



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

DATA WAREHOUSE & DATA MINING (ACCA6026)

Muhammad Kamrul Hasan Sohel

MODULE CODE: ACCA6026

STUDENT ID: 2316411

LECTURER NAME: RAJEEV NATH

UNIVERSITY OF WALES TRINITY SAINT DAVID

Abstract

This report applies data-mining methods to an online retail customer dataset to predict churn and identify behavioural segments. The dataset contains 1,000 customers and 15 demographic, transactional and engagement attributes. Excel was used for descriptive exploration and business-facing charts, while the principal mining was performed in WEKA using J48, Random Forest, Logistic Regression and JRip with stratified 10-fold cross-validation. JRip was the strongest classifier, achieving 52.5% accuracy, 68.1% churn recall, an F1-score of 0.601 and ROC AUC of 0.527. These results remain close to chance, supported by near-zero Kappa values and very low Gain Ratio scores. SimpleKMeans produced three customer profiles, but classes-to-clusters evaluation showed weak correspondence with churn. The study therefore recommends improved behavioural and time-series data, cautious use of model scores, and controlled retention experiments rather than automated churn decisions.

Table of Contents

Abstract	1
Introduction	3
1.1 Purpose and Practice of Data Mining	3
1.2 Dataset Description and Exploration	3
<i>Table 1. Dataset attributes, types and analytical roles</i>	4
<i>Table 2. Selected descriptive statistics from the Excel exploration</i>	5
<i>Figure 2. Numerical correlations with churn; all relationships are close to zero</i>	6
<i>Figure 4. WEKA class distribution: 526 TRUE and 474 FALSE churn labels</i>	7
<i>Figure 5. Tableau Public dashboard for exploratory churn analysis across returns, total spend, annual income and purchase regency</i>	8
1.3 Data Pre-processing and Post-processing	8
<i>Figure 6. WEKA Gain Ratio attribute ranking; discriminative scores are near zero</i>	9
1.4 Sample Runs and Analysis Methods	10
<i>Table 3. WEKA classification performance using stratified 10-fold cross-validation</i>	10
<i>Figure 7. WEKA J48 output: 49.8% accuracy and ROC AUC 0.492</i>	11
<i>Figure 9. WEKA Random Forest output: 49.8% accuracy and 60.6% churn recall</i>	12
<i>Figure 11. WEKA JRip output: best overall result with 52.5% accuracy and 68.1% churn recall</i>	13
<i>Figure 12. Comparison of accuracy, churn recall, F1 and ROC AUC across classifiers</i> ..	13
Behavioural segmentation	13
<i>Table 4. K-means behavioural segment profiles and observed churn rates</i>	14
<i>Figure 14. K-means cluster assignments by Total_Spend and Last_Purchase_Days_Ago</i> ..	15
1.5 Business Recommendations	16
1.6 Reflective Commentary	17
Conclusion	17
References	18
Appendix A – Additional Analytical Evidence	19

Introduction

Customer churn is commercially important because a retailer that can recognise disengagement early can prioritise retention offers, service recovery and personalised communication. Modern data mining combines statistical learning, machine learning and visual analytics to convert historical customer records into descriptive and predictive evidence [1]. Recent churn research also emphasises the value of comparing multiple models and considering metrics beyond accuracy because the business cost of missing a churning customer can differ from the cost of contacting a retained customer [2], [3]. This report follows the assignment structure and uses WEKA as the main mining workbench, supported by Excel-based exploration and dashboard-oriented visual presentation.

1.1 Purpose and Practice of Data Mining

Data mining is the disciplined process of discovering useful patterns, relationships and predictive rules from data. It extends conventional reporting by moving from questions such as “what happened?” to “which customers behave similarly?” and “which customers are likely to churn?”. A robust workflow includes data understanding, preparation, modelling, evaluation and interpretation; the latest edition of Witten et al. highlights preprocessing and credibility of evaluation as core parts of practical machine learning [1]. WEKA supports these activities through preprocessing, classification, clustering, attribute selection and visualisation components [4], [5].

In retail, descriptive mining can reveal spend distributions, purchasing frequency, recency, returns and service-contact patterns. Supervised classification learns from a labelled outcome such as Target_Churn and can rank customers for retention activity. Unsupervised clustering has a different purpose: it forms customer groups without using the churn label, allowing managers to recognise profiles such as high-value, inactive or dissatisfied customers. These two methods complement each other. A classifier supports individual risk prioritisation, whereas segmentation supports differentiated campaign design. However, predictive outputs are decision support rather than certainty. Accuracy must be considered alongside precision, recall, F1, ROC AUC and agreement statistics, because a model can appear acceptable while still missing commercially important churners [2], [3].

1.2 Dataset Description and Exploration

The supplied CSV contains 1,000 instances and 15 attributes. Table 1 summarises each attribute, its data type and analytical role, while Figures 3 and 4 provide WEKA evidence of the imported schema and the near-balanced churn class distribution. Customer_ID is a unique numeric identifier; it was removed from modelling because it does not represent customer behaviour. Target_Churn is the nominal Boolean class, with 526 churned customers (52.6%) and 474 retained customers (47.4%). Gender and Promotion_Response are nominal; Email_Opt_In and Target_Churn are Boolean/nominal; the remaining predictors are numeric. No missing values or duplicate rows were detected. The class split is sufficiently balanced that synthetic oversampling was not justified.

Attribute	Type	Role
Customer_ID	Numeric	Identifier – removed
Age	Numeric	Predictor
Gender	Nominal	Predictor
Annual_Income	Numeric	Predictor
Total_Spend	Numeric	Predictor
Years_as_Customer	Numeric	Predictor
Num_of_Purchases	Numeric	Predictor
Average_Transaction_Amount	Numeric	Predictor
Num_of>Returns	Numeric	Predictor
Num_of_Support_Contacts	Numeric	Predictor
Satisfaction_Score	Numeric (1–5)	Predictor
Last_Purchase_Days_Ago	Numeric	Predictor
Email_Opt_In	Boolean/Nominal	Predictor
Promotion_Response	Nominal	Predictor
Target_Churn	Boolean/Nominal	Class

Table 1. Dataset attributes, types and analytical roles

Attribute	Min	Mean	Median	Max	Std. dev.
Age	18.00	43.27	43.00	69.00	15.24
Annual_Income	20.01	111.96	114.14	199.73	52.84
Total_Spend	108.94	5080.79	4986.19	9999.64	2862.12
Num_of_Purchases	1.00	49.46	49.00	99.00	28.54
Average_Transaction_Amount	10.46	266.88	270.10	499.57	145.87
Satisfaction_Score	1.00	2.97	3.00	5.00	1.39
Last_Purchase_Days_Ago	1.00	182.89	180.50	364.00	104.39

Table 2. Selected descriptive statistics from the Excel exploration

As summarised in Table 2, exploration showed broad variation in spending and recency: mean Total_Spend was 5,080.79 and mean Last_Purchase_Days_Ago was 182.89. Satisfaction ranged from 1 to 5 with a mean of 2.97. Churn rates were very similar across promotion-response groups (51.5%–53.6%), and all numerical correlations with churn were weak; the largest absolute Pearson correlation was only 0.036 for Annual_Income. This early evidence suggested that the supplied predictors might not contain a strong churn signal. Figure 1 brings the Excel class-balance, promotion-response and model-comparison evidence together in one dashboard, while Figure 2 shows that all numerical correlations with Target_Churn remain close to zero.

