Machine Learning Report: Telco Customer Churn

Created by: Solis, Alexander Jon S. Solis & Virata, Sean Maverick A.

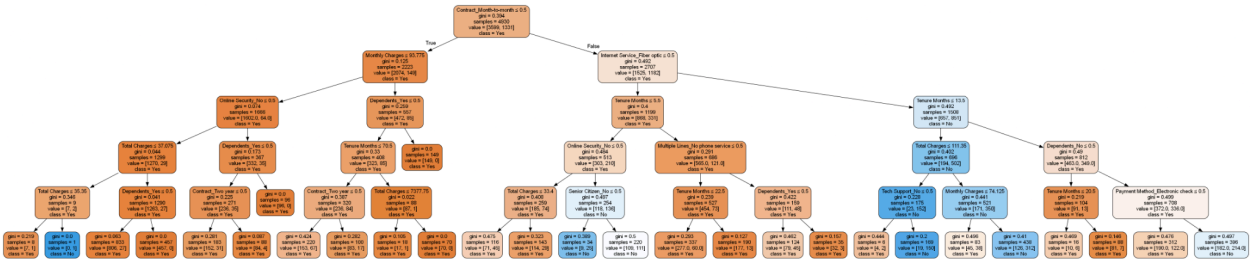
This project uses the Telco Customer Churn dataset. It contains demographic, account, and service-related features of telecom customers, along with a target variable Churn Label (Yes/No), which indicates whether a customer has churned. These are the following columns in the dataset:

- **CustomerID:** A unique ID that identifies each customer.
- **Count:** A value used in reporting/dashboarding to sum up the number of customers in a filtered set.
- **Country:** The country of the customer's primary residence.
- **State:** The state of the customer's primary residence.
- **City:** The city of the customer's primary residence.
- **Zip Code:** The zip code of the customer's primary residence.
- **Lat Long:** The combined latitude and longitude of the customer's primary residence.
- **Latitude:** The latitude of the customer's primary residence.
- **Longitude:** The longitude of the customer's primary residence.
- **Gender:** The customer's gender: Male, Female
- **Senior Citizen:** Indicates if the customer is 65 or older: Yes, No
- **Partner:** Indicate if the customer has a partner: Yes, No
- **Dependents:** Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.
- **Tenure Months:** Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above.
- **Phone Service:** Indicates if the customer subscribes to home phone service with the company: Yes, No
- **Multiple Lines:** Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No
- **Internet Service:** Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable.
- **Online Security:** Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No
- **Online Backup:** Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No
- **Device Protection:** Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No
- **Tech Support:** Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No
- **Streaming TV:** Indicates if the customer uses their Internet service to stream television programing from a third party provider: Yes, No. The company does not charge an additional fee for this service.
- **Streaming Movies:** Indicates if the customer uses their Internet service to stream movies from a third party provider: Yes, No. The company does not charge an additional fee for this service.

- **Contract:** Indicates the customer's current contract type: Month-to-Month, One Year, Two Year.
- **Paperless Billing:** Indicates if the customer has chosen paperless billing: Yes, No
- **Payment Method:** Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check
- **Monthly Charge:** Indicates the customer's current total monthly charge for all their services from the company.
- **Total Charges:** Indicates the customer's total charges, calculated to the end of the quarter specified above.
- **Churn Label:** Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value.
- **Churn Value:** 1 = the customer left the company this quarter. 0 = the customer remained with the company. Directly related to Churn Label.
- **Churn Score:** A value from 0-100 that is calculated using the predictive tool IBM SPSS Modeler. The model incorporates multiple factors known to cause churn. The higher the score, the more likely the customer will churn.
- **CLTV:** Customer Lifetime Value. A predicted CLTV is calculated using corporate formulas and existing data. The higher the value, the more valuable the customer. High value customers should be monitored for churn.
- **Churn Reason:** A customer's specific reason for leaving the company. Directly related to Churn Category.

After loading the data, we preprocessed it by removing unnecessary columns which are the following: 'Lat Long', 'Latitude', 'Longitude', 'Count', 'Streaming TV', 'Streaming Movies', 'Zip Code', 'Churn Reason', 'CustomerID', 'Country', 'State', 'City'. We also converted categorical variables into numeric form using one-hot encoding using this code: **pd.get_dummies(df_copy, columns=['Gender', 'Senior Citizen', 'Partner', 'Dependents', 'Phone Service', 'Multiple Lines', 'Internet Service', 'Online Security', 'Online Backup', 'Device Protection', 'Tech Support', 'Contract', 'Paperless Billing', 'Payment Method'], drop_first=False, dummy_na=False, dtype=int)**. We also cleaned the Total Charges column by replacing blanks with NaN, converting it to float, and filling missing values with the mean. We also dropped leakage columns which are the Churn Value, Churn Score, CLTV.

