جامعة الأمير سلطان
PRINCE SULTAN UNIVERSITY

CCIS كلية علوم الحاسب والمعلومات
COLLEGE OF COMPUTER & INFORMATION SCIENCES

Data Mining Project

Prepared by

OSAMAH ALDHUWAYHI

Hussam Aldossary

MAHMOOD JAMALULLAIL

Supervised by

Dr. Esam Mohammed Othman

Table of Contents

1. INTRODUCTION 3
   1.1. PURPOSE 3
   1.2. SCOPE 3
2. Paper Summary 4
   2.1. Research Problem 4
   2.2. Methodology Used in the Paper 4
   2.3. Algorithms Used 6
   2.4. Results Achieved in the Paper 6
3. Dataset Choice and Preprocessing 8
   3.1. Dataset Description 8
   3.2. Justification for Dataset Selection 10
   3.3. Data Preprocessing 11
4. Methodology and Experimental Setup 12
   4.1. Selected Mining Technique 12
   4.2. Algorithms Applied 12
   4.3. Experimental Setup 13
   4.4. Evaluation Metrics 16
5. Results and Analysis 17
   5.1. Classification Accuracy and Detailed Metrics 17
   5.2. Confusion Matrices for Each Classifier 18
   5.3. Comparison with Paper's Reported Results 20
   5.4. Summary result 21
6. REFERENCES 22
7. Appendix 23

1 INTRODUCTION

1.1 Purpose

In today's competitive market, service-based companies are highly dependent on customer satisfaction and long-term value generation. Some companies' performance decreases significantly due to the customer's unexpected behavior, especially ‘Churn’, where the customer decides to exit the company to its competitors without any prior notice. So, our project aims to understand, analyze, and predict the customer behavior from the dataset called “Telco Customer Churn “to lower the unexpected behavior and to facilitate risk mitigation for the company with actionable insights. Furthermore, to gain more valuable insights and cover the hidden patterns that the paper did not recover.

1.2 SCOPE

For the scope, our source is limited to the “Telco Customer Churn” dataset, where it covers more than 7,000+ customer records and 20+ distinct attributes. The software tools we will use is WEKA software, where we will do the preprocessing, classification and evaluation technique. Our project focuses on Classification (predictive model) to justify whether the customer will churn by conducting a binary classification using (“Yes”) or stay as (“No”). Moreover, our scope includes a direct comparison, where we will compare our results that are generated in WEKA with the findings that are presented in the selected research paper. The deliverables of the project consists of 6 pages that are uploaded as a PDF report, the modified ARFF dataset and visual evidence (Screenshorts) of the WEKA experimental results.

2 Paper Summary

2.1 Research Problem

The telecommunications industry faces intense competition, and customer churn (voluntarily switching to another provider) is a major revenue threat. The paper addresses the problem of predicting which customers are likely to churn so that telecom companies can take proactive retention actions. The challenge is to build an accurate prediction model using big data (70+ Terabytes) from multiple sources (call details, SMS, internet usage, complaints, IMEI data, etc.) over a 9-month period.

2.2 Methodology Used in the Paper

The authors used a supervised classification approach on a big data platform (Hortonworks Data Platform – HDP) with Spark as the execution engine. The key methodological steps were:

1. Data Acquisition & Storage – Used Apache Flume (for unstructured/semi-structured data) and Apache SQOOP (for structured relational data) to move data into HDFS. Data was stored in Parquet format with Snappy compression for efficiency.

2. Feature Engineering – This was the most time-consuming phase. They engineered over 2,000 features from:

- Statistical features: Aggregated values (average, count, sum, max, min) per customer per day/week/month from CDRs (calls, SMS, MMS, internet), plus complaint data, IMEI data, and marketing-derived features.
- Social Network Analysis (SNA) features: Built a social graph with 15 million nodes (customers) and 2.5 billion edges (interactions). Extracted features like PageRank, SenderRank, In/Out Degree, Neighbor Connectivity, Local Clustering Coefficient, Jaccard Similarity, and Cosine Similarity.

3. Feature Transformation & Selection – Removed constant/nearly constant features, handled missing values (deleted records with >90% missing, deleted features with >70% missing, filled categorical missing with "Other", filled numerical missing with mean), removed correlated features using Pearson correlation, and reduced categorical features to top 32 categories.

4. Handling Class Imbalance – The churn class was only 5% of the dataset. They experimented with oversampling (duplicating churn class), undersampling (reducing active class), and no rebalancing.

5. Classification – Used 10-fold cross-validation with 70% training / 30% testing split. Applied four tree-based algorithms:

- Decision Tree (optimized depth=20, nodes=398)
- Random Forest (optimized trees=200)
- Gradient Boosted Machine (GBM) (optimized trees=200)
- XGBOOST (optimized trees=180)

2.3 Algorithms Used

Algorithm	Type	Optimized Hyperparameters
Decision Tree	**Tree-based**	**Depth = 20, Max Nodes = 398**
Random Forest	**Ensemble (Bagging)**	**Number of trees = 200**
GBM	**Ensemble (Boosting)**	**Number of trees = 200**
XGBOOST	**Ensemble (Boosting)**	**Number of trees = 180**

2.4 Results Achieved in the Paper

Algorithm	AUC (Statistical Features Only)	AUC (SNA Features Only)	AUC (Both Features)
XGBOOST	84%	75.3%	93.3%
GBM	82%	71%	90.8%
Random Forest	79%	69%	87.76%
Decision Tree	76%	67.2%	83%

Key Findings:

1. **XGBoost Performance:**
    - XGBoost outperformed all other algorithms with an AUC of 93.3%.
2. **Impact of SNA Features:**
    - Adding Social Network Analysis (SNA) features significantly improved performance, raising XGBoost's AUC from 84% to 93.3%.
3. **Optimal Sliding Windows:**
    - Statistical features performed best using the last 6 months of data.
    - SNA features performed best using the last 4 months of data.
4. **Rebalancing Strategies:**
    - XGBoost and GBM achieved best results without rebalancing.
    - Random Forest and Decision Tree performed best with undersampling.
5. **Top 5 Features for Churn Prediction:**
    - **MTN Cosine Similarity (SNA):** Higher similarity to competitors' networks increases churn risk.
    - **Days Since Last Outgoing Transaction (Statistical):** Longer inactivity increases churn likelihood.
    - **Total Balance (Statistical):** Lower balances are associated with higher churn.
    - **Average Radio Access Type (Statistical):** More 2G usage (slower internet) increases churn risk.
    - **Local Cluster Coefficient (SNA):** Very low values (typical for business users) indicate lower churn probability.