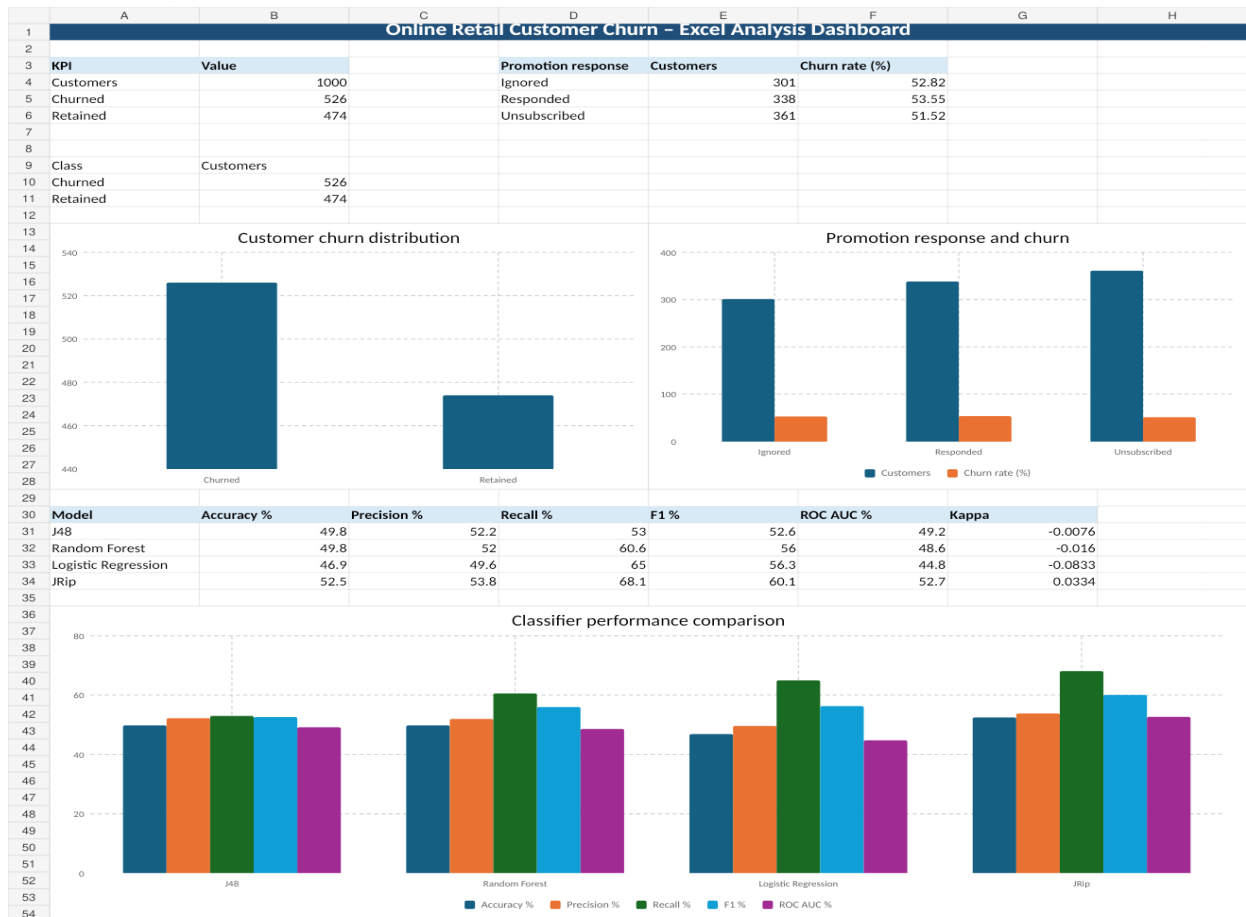


Figure 1. Excel analysis dashboard summarising class balance, promotion response and model results

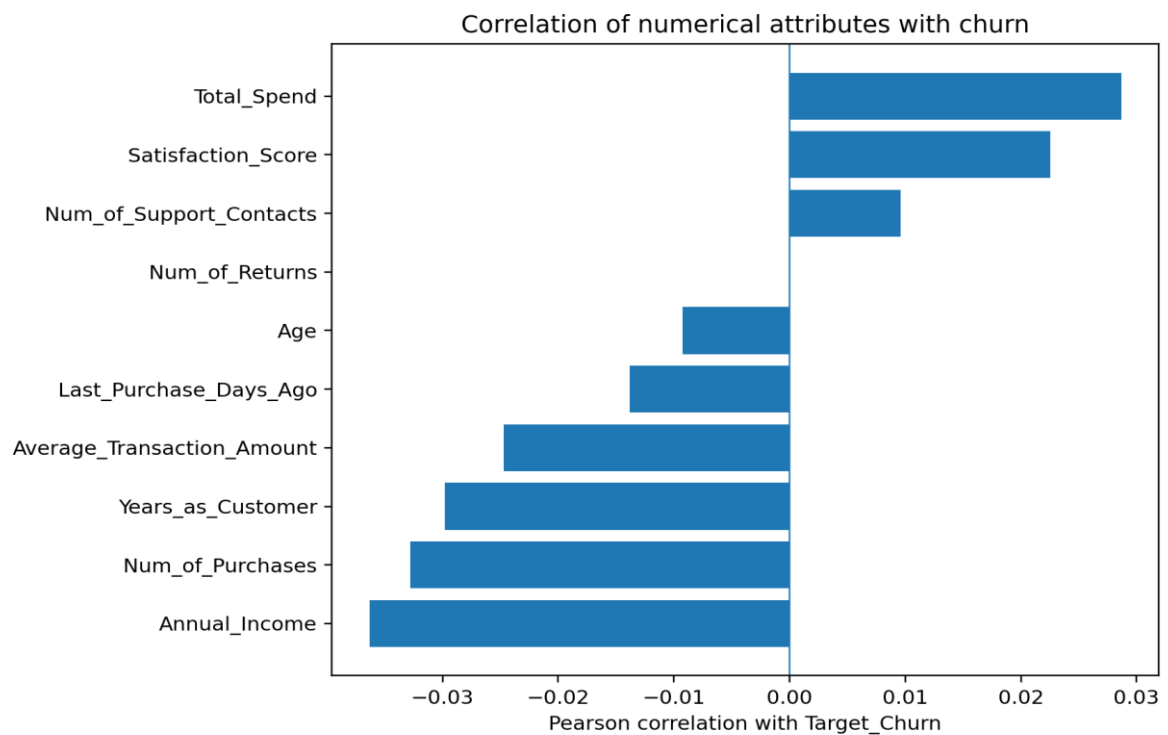


Figure 2. Numerical correlations with churn; all relationships are close to zero

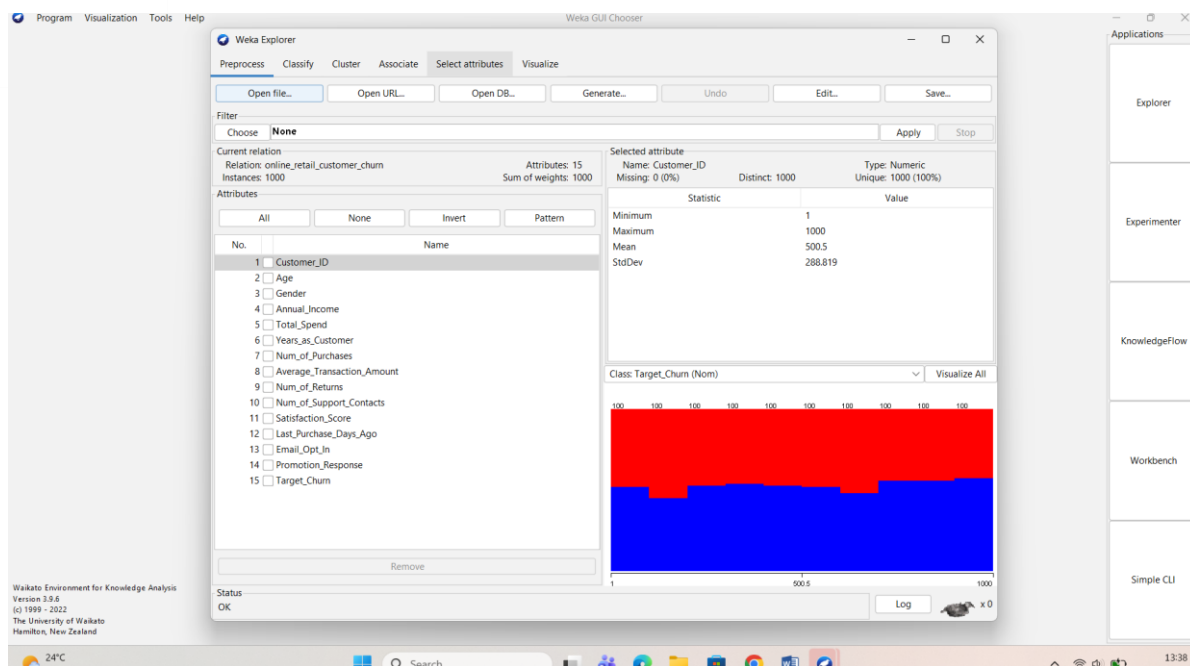


Figure 3. WEKA Preprocess view confirming 1,000 instances and 15 original attributes