For this project, we used two models: the Decision Tree Classifier and the K-Nearest Neighbors(KNN) Classifier. The Decision Tree was used with parameters **max_depth=5** and **random_state=1** because it is highly interpretable, allowing us to understand the rules and feature importance behind churn predictions. This makes it valuable for identifying the key factors influencing the customer decisions, such as the contract type or monthly charges. On the other hand, KNN was applied with **n_neighbors=3** and scaled features using StandardScaler. KNN is well-suited for this use case because it classifies customers based on similarities with others, effectively capturing complex, nonlinear relationships in the data. It also achieved slightly better recall in identifying churners that are at risk of leaving. Together, these models balance interpretability and predictive power, making them strong candidates for churn analysis.



Towards the results for the models' performance, the Decision Tree achieved an overall accuracy of 0.79, meaning it correctly classified nearly 8 out of 10 customers. For non-churners (class No), the model performed strongly with precision and recall of 0.86, showing it is reliable at identifying customers who are likely to stay. However, for churners (class Yes), the performance was weaker, with a precision of 0.59, recall of 0.58, and an F1-score of 0.58. This indicates that while the model captures some customers at risk of leaving, it still misses a significant portion (false negatives) and occasionally misclassifies loyal customers as churners (false positives). The confusion matrix confirms this pattern, with 227 churners misclassified as non-churners and 217 non-churners misclassified as churners. Overall, the Decision Tree balances interpretability with solid predictive power, but its recall for churners leaves room for improvement.

Meanwhile, the KNN classifier achieved an accuracy of 0.75, slightly lower than the Decision Tree. For non-churners (class No), the model had good performance with precision and recall around 0.83–0.84. However, its performance for churners (class Yes) was weaker, with a precision of 0.51, recall of 0.49, and an F1-score of 0.50. This suggests that while KNN can identify churners to some extent, it struggles more than the Decision Tree, both missing actual churners and generating more false alarms. The confusion matrix shows 273 churners misclassified as non-churners and 255 non-churners incorrectly predicted as churners. Compared to the Decision Tree, KNN provides slightly worse results across all metrics, especially in detecting churners, highlighting that it may not be the best standalone model for this dataset.

Both models produced comparable results, with accuracy scores around 82–83% and F1-scores for churn ranging between 0.56 and 0.57. If the business objective is to catch as many churners as possible, which is essential for customer retention campaigns, KNN may be slightly preferable due to its higher recall. However, if the goal is to maintain interpretability and balanced predictions, the Decision Tree is a better option since it provides clear decision rules and valuable insights into the factors influencing churn. Moving forward, model performance can be enhanced by tuning hyperparameters (such as `max_depth` for the Decision Tree and `n_neighbors` for KNN), exploring ensemble methods like Random Forest or Gradient Boosting for improved recall and F1-scores, and applying cost-sensitive learning to give more weight to churn detection if identifying potential churners is a higher business priority than overall accuracy.

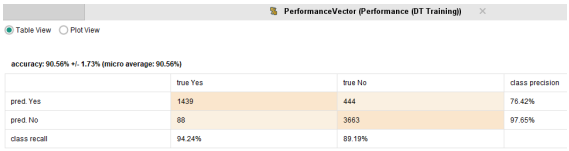

PerformanceVector (Performance (DT Training))

Table View | Plot View

accuracy: 90.56% +/- 1.73% (micro average: 90.56%)

	true Yes	true No	class precision
pred. Yes	1439	444	76.42%
pred. No	88	3663	97.65%
class recall	94.24%	89.19%	

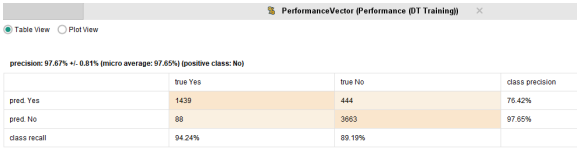

PerformanceVector (Performance (DT Training))

Table View | Plot View

precision: 97.67% +/- 0.81% (micro average: 97.65%) (positive class: No)

	true Yes	true No	class precision
pred. Yes	1439	444	76.42%
pred. No	88	3663	97.65%
class recall	94.24%	89.19%	

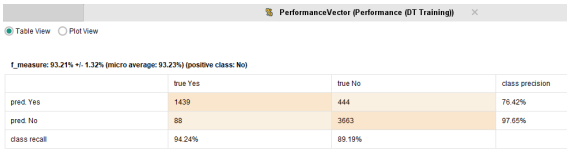

PerformanceVector (Performance (DT Training))

Table View | Plot View

f_measure: 93.21% +/- 1.32% (micro average: 93.23%) (positive class: No)

	true Yes	true No	class precision
pred. Yes	1439	444	76.42%
pred. No	88	3663	97.65%
class recall	94.24%	89.19%	

The Altair AI Studio, also known as RapidMiner, is also used to create a model that can predict customer churning using the Decision Tree classifier and compare it to the KNN model. Using the Decision Tree classifier, the model achieved an accuracy of 90.67%, a precision of 97.67%, and an F1-score of 93.21% during its training. The high accuracy and precision highlight the model's ability to predict customer churning during training.