3 Dataset Choice and Preprocessing

3.1 Dataset Description

The dataset used in this project is the Telco Customer Churn dataset, publicly available on Kaggle. It contains 7,043 customer records with 21 attributes (20 features + 1 target variable). The data represents a telecommunications company's customer base, including demographic information, account details, service subscriptions, and usage patterns.

The target variable is Churn, which indicates whether a customer has left the company within the last month. The dataset has a class imbalance issue, with approximately 73.5% Non-Churn (No) and 26.5% Churn (Yes) .

Attribute summary:

Attribute Name	Type	Description
customerID	**Nominal (String)**	**Unique identifier for each customer – removed**
gender	**Nominal**	**Male / Female**
SeniorCitizen	**Binary (0/1)**	**1 = Senior citizen, 0 = Not senior**
Partner	**Nominal**	**Yes / No – whether customer has a partner**
Dependents	**Nominal**	**Yes / No – whether customer has dependents**
tenure	**Numeric**	**Number of months the customer has stayed with the company**
PhoneService	**Nominal**	**Yes / No – whether customer has phone service**

MultipleLines	**Nominal**	**Yes / No / No phone service – whether customer has multiple lines**
InternetService	**Nominal**	**DSL / Fiber optic / No – type of internet service**
OnlineSecurity	**Nominal**	**Yes / No / No internet service – whether customer has online security**
OnlineBackup	**Nominal**	**Yes / No / No internet service – whether customer has online backup**
DeviceProtection	**Nominal**	**Yes / No / No internet service – whether customer has device protection**
TechSupport	**Nominal**	**Yes / No / No internet service – whether customer has tech support**
StreamingTV	**Nominal**	**Yes / No / No internet service – whether customer has streaming TV**
StreamingMovies	**Nominal**	**Yes / No / No internet service – whether customer has streaming movies**
Contract	**Nominal**	**Month-to-month / One year / Two year – type of contract**
PaperlessBilling	**Nominal**	**Yes / No – whether customer uses paperless billing**
PaymentMethod	**Nominal**	**Electronic check / Mailed check / Bank transfer (automatic) / Credit card (automatic)**

MonthlyCharges	**Numeric**	**Monthly amount charged to the customer (USD)**
TotalCharges	**Numeric**	**Total amount charged to the customer (USD) – contains missing values**
Churn (target)	**Nominal**	**Yes / No – whether the customer churned**

3.2 Justification for Dataset Selection

1. Relevance to the Paper – The paper focuses on customer churn prediction in the telecom industry. The Telco Customer Churn dataset directly addresses the same business problem, making it suitable for comparing results with the paper's findings.
2. Public Availability – The dataset is freely available on Kaggle, ensuring reproducibility of experiments.
3. Appropriate Size for WEKA – With 7,043 instances and 21 attributes, the dataset is small enough to run efficiently in WEKA on a standard computer, unlike the paper's 70+ TB big data environment.
4. Mix of Data Types – The dataset includes both numerical (tenure, MonthlyCharges, TotalCharges) and nominal (gender, InternetService, Contract, etc.) attributes, allowing demonstration of various preprocessing and classification techniques.
5. Class Imbalance – The dataset exhibits class imbalance (26.5% churn), which is a real-world challenge discussed in the paper. This allows us to experiment with balancing techniques (e.g., SMOTE) in the extended analysis section.
6. Missing Values – The TotalCharges attribute contains missing values, providing an opportunity to demonstrate data cleaning.
7. Direct Comparison – Many published churn prediction studies have used this exact dataset, making it easier to benchmark our WEKA results against existing literature.

3.3 Data Preprocessing

Step	Action	Tool/Method
1	Load CSV	WEKA Explorer – Open file
2	Handle missing values (TotalCharges)	ReplaceMissingValues filter (replace with mean)
3	Remove customerID	Manual attribute removal
4	Verify attribute types	WEKA automatic detection + manual check
5	Save as ARFF	WEKA – Save button

4 Methodology and Experimental Setup

4.1 Selected Mining Technique

Classification is the standard approach for churn prediction, as evidenced by the paper and numerous other studies.The data mining technique selected for this project is supervised classification. This is appropriate because:

1. The target variable (Churn) is nominal with two classes (Yes / No).
2. The goal is to predict whether a new customer will churn based on historical labeled data.
3. Classification is the standard approach for churn prediction, as evidenced by the paper and numerous other studies.

4.2 Algorithms Applied

Algorithm	Category	Why Selected	Corresponding Algorithm in Paper
Random Forest	**Ensemble (Bagging)**	**Handles non-linear relationships well; robust to overfitting; top performer in many churn studies**	**Random Forest (used in paper, 87.76% AUC)**
J48 (Decision Tree)	**Tree-based**	**Highly interpretable; produces a visual tree; good baseline**	**Decision Tree (used in paper, 83% AUC)**
Logistic Regression	**Linear**	**Simple, fast, interpretable; provides probability scores**	**Not directly used in paper (paper focused on tree-based methods)**
Naive Bayes	**Probabilistic**	**Fast, simple baseline; assumes feature independence**	**–**

4.3 Experimental Setup

All experiments were conducted using **WEKA** on the preprocessed dataset **telco_churn.arff** , which contains 7,043 instances, 19 features, and the target variable **Churn**.

4.3.1 Evaluation Method: Stratified Cross-Validation

To ensure reliable performance estimation, **10-fold stratified cross-validation** was employed.

- The dataset is divided into 10 equally sized folds.
- In each iteration, 9 folds are used for training and 1 fold for testing.
- This process is repeated 10 times, and the results are averaged.
- Stratification ensures that each fold preserves the original class distribution (approximately 73.5% *No* and 26.5% *Yes*).