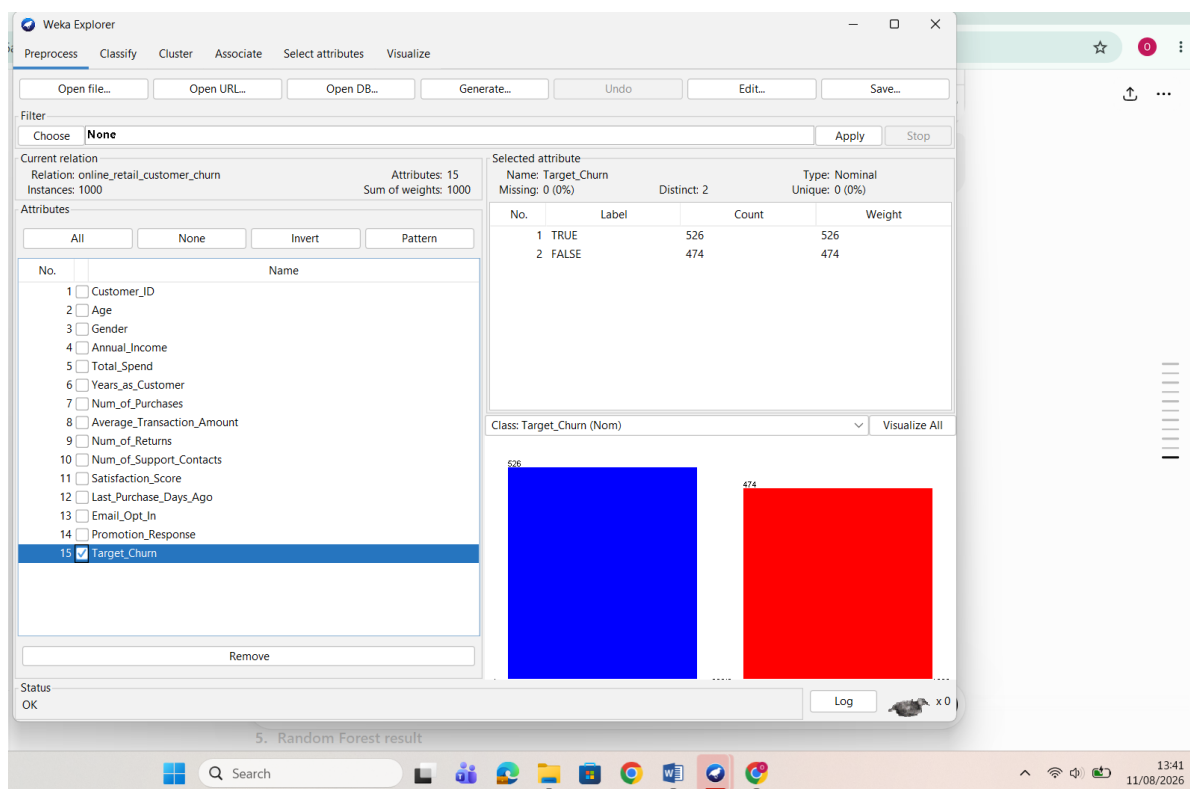


Figure 4. WEKA class distribution: 526 TRUE and 474 FALSE churn labels

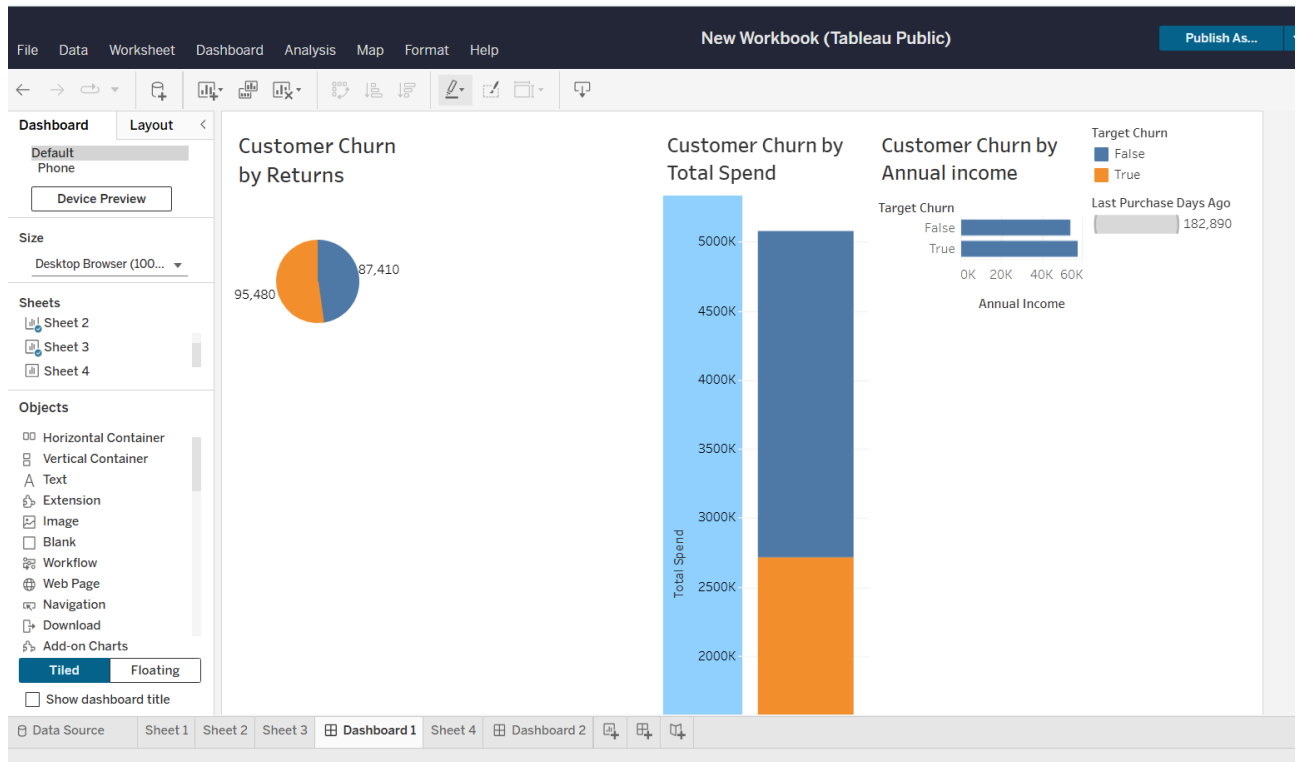


Figure 5. Tableau Public dashboard for exploratory churn analysis across returns, total spend, annual income and purchase regency

A supplementary dashboard was created in Tableau Public from the same CSV to strengthen business-facing visual exploration, as shown in Figure 5. The dashboard combines churn status with returns, total spend, annual income and days since the last purchase, allowing managers to compare customer behaviour interactively. This Tableau stage was used for descriptive presentation rather than predictive modelling; WEKA remained the principal data-mining environment, consistent with the assignment requirement. The dashboard complements the Excel summaries by presenting several churn-related measures in a single management view [6].