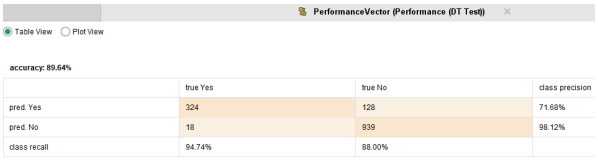

PerformanceVector (Performance (DT Test))

Table View | Plot View

accuracy: 89.64%

	true Yes	true No	class precision
pred. Yes	324	128	71.68%
pred. No	18	939	98.12%
class recall	94.74%	88.00%	

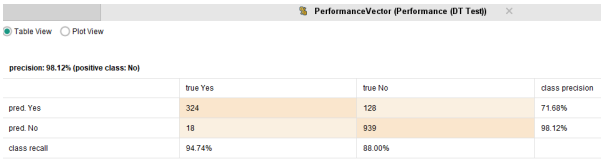

PerformanceVector (Performance (DT Test))

Table View | Plot View

precision: 98.12% (positive class: No)

	true Yes	true No	class precision
pred. Yes	324	128	71.68%
pred. No	18	939	98.12%
class recall	94.74%	88.00%	

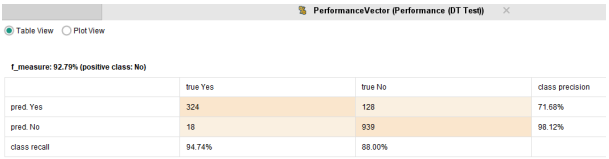
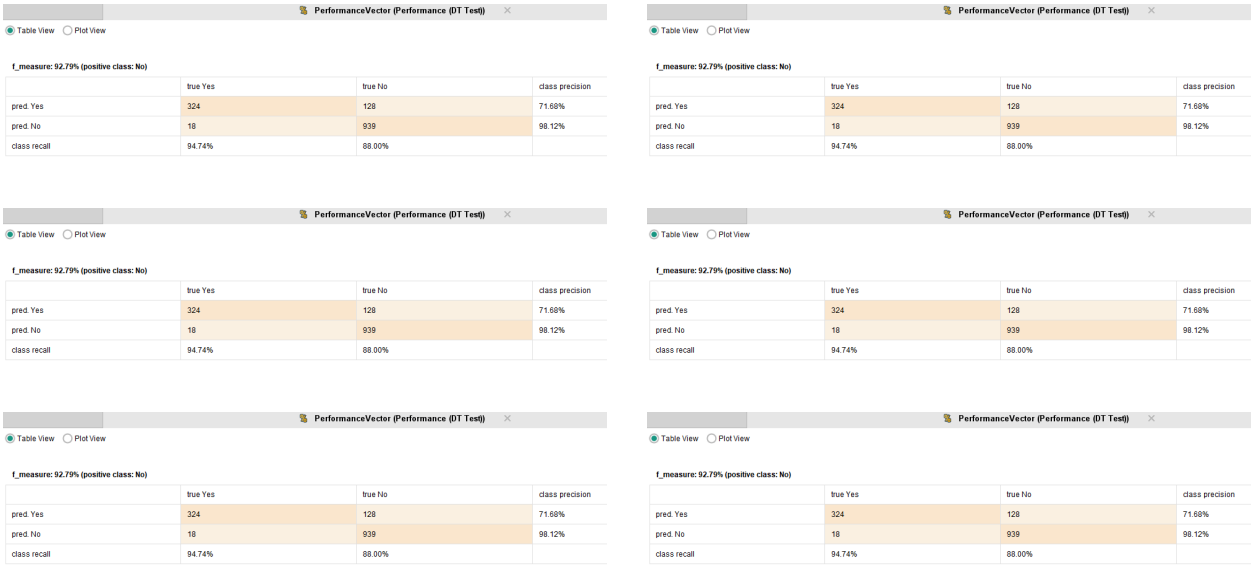

PerformanceVector (Performance (DT Test))

Table View | Plot View

f_measure: 92.79% (positive class: No)

	true Yes	true No	class precision
pred. Yes	324	128	71.68%
pred. No	18	939	98.12%
class recall	94.74%	88.00%	

However, the high accuracy slightly dropped when the test dataset was used. During the validation test, it achieved an accuracy of 89.64%, a precision of 98.12%, and an F1-score of 92.79%. The minimal difference in the accuracy and precision results indicates that the model is not overfitting with the dataset.



Against the KNN model, the Decision Tree model has performed better. During the training validation, the KNN has achieved an accuracy of 87.72%, a precision of 89.60%, and an F1-score of 91.78%. The test validation result is only slightly better compared to the training validation, with an F1-score of 92.59%. Alone, the KNN model has achieved high performance and demonstrated its ability to accurately determine the churning customers. However, the Decision Tree model has shown a marginal performance advantage in accuracy, precision, and F1-score over the KNN model. This makes the Decision Tree a better classifier when using the Altair AI Studio. With that being said, the difference between the two is minimal to be able to determine a superior model.

In hindsight, depending on the program used to create the models, they can either perform similarly with a minute difference in performance or have some perceivable differences between the KNN and Decision tree algorithm. However, this project also demonstrated that it is feasible to create an model that can accurately predict a customer churning using two different algorithms in two completely different platforms.