This approach is particularly important for imbalanced datasets, as it prevents biased evaluation caused by uneven class representation. Compared to a simple train/test split, cross-validation provides a more robust and less variance-prone estimate of model performance. It is also consistent with the evaluation strategy used in the reference paper.

4.3.2 Parameter Tuning

To enhance model performance and meet project requirements, manual parameter tuning was performed instead of relying solely on default settings. The final configurations are summarized below:

Algorithm	Parameter	Value	The Reason
Naive Bayes	useSupervisedDiscretization	True	Improves handling of numerical features by converting them into intervals based on class information.
Logistic Regression	maxIts	200	Ensures convergence of the optimization process.
	ridge	1.0E-7	Prevents overfitting by stabilizing coefficient estimates.
J48	minNumObj	5	Reduces overfitting by requiring a minimum number of instances per leaf.
	confidenceFactor	0.15	Produces more aggressive pruning, resulting in a simpler and more generalizable tree.
Random Forest	numIterations	150	Provides a balance between accuracy and computational efficiency.
	maxDepth	12	Limits tree complexity to prevent overfitting.
	seed	1	Ensures reproducibility of results.

4.3.3 Handling Class Imbalance

The dataset exhibits class imbalance, with approximately 73.5% non-churn and 26.5% churn instances. This imbalance can lead to biased models that favor the majority class.

To address this issue, the following strategies were applied:

1. **Stratified cross-validation** to maintain class proportions in all folds.
2. Use of multiple evaluation metrics beyond accuracy, including precision, recall, F1-score, and ROC Area (AUC).
3. In the extended analysis (Section 6), **SMOTE (Synthetic Minority Over-sampling Technique)** is applied to further balance the dataset and assess its impact on performance.

4.3.4 Experimental Procedure

The experiments were conducted in WEKA Explorer using the following steps:

1. Load the telco_churn.arff dataset in the Preprocess tab.
2. Navigate to the **Classify** tab.
3. Select the desired classifier (e.g., trees.RandomForest).
4. Configure the parameters according to Section 4.3.2.
5. Choose **Cross-validation** under Test Options and set the number of folds to 10.
6. Ensure that Churn is set as the target class.
7. Execute the model by clicking **Start**.
8. Record the results, including accuracy, precision, recall, F1-score, ROC Area, and the confusion matrix.

4.4 Evaluation Metrics

The following metrics were used to evaluate model performance:

Metric	Definition	Business Relevance
Accuracy	(TP + TN) / (TP + TN + FP + FN)	Measures overall correctness but may be misleading for imbalanced datasets.
Precision (Churn = Yes)	TP / (TP + FP)	Indicates how many predicted churners actually churned, minimizing unnecessary retention costs.
Recall (Churn = Yes)	TP / (TP + FN)	Measures how effectively actual churners are identified.
F1-Score (Churn)	Harmonic mean of precision and recall	Balances precision and recall.
ROC Area (AUC)	Area under the ROC curve	Evaluates the model's ability to distinguish between classes; robust to imbalance.
Confusion Matrix	TP, TN, FP, FN counts	Provides detailed insight into prediction errors.

5. Results and Analysis

This section presents a comprehensive analysis of the experimental results obtained from the four classification models—Naive Bayes, Logistic Regression, J48, and Random Forest. All models were evaluated using 10-fold stratified cross-validation on the preprocessed Telco Customer Churn dataset. The evaluation includes overall accuracy, class-specific performance metrics, confusion matrices, and ROC Area (AUC). Furthermore, the findings are compared with those reported in Ahmad et al. (2019) churn prediction study to provide broader context and insight.

5.1 Classification Accuracy and Detailed Metrics

The overall performance of the classifiers is summarized in the table below:

Algorithm	**Accuracy (%)**	**Kappa**	**Weighted F-Mesure**	**ROC Area (AUC)**	**Precision (Yes)**	**Recall (Yes)**
Naive Bayes	73.08	0.4238	0.746	0.824	0.496	**0.803**
Logistic Regression	**80.22**	**0.4644**	**0.796**	**0.845**	**0.653**	0.545
J48	79.24	0.4323	0.785	0.815	0.635	0.512
Random Forest	79.55	0.4400	0.788	0.833	0.643	0.515

Key Observations:

- **Logistic Regression** emerged as the best-performing model overall, achieving the highest accuracy (80.22%), Kappa statistic, F-measure, and AUC (0.845). This demonstrates that even a relatively simple linear model can perform exceptionally well when the dataset features are informative.
- **Random Forest** and **J48** produced highly comparable results, with Random Forest slightly outperforming J48 across most metrics. This aligns with the well-established advantage of ensemble methods over single decision trees.
- **Naive Bayes**, although yielding the lowest accuracy (73.08%), achieved the **highest recall (0.803)** for the churn class. This indicates its strong ability to detect actual churners, albeit at the expense of lower precision.

From a business perspective, recall is particularly critical, as failing to identify a potential churner may result in direct revenue loss. Therefore, models with higher recall may still be valuable despite lower overall accuracy.

5.2 Confusion Matrices for Each Classifier

The confusion matrices below provide a detailed breakdown of prediction outcomes. Rows represent the actual class, while columns represent the predicted class.

Naive Bayes

Actual \ Predicted	No	Yes
No	3646	1528
Yes	368	1501

Logistic Regression

Actual \ Predicted	No	Yes
No	4632	542
Yes	851	1018

J48 (Decision Tree)

Actual \ Predicted	No	Yes
No	4625	549
Yes	913	956

Random Forest

Actual \ Predicted	No	Yes
No	4640	534
Yes	906	963

Interpretation and Insights:

- **Naive Bayes** demonstrates exceptional sensitivity to churners, correctly identifying 1,501 out of 1,869 cases. However, it produces a large number of false positives (1,528), which could lead to unnecessary allocation of retention resources.
- **Logistic Regression** provides the most balanced performance, maintaining relatively low false positives (542) while still capturing a reasonable number of churners. This balance explains its superior overall metrics.
- **Random Forest** is the most conservative model, producing the lowest number of false positives (534). While this minimizes unnecessary interventions, it also results in a higher number of missed churners.
- **J48** performs similarly to Random Forest, reinforcing the idea that decision tree-based methods are competitive, though slightly less effective than ensemble approaches.