1.3 Data Pre-processing and Post-processing

Pre-processing began with schema inspection, missing-value checks, duplicate checks and range validation. Because no missing values were present, imputation was unnecessary. Customer_ID was removed before modelling; leaving a unique identifier in the feature set can allow accidental record-specific splits without business meaning. Target_Churn was set as the class attribute. WEKA can import the CSV directly and supports filters, classifiers, clusterers and attribute-selection processes [4].

Categorical variables were retained as nominal values for tree/rule models. J48, Random Forest and JRip can directly split or construct rules using nominal attributes. Logistic Regression requires an internally suitable numeric representation; WEKA handles nominal-to-binary processing within its modelling workflow. Scale differences were considered for clustering

because `Annual_Income` and `Total_Spend` have much larger numeric ranges than `Satisfaction_Score`. In a production workflow, `Standardize` or `Normalize` should be applied within the modelling pipeline so that transformations are learned only from training data [1]. The supplied WEKA clustering run was interpreted conservatively because K-means with Euclidean distance is sensitive to scale.

Class balancing was assessed but not applied. The 52.6:47.4 distribution is close to balanced, so SMOTE or oversampling would add synthetic information without addressing the central problem: weak relationships between predictors and `Target_Churn`. Stratified 10-fold cross-validation was used for classification so each record was tested once while class proportions were approximately preserved across folds. This is preferable to reporting a single arbitrary percentage split. WEKA's documentation identifies cross-validation as a standard classifier evaluation mode and supports confusion matrices and per-class statistics [4].

Figure 6 shows the `GainRatioAttributeEval` with Ranker output used for feature selection. Gender ranked first (0.00057228), followed by `Promotion_Response` (0.00013516) and `Email_Opt_In` (0.000000751); the remaining attributes received zero in this run. These values are exceptionally small, providing independent evidence that no single feature has strong discriminatory power. Post-processing therefore focused on uncertainty rather than forcing a positive story: confusion matrices were translated into business-relevant false-positive/false-negative implications, model metrics were compared in one table, and K-means centroids were converted into descriptive customer profiles. Excel charts were used for descriptive communication, while a Tableau-oriented dashboard design is appropriate for combining multiple views into one management screen; Tableau defines a dashboard as a collection of connected views for simultaneous comparison [6].

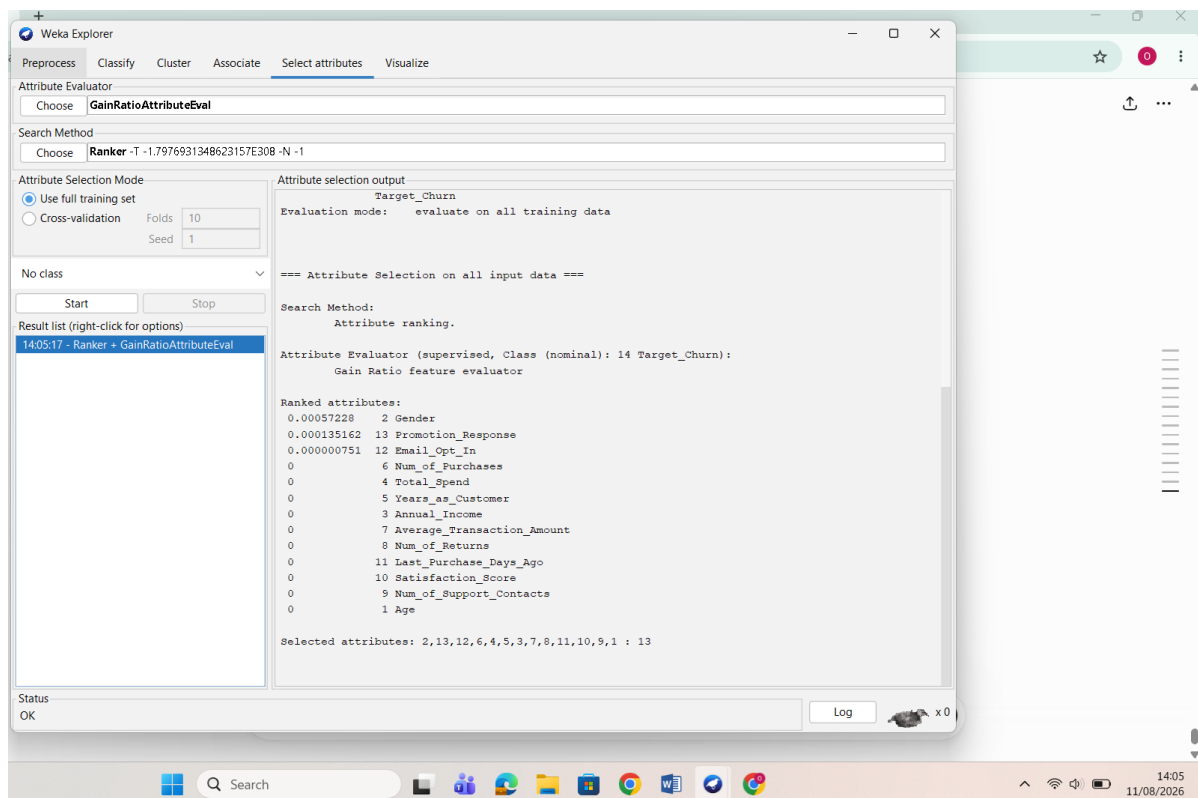


Figure 6. WEKA Gain Ratio attribute ranking; discriminative scores are near zero

1.4 Sample Runs and Analysis Methods

Four classifiers were compared under the same 10-fold cross-validation design, with the consolidated results reported in Table 3 and the corresponding WEKA outputs presented in Figures 7–11. J48 is WEKA's implementation of the C4.5 decision-tree family and offers transparent hierarchical rules [7]. Random Forest combines many randomised trees to reduce the variance of a single tree [8]. Logistic Regression provides a linear probabilistic benchmark and is useful when managers want a relatively interpretable model. JRip learns a compact set of propositional rules, making it attractive when understandable IF–THEN logic is required. Recent churn studies similarly compare multiple algorithms and emphasise recall, F1 and ROC-based evaluation rather than accuracy alone [2], [3].

Model	Accuracy	Churn precision	Churn recall	Churn F1	ROC AUC	Kappa
J48	49.8%	52.2%	53.0%	0.526	0.492	-0.0076
Random Forest	49.8%	52.0%	60.6%	0.560	0.486	-0.0160
Logistic Regression	46.9%	49.6%	65.0%	0.563	0.448	-0.0833
JRip	52.5%	53.8%	68.1%	0.601	0.527	0.0334

Table 3. WEKA classification performance using stratified 10-fold cross-validation

Table 3 shows that JRip was the strongest overall model, correctly classifying 525 of 1,000 cases (52.5%); this is also visible in the WEKA output in Figure 11. Figure 12 then compares the principal performance measures across all four classifiers. For churners it achieved precision 0.538, recall 0.681 and F1 0.601, with ROC AUC 0.527 and Kappa 0.0334. Its confusion matrix contained 358 true positives, 168 false negatives, 307 false positives and 167 true negatives. The high churn recall is useful if the objective is to identify as many at-risk customers as possible, but 307 false positives would create substantial unnecessary retention contact. The ROC AUC only slightly exceeds 0.5, so the ranking ability remains weak.

