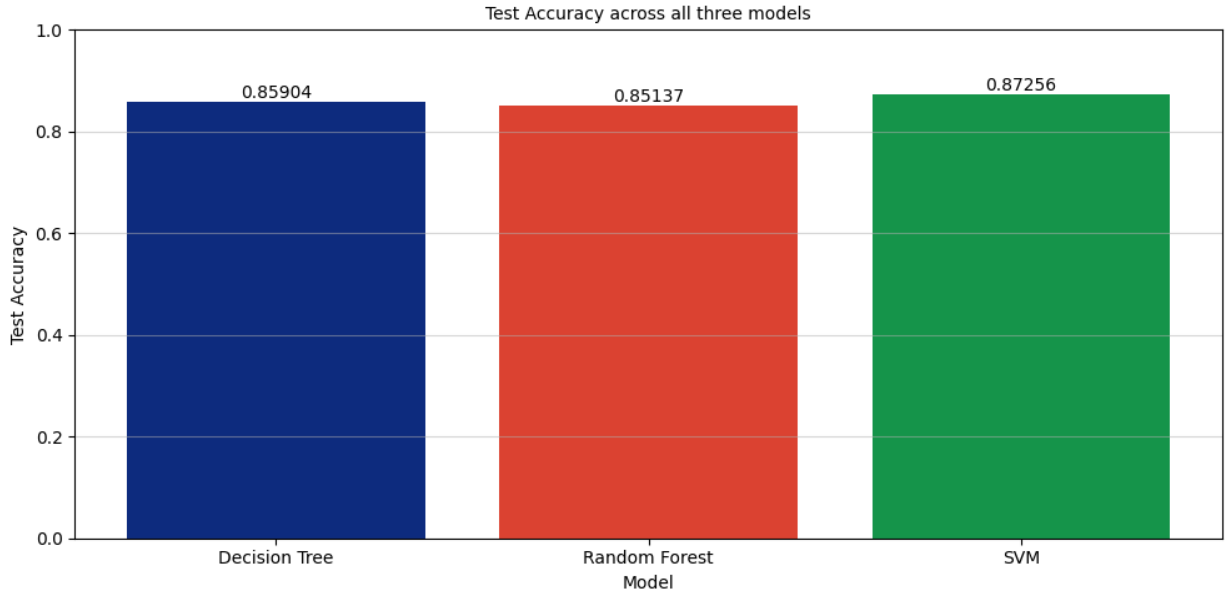






Since we're evaluating accuracy on the training set, a trend of increasing accuracy with larger hyperparameter values is expected. Increasing n_estimators and max_depth expands model flexibility and capacity which leads to overfitting. Similarly, using an RBF kernel instead of a linear one makes the SVM more flexible (higher variance), which also tends to overfit the training data.

Question 2



Model	Test Accuracy	Best Hyperparameters
SVM	0.8726	{'C': 10, 'gamma': 'scale', 'kernel': 'linear', 'tol': 0.0001}
Decision Tree	0.8590	{'criterion': 'gini', 'max_depth': None, 'max_leaf_nodes': 100, 'min_samples_leaf': 1}
Random Forest	0.8514	{'bootstrap': True, 'max_depth': None, 'n_estimators': 200}

Note: testing more SVM parameters was not possible due to the long computation time.

Question 3

The SVM achieved the best test score of 0.87256 (best out of the three models), the kernel used was the linear kernel with C = 10, this is an indication of the data being largely linearly separable. There is not a very big difference in accuracy across three models, so a definitive conclusion cannot be provided but as a hypothesis it seems that the trees have been trained to be slightly more sensitive to noise/outliers and SVM achieved better generalization due to the hyperparameters used.