Overall, the confusion matrices highlight an important trade-off: models that aggressively detect churners tend to increase false positives, whereas conservative models reduce false alarms but miss more actual churn cases.

5.3 Comparison with the Research Study

The results of this project can be meaningfully compared with those reported in Ahmad et al. (2019) churn prediction study. In that study, the authors utilized a large-scale big data environment (Spark and Hadoop) and engineered over 2,000 features, including Social Network Analysis (SNA) attributes. Their best-performing model, XGBoost, achieved an AUC of 93.3%.

A comparison is shown below:

Model (This Study)	Accuracy	AUC	Paper's Equivalent	Paper's AUC
Logistic Regression	80.22%	0.845	Statistical baseline	~84%
Random Forest	79.55%	0.833	Random Forest	87.76%
J48	79.24%	0.815	Decision Tree	83%
Naive Bayes	73.08%	0.824	Not included	–
Best in paper	–	–	XGBoost + SNA	**93.3%**

Reasons for Performance Differences:

Several key factors explain why the results in this project are lower than those reported in the paper:

- The paper utilized a **massive dataset** (10 million customers), whereas this project used only 7,043 instances.
- Over **2,000 engineered features** were used in the paper, compared to only 19 features in this dataset.
- The inclusion of **Social Network Analysis (SNA)** significantly enhanced predictive performance in the paper.
- The paper employed **XGBoost**, a highly advanced algorithm not available in standard WEKA.
- Differences in class imbalance (5% vs. 26.5%) also influenced model behavior and evaluation metrics.

Key Takeaways:

- Ensemble methods consistently outperform single decision trees.
- AUC is a reliable metric for evaluating imbalanced classification problems.
- Feature engineering plays a crucial role in improving predictive performance.
- Even simple models, such as Logistic Regression, can achieve strong results when applied to well-structured data.

5.4 Summary of Results

In summary, the experimental findings can be interpreted as follows:

- **Best overall model:** Logistic Regression (80.22% accuracy, AUC = 0.845).
- **Best at detecting churners:** Naive Bayes (highest recall = 0.803).
- **Most conservative model:** Random Forest (lowest false positives).
- **Performance ranking (AUC):** Logistic Regression > Random Forest > Naive Bayes > J48.

REFERENCES

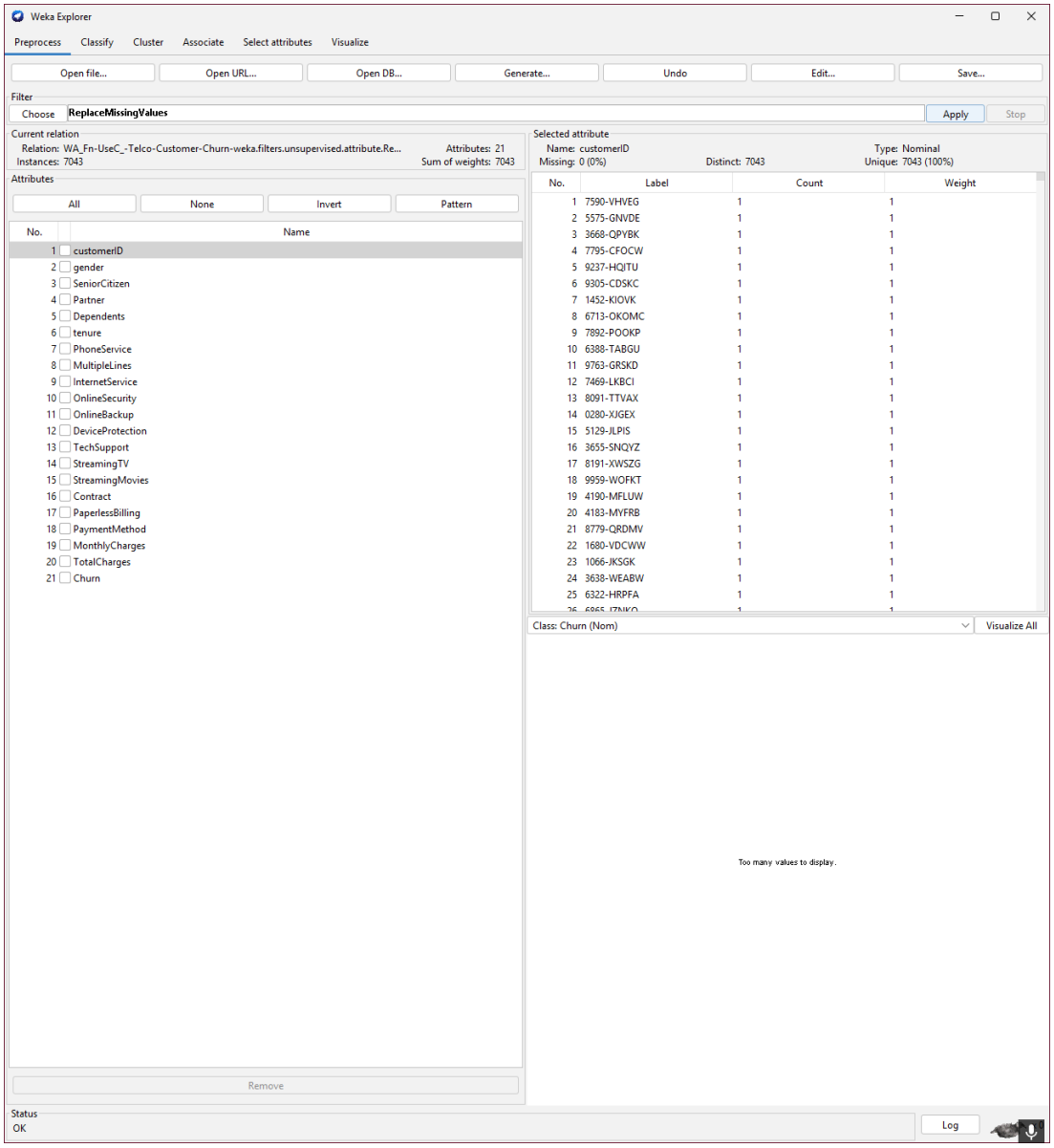
Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6(1), 1–24. **https://doi.org/10.1186/s40537-019-0191-6**