J48 and Random Forest both achieved 49.8% accuracy. Random Forest improved churn recall to 60.6%, whereas J48 achieved 53.0%. Logistic Regression produced the lowest accuracy (46.9%) but relatively high churn recall (65.0%), demonstrating why accuracy alone is insufficient. Kappa was negative for J48, Random Forest and Logistic Regression, while JRip's Kappa was only 0.0334. In practical terms, none of the models demonstrates dependable separation of churners from retained customers. This contrasts with recent research that reports high performance on richer churn datasets, suggesting that model quality depends strongly on informative variables and target construction rather than algorithm choice alone [2], [3].

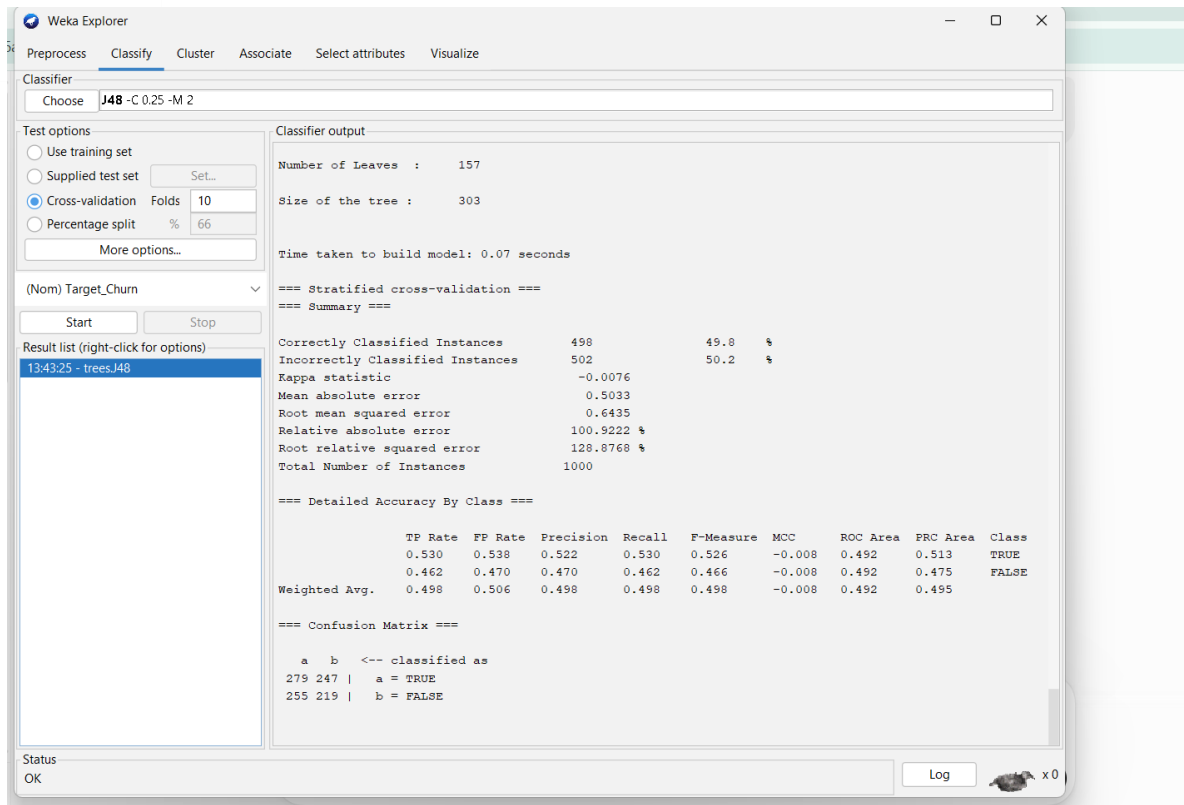


Figure 7. WEKA J48 output: 49.8% accuracy and ROC AUC 0.492

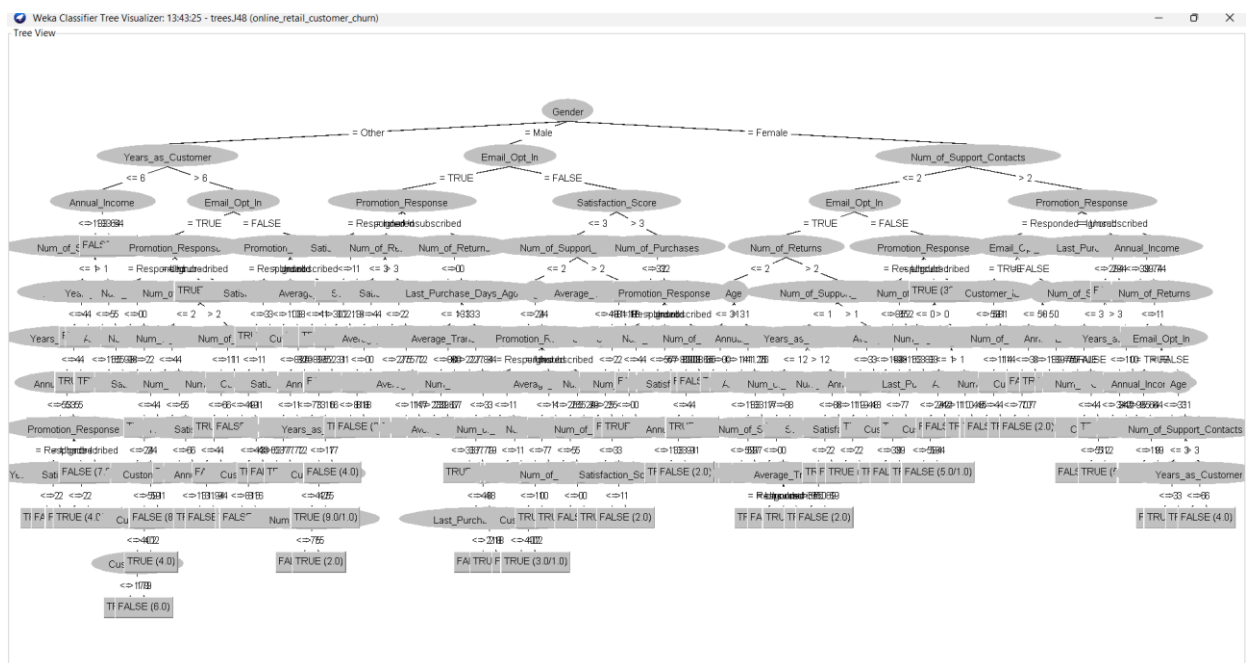


Figure 8. J48 decision tree; the very large tree illustrates complexity without strong predictive gain