IBM. (2019). *Telco customer churn dataset* [Data file]. Kaggle**.**
https://www.kaggle.com/datasets/blastchar/telco-customer-churn

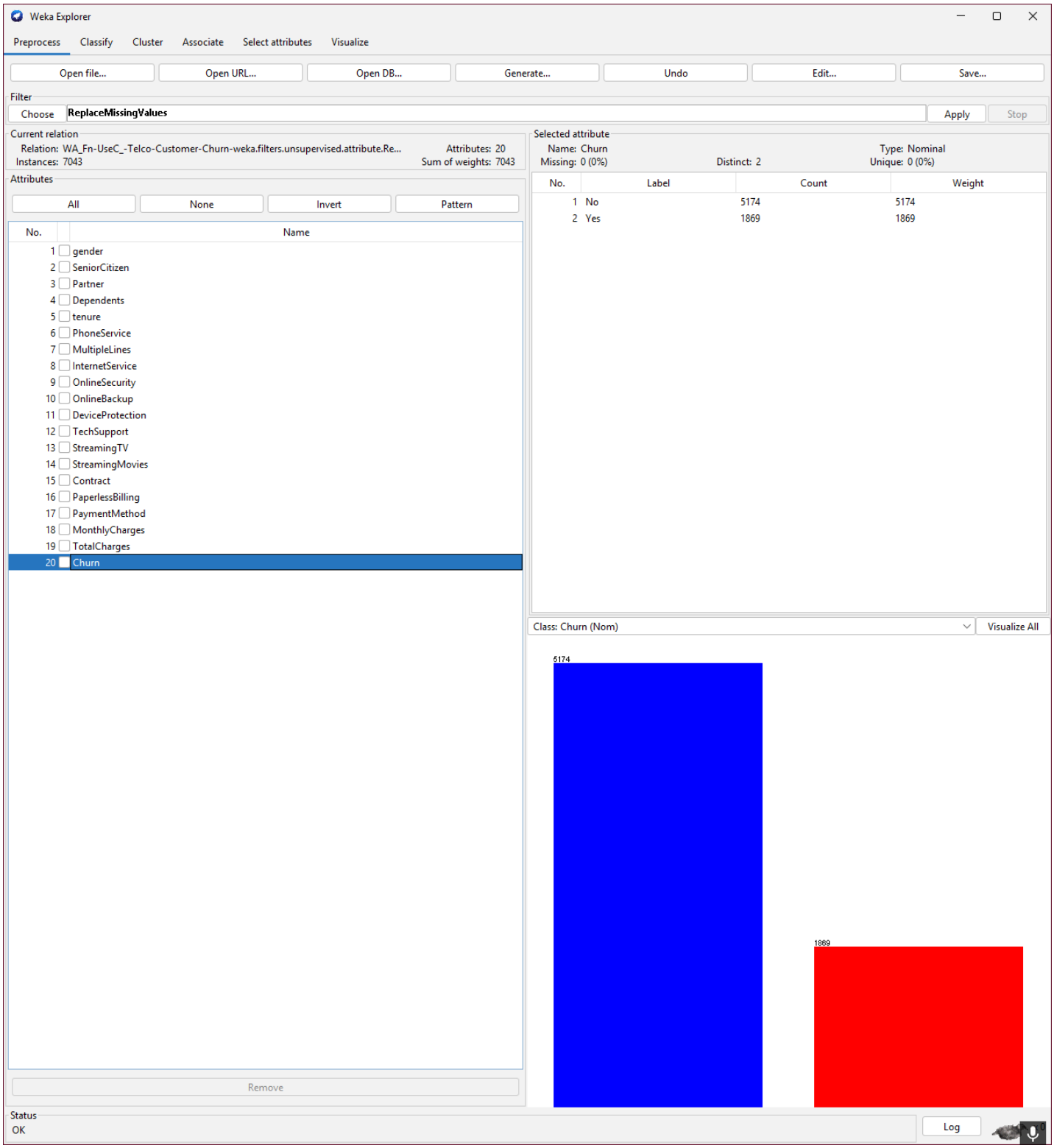
The University of Waikato. (2024). *WEKA: Waikato Environment for Knowledge Analysis* [Computer software and documentation]. **https://www.cs.waikato.ac.nz/ml/weka/**

7 appendix

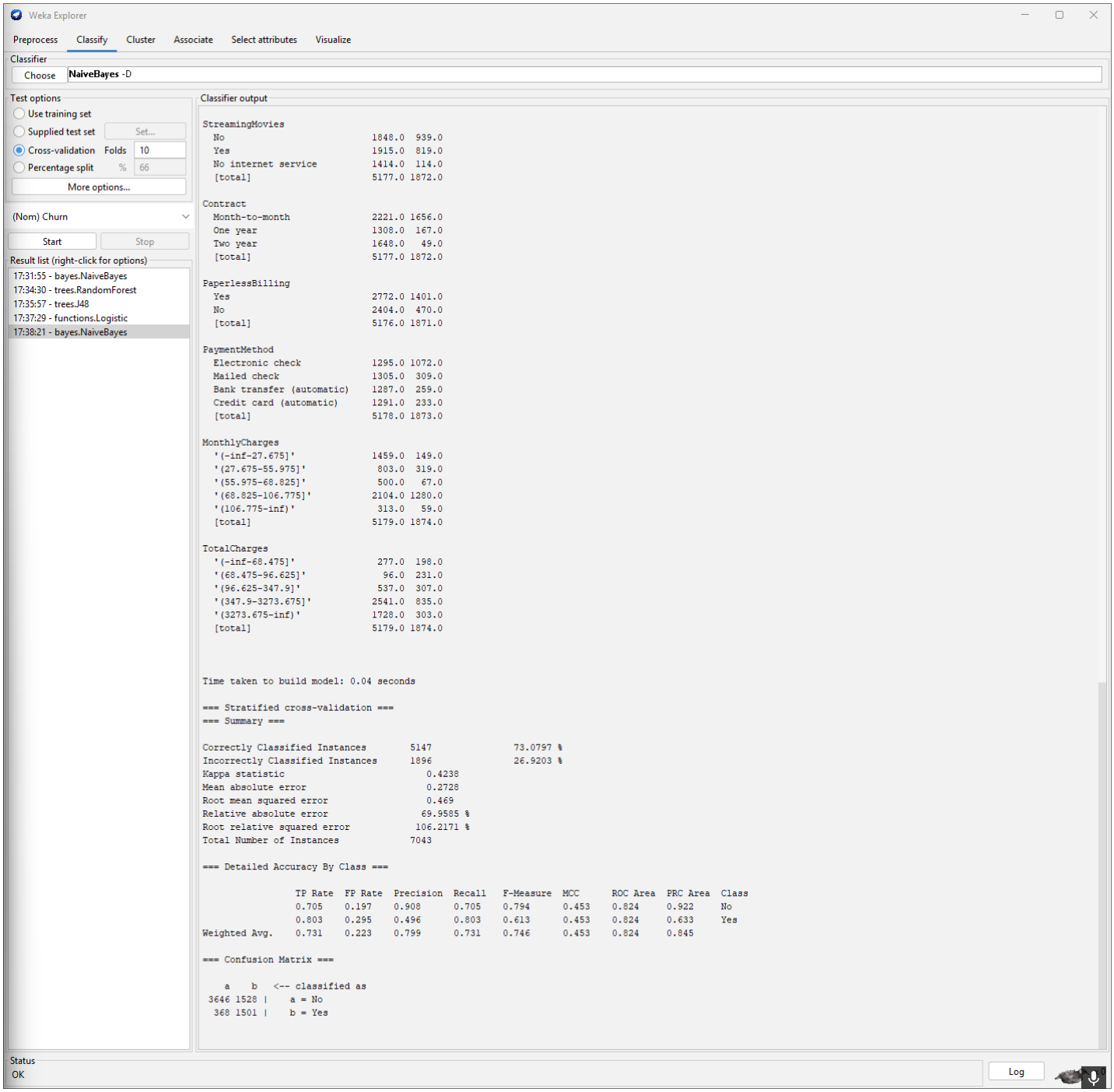
Replace Missing Values:



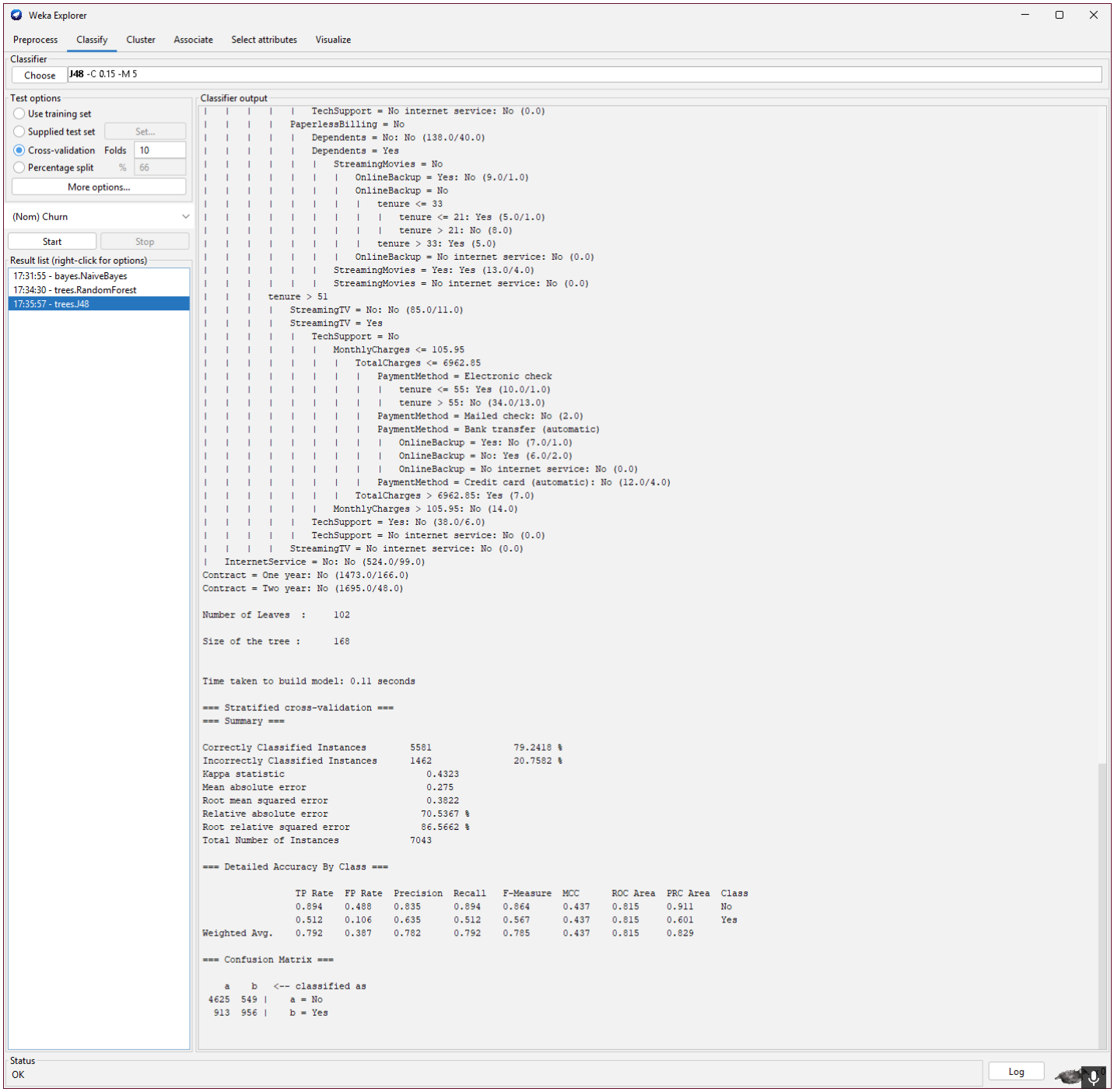
Remove CustomerID:



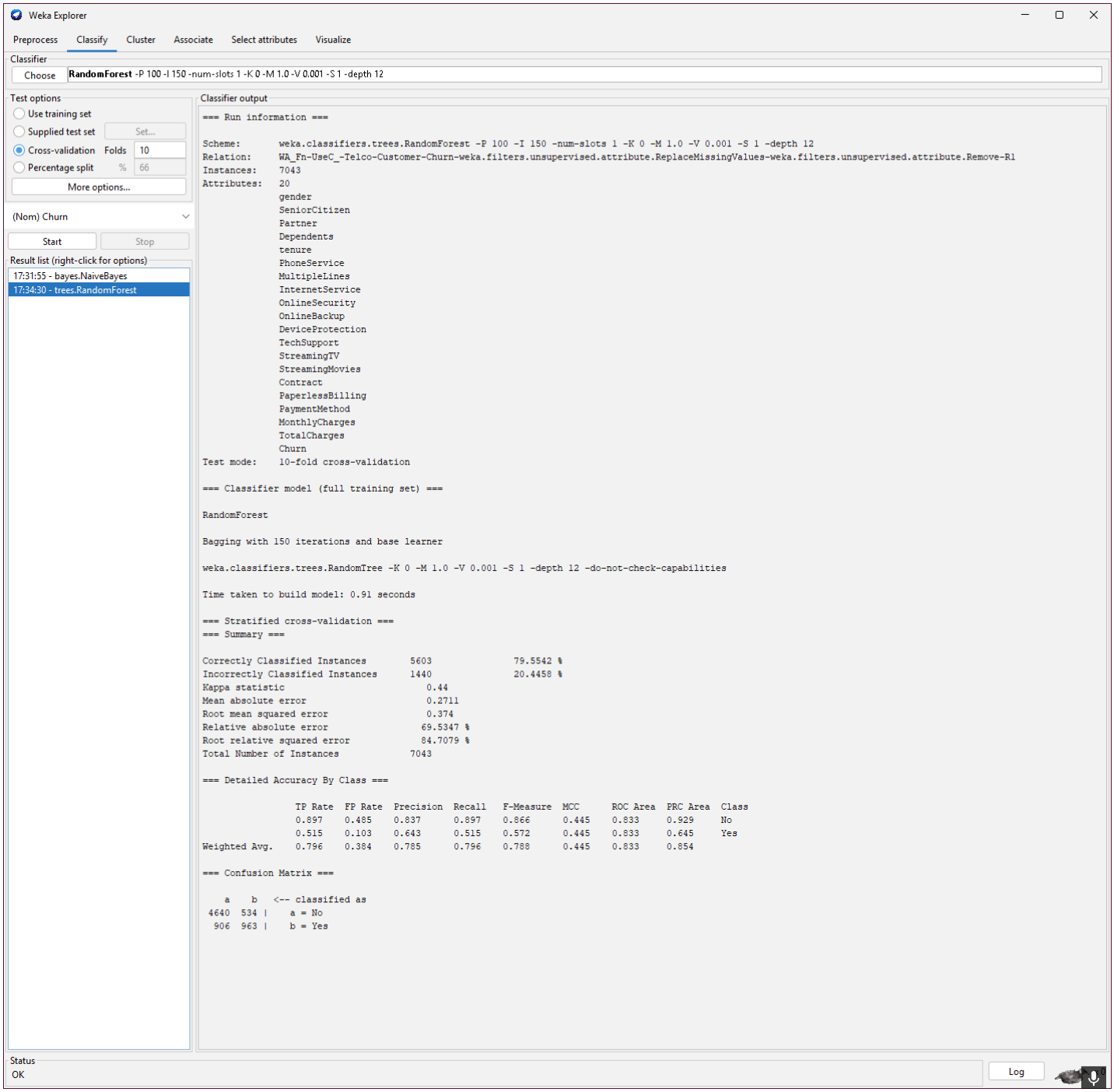
NaiveBayes


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes -D
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Churn
Start
Stop
Result list (right-click for options)
17:31:55 - bayes.NaiveBayes
17:34:30 - trees.RandomForest
17:35:57 - trees.J48
17:37:29 - functions.Logistic
17:38:21 - bayes.NaiveBayes
Classifier output
StreamingMovies
No 1848.0 939.0
Yes 1915.0 819.0
No internet service 1414.0 114.0
[total] 5177.0 1872.0
Contract
Month-to-month 2221.0 1656.0
One year 1308.0 167.0
Two year 1648.0 49.0
[total] 5177.0 1872.0
PaperlessBilling
Yes 2772.0 1401.0
No 2404.0 470.0
[total] 5176.0 1871.0
PaymentMethod
Electronic check 1295.0 1072.0
Mailed check 1305.0 309.0
Bank transfer (automatic) 1287.0 259.0
Credit card (automatic) 1291.0 233.0
[total] 5178.0 1873.0
MonthlyCharges
'(-inf-27.675]' 1459.0 149.0
'(27.675-55.975]' 803.0 319.0
'(55.975-68.825]' 500.0 67.0
'(68.825-106.775]' 2104.0 1280.0
'(106.775-inf)' 313.0 59.0
[total] 5179.0 1874.0
TotalCharges
'(-inf-68.475]' 277.0 198.0
'(68.475-96.625]' 96.0 231.0
'(96.625-347.9]' 537.0 307.0
'(347.9-3273.675]' 2541.0 835.0
'(3273.675-inf)' 1728.0 303.0
[total] 5179.0 1874.0
Time taken to build model: 0.04 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 5147 73.0797 %
Incorrectly Classified Instances 1896 26.9203 %
Kappa statistic 0.4238
Mean absolute error 0.2728
Root mean squared error 0.469
Relative absolute error 69.9585 %
Root relative squared error 106.2171 %
Total Number of Instances 7043
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.705 0.197 0.908 0.705 0.794 0.453 0.824 0.922 No
0.803 0.295 0.496 0.803 0.613 0.453 0.824 0.633 Yes
Weighted Avg. 0.731 0.223 0.799 0.731 0.746 0.453 0.824 0.845
=== Confusion Matrix ===
a b <-- classified as
3646 1528 | a = No
368 1501 | b = Yes
Status
OK
Log