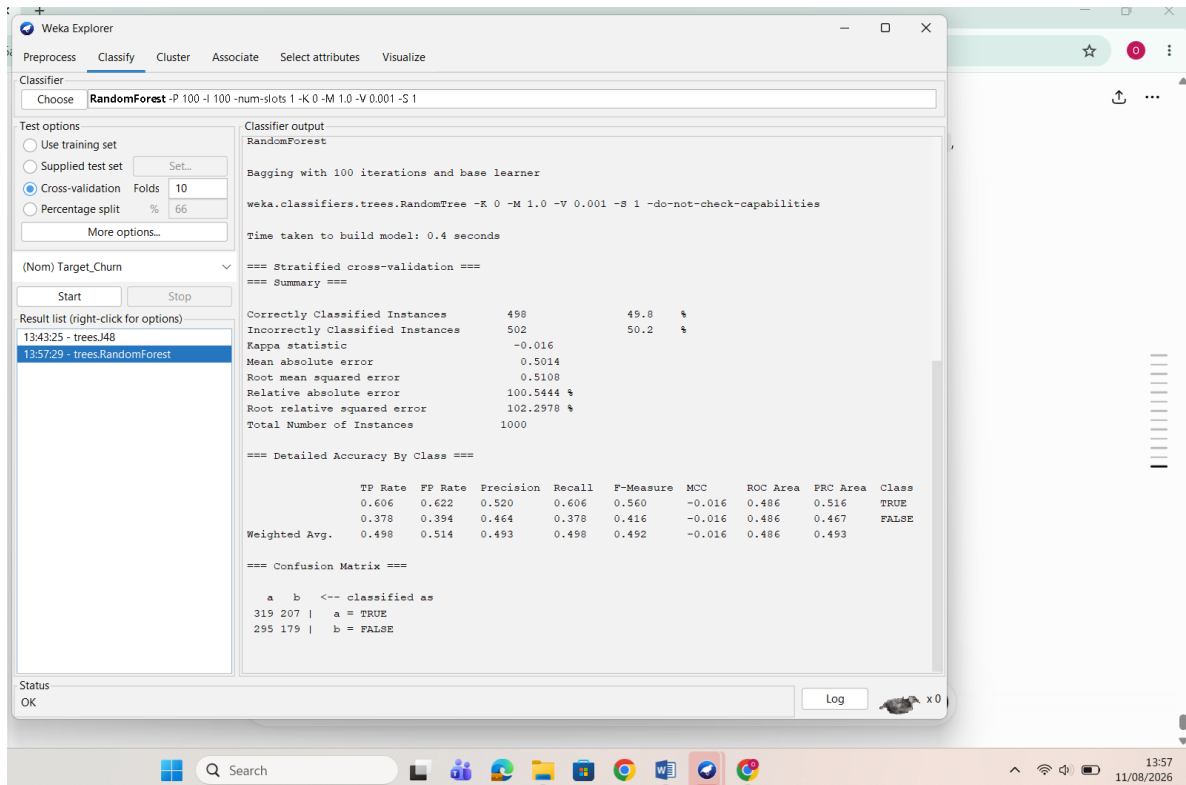


Figure 9. WEKA Random Forest output: 49.8% accuracy and 60.6% churn recall

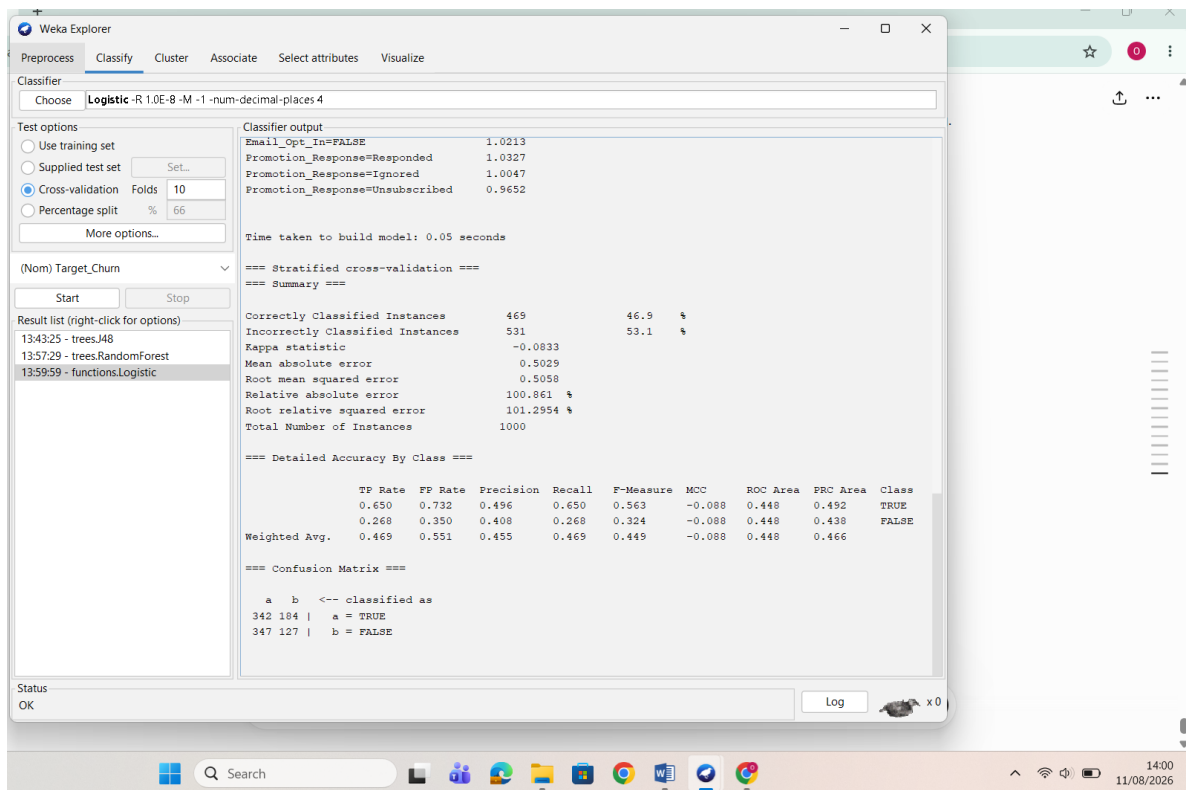


Figure 10. WEKA Logistic Regression output: 46.9% accuracy and 65.0% churn recall

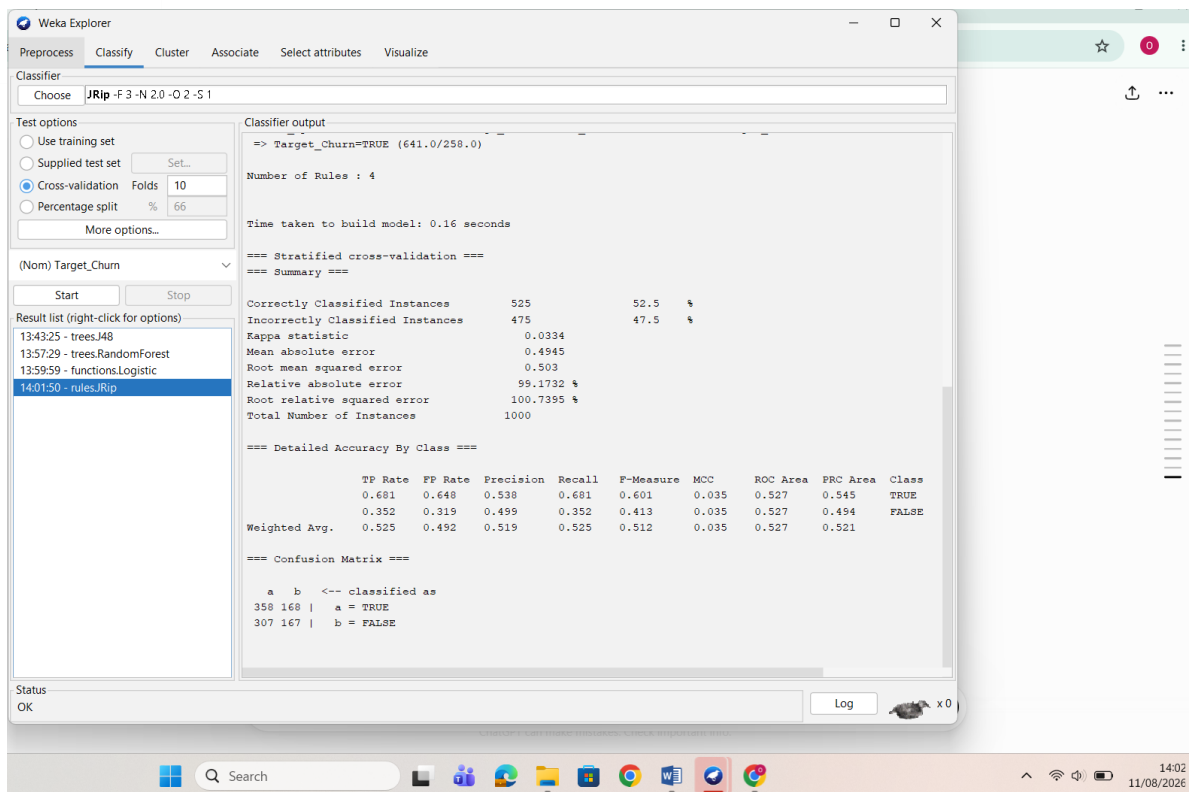


Figure 11. WEKA JRip output: best overall result with 52.5% accuracy and 68.1% churn recall