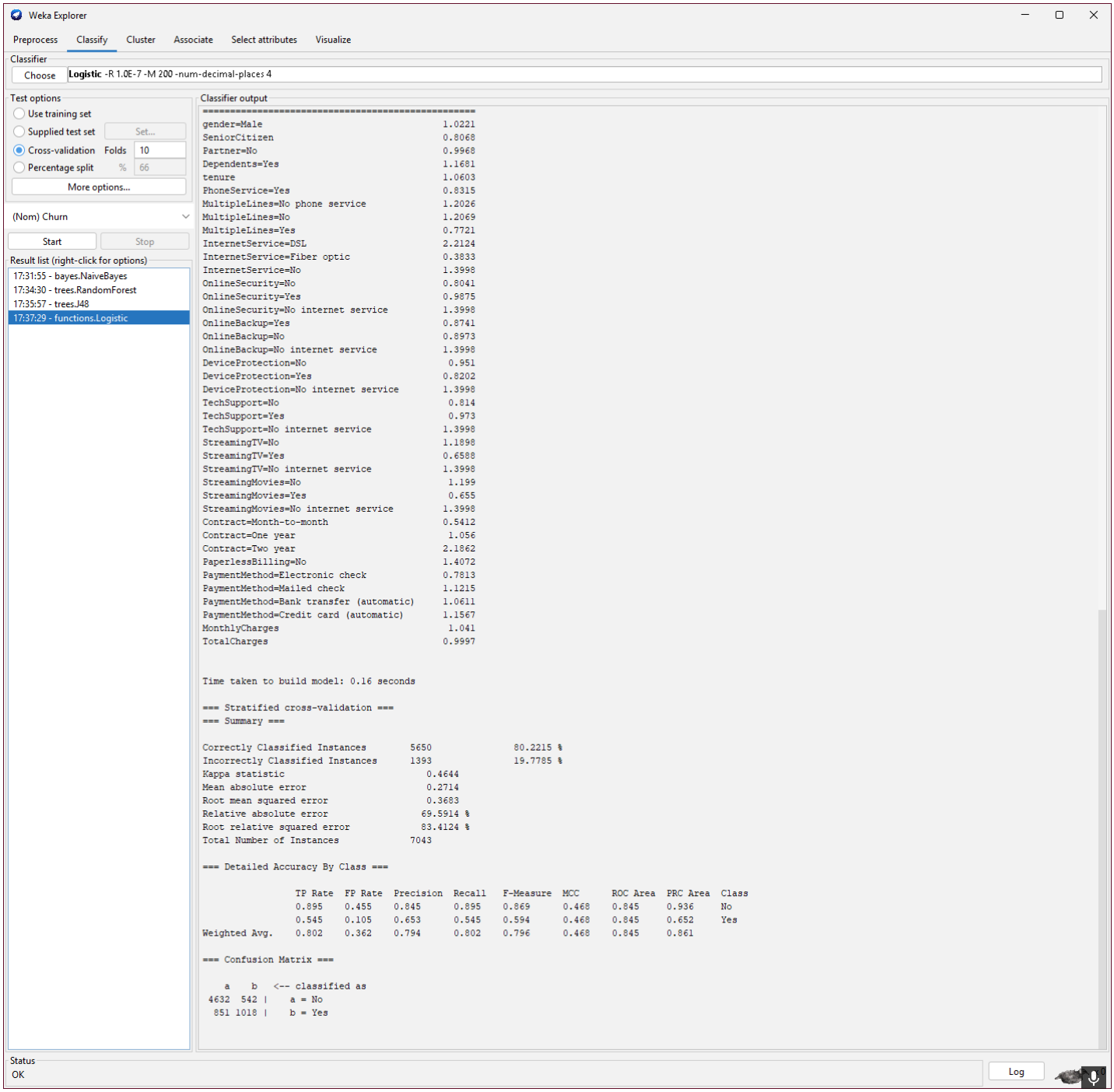
J48



RandomForest



LogisticRegression


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-7 -M 200 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Churn
Start
Stop
Result list (right-click for options)
17:31:55 - bayes.NaiveBayes
17:34:30 - trees.RandomForest
17:35:57 - trees.J48
17:37:29 - functions.Logistic
Classifier output
==========
gender=Male 1.0221
SeniorCitizen 0.8068
Partner=No 0.9968
Dependents=Yes 1.1681
tenure 1.0603
PhoneService=Yes 0.8315
MultipleLines=No phone service 1.2026
MultipleLines=No 1.2069
MultipleLines=Yes 0.7721
InternetService=DSL 2.2124
InternetService=Fiber optic 0.3833
InternetService=No 1.3998
OnlineSecurity=No 0.8041
OnlineSecurity=Yes 0.9875
OnlineSecurity=No internet service 1.3998
OnlineBackup=Yes 0.8741
OnlineBackup=No 0.8973
OnlineBackup=No internet service 1.3998
DeviceProtection=No 0.951
DeviceProtection=Yes 0.8202
DeviceProtection=No internet service 1.3998
TechSupport=No 0.814
TechSupport=Yes 0.973
TechSupport=No internet service 1.3998
StreamingTV=No 1.1898
StreamingTV=Yes 0.6588
StreamingTV=No internet service 1.3998
StreamingMovies=No 1.199
StreamingMovies=Yes 0.655
StreamingMovies=No internet service 1.3998
Contract=Month-to-month 0.5412
Contract=One year 1.056
Contract=Two year 2.1862
PaperlessBilling=No 1.4072
PaymentMethod=Electronic check 0.7813
PaymentMethod=Mailed check 1.1215
PaymentMethod=Bank transfer (automatic) 1.0611
PaymentMethod=Credit card (automatic) 1.1567
MonthlyCharges 1.041
TotalCharges 0.9997
Time taken to build model: 0.16 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 5650 80.2215 %
Incorrectly Classified Instances 1393 19.7785 %
Kappa statistic 0.4644
Mean absolute error 0.2714
Root mean squared error 0.3683
Relative absolute error 69.5914 %
Root relative squared error 83.4124 %
Total Number of Instances 7043
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.895 0.455 0.845 0.895 0.869 0.468 0.845 0.936 No
0.545 0.105 0.653 0.545 0.594 0.468 0.845 0.652 Yes
Weighted Avg. 0.802 0.362 0.794 0.802 0.796 0.468 0.845 0.861
=== Confusion Matrix ===
a b <-- classified as
4632 542 | a = No
851 1018 | b = Yes
Status
OK
Log