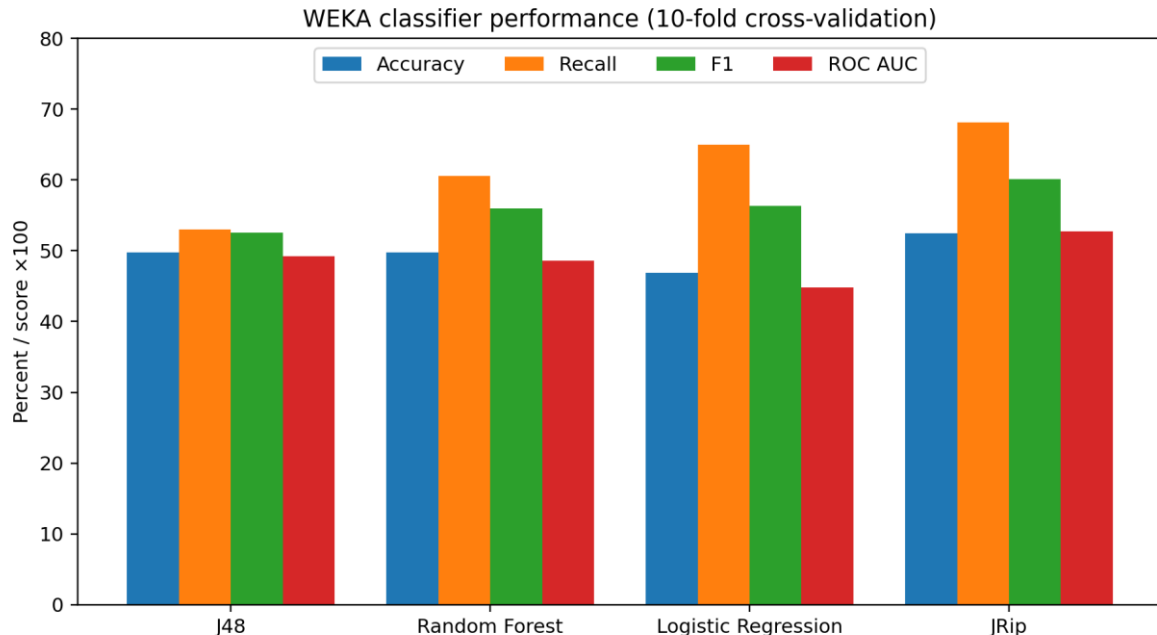


Figure 12. Comparison of accuracy, churn recall, F1 and ROC AUC across classifiers

Behavioural segmentation

SimpleKMeans was run with three clusters after excluding Customer_ID and Target_Churn from cluster construction. The resulting centroid profiles are shown in Figure 13 and summarised numerically in Table 4. This separation is methodologically important: the churn

label should evaluate the segments, not create them. Cluster 0 contained 267 customers and had the highest mean income (128.28) and spend (5,593.10), with an older mean age of 49.7. Cluster 1 was largest (391) and combined the highest purchase frequency (51.85), average transaction amount (306.15) and satisfaction (3.11), but also the longest inactivity (188.16 days). Cluster 2 contained 342 younger, lower-income customers, with the lowest satisfaction (2.80) and more recent purchasing (173.66 days).

Cluster	Size	Age	Income	Spend	Purchases	Satisfaction	Days inactive	Churn rate
Cluster 0	267 (27%)	49.7	128.28	5,593.10	49.35	3.00	186.99	50.6%
Cluster 1	391 (39%)	41.7	113.71	4,829.95	51.85	3.11	188.16	51.7%
Cluster 2	342 (34%)	40.0	97.23	4,967.61	46.80	2.80	173.66	55.3%

Table 4. K-means behavioural segment profiles and observed churn rates

Classes-to-clusters evaluation produced churn rates of 50.6%, 51.7% and 55.3% across the three groups, with WEKA reporting 62.2% incorrectly clustered instances when mapping clusters to the binary class. Figure 14, which plots Total_Spend against Last_Purchase_Days_Ago, shows substantial colour overlap, while Figure 15 confirms the weak classes-to-clusters alignment. Therefore, the clusters are useful as descriptive campaign profiles but not as clean proxies for churn status. K-means is valuable for segmentation, but cluster usefulness should be judged by business interpretability and validation rather than assuming that every cluster is a naturally distinct churn group [9].

Weka Explorer

Preprocess Classify **Cluster** Associate Select attributes Visualize

Clusterer: Choose **SimpleKMeans** -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

☒ Use training set

☐ Supplied test set Set...

☐ Percentage split % 66

☐ Classes to clusters evaluation (Nom) Target_Churn

☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

14:09:08 - SimpleKMeans

14:12:07 - SimpleKMeans

Cluster output

Cluster 2: 22,Other,28.87,3982.58,10,20,251.79,2,1,1,323,FALSE,Responded

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Full Data (1000.0)	Cluster# 0 (267.0)	1 (391.0)	2 (342.0)
Age	43.267	49.6742	41.7263	40.0263
Gender	Female	Female	Female	Other
Annual_Income	111.963	128.282	113.7075	97.2282
Total_Spend	5080.7927	5593.1007	4829.95	4967.6149
Years_as_Customer	9.727	9.0861	10.89	8.8977
Num_of_Purchases	49.456	49.3483	51.8542	46.7982
Average_Transaction_Amount	266.8765	253.3253	306.1458	232.5604
Num_of>Returns	4.612	4.8539	4.7059	4.3158
Num_of_Support_Contacts	1.934	1.8652	2.0384	1.8684
Satisfaction_Score	2.974	3	3.1074	2.8012
Last_Purchase_Days_Ago	182.89	186.9888	188.1611	173.6637
Email_Opt_In	TRUE	TRUE	TRUE	TRUE
Promotion_Response	Unsubscribed	Responded	Unsubscribed	Responded

Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	267 (27%)
1	391 (39%)
2	342 (34%)

Status
OK

Figure 13. WEKA SimpleKMeans centroids after excluding Target_Churn from cluster construction

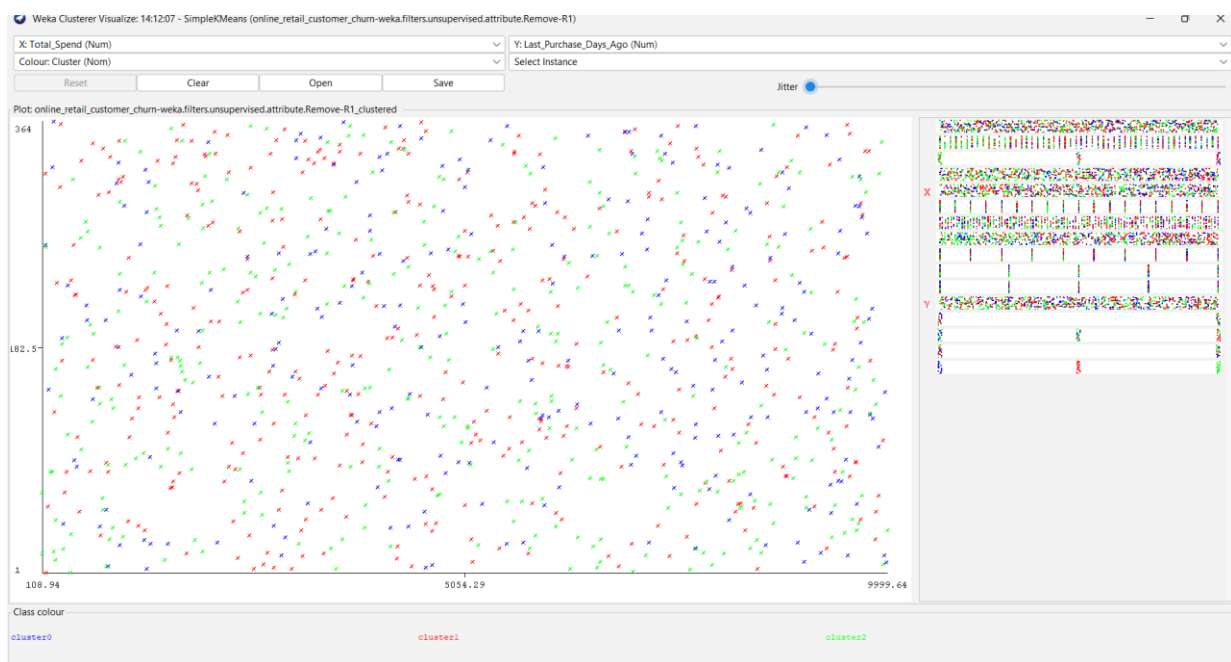


Figure 14. K-means cluster assignments by Total_Spend and Last_Purchase_Days_Ago

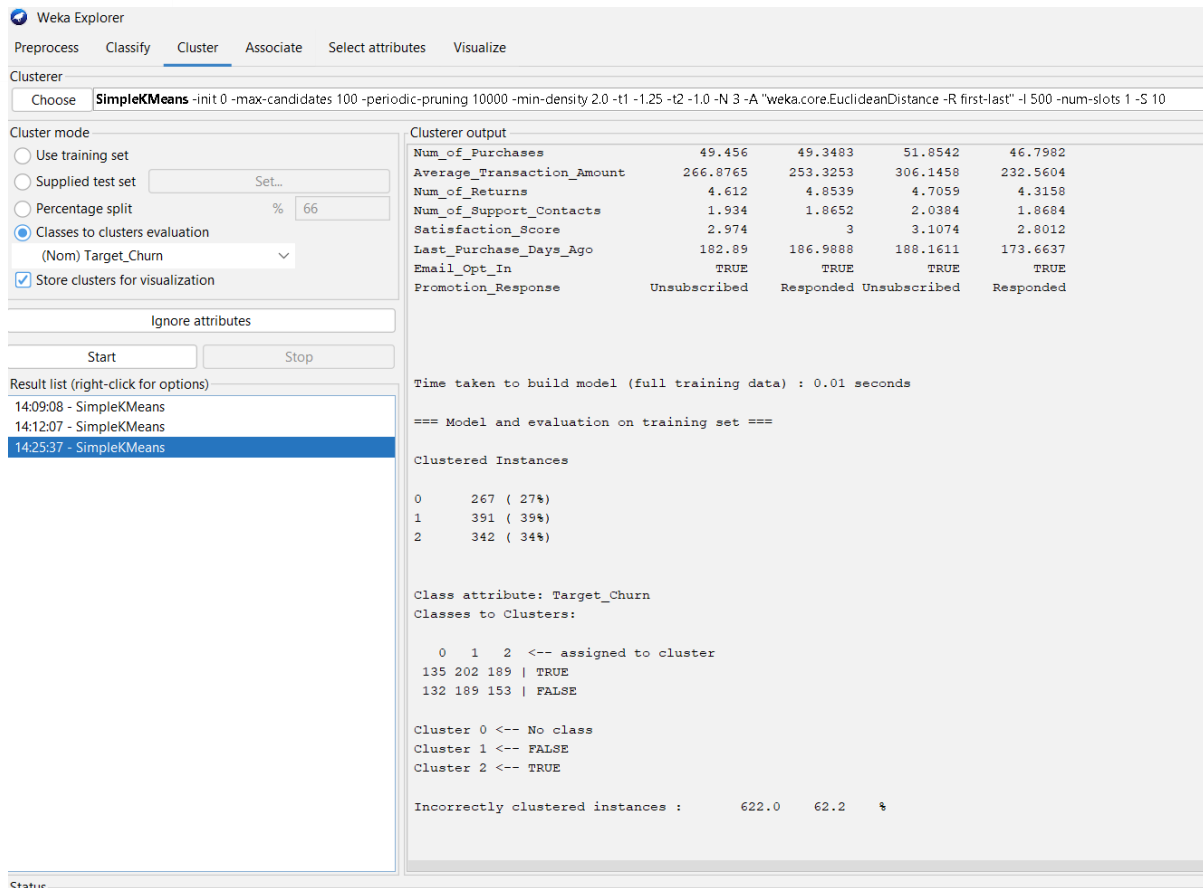


Figure 15. WEKA classes-to-clusters evaluation; the segments align weakly with churn

1.5 Business Recommendations

The current models should not be used to automatically decide which customers receive costly retention incentives. JRip is the best tested model, but AUC 0.527 and Kappa 0.0334 indicate only marginal discrimination. Instead, the retailer should improve the feature set with time-stamped purchase frequency, declining spend trends, website/app sessions, basket abandonment, discount dependence, complaint categories, delivery delays, refund value, loyalty activity and customer lifetime value. Churn should also be operationally defined using a defensible inactivity window and validated on a later time period.

For immediate low-risk action, the three K-means profiles can guide controlled campaign tests. Cluster 2 has the highest observed churn rate and lowest satisfaction, so it is suitable for service-recovery or feedback experiments; Cluster 0 combines higher income and spend and could receive value-preserving reactivation rather than blanket discounts. Campaigns should include control groups and be judged on incremental margin, repeat purchase, cost per reactivated customer and retention lift. Excel can support routine monitoring, while a Tableau dashboard can combine risk, segment and campaign views for managers [10], [6].

1.6 Reflective Commentary

The analysis showed that a technically correct data-mining workflow does not guarantee a commercially useful model. Strengths included a clean and balanced dataset, consistent 10-fold cross-validation, comparison of four different classifier families, use of per-class metrics, transparent confusion matrices and a separate unsupervised segmentation stage. The results were also presented through both WEKA evidence and Excel-based visual summaries.

The main limitation was not software or algorithm choice but weak predictive signal. The attribute-ranking scores were almost zero, AUC values were close to 0.5, and the clusters overlapped strongly. The dataset is also cross-sectional rather than longitudinal, so it cannot directly represent changes in behaviour before churn. If repeating the study, I would first verify how Target_Churn was generated, engineer recency-frequency-monetary trends over multiple periods, standardise variables consistently for distance-based methods, test cost-sensitive thresholds, and reserve a future-time holdout set. I would then use explainability and controlled retention experiments before recommending deployment.

Conclusion

The analysis met both objectives: customer churn was modelled using classification and customer behaviour was explored using K-means segmentation. JRip performed best but remained only slightly above chance, while J48, Random Forest and Logistic Regression were also weak. Gain Ratio ranking confirmed limited feature signal. Three interpretable behavioural clusters were identified, yet their churn rates differed only modestly. The most defensible business conclusion is therefore to improve behavioural and temporal data, use the present findings for exploratory segmentation and campaign testing, and avoid high-impact automated churn decisions until out-of-sample performance improves.

References

- [1] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, and J. Foulds, Data Mining: Practical Machine Learning Tools and Techniques, 5th ed. Cambridge, MA, USA: Morgan Kaufmann, 2025.
- [2] N. Ahmad, M. J. Awan, H. Nobanee, A. M. Zain, A. Naseem, and A. Mahmoud, “Customer Personality Analysis for Churn Prediction Using Hybrid Ensemble Models and Class Balancing Techniques,” *IEEE Access*, vol. 12, pp. 1865–1879, 2024, doi: 10.1109/ACCESS.2023.3334641.
- [3] N. M. AbdelAziz et al., “A Comprehensive Evaluation of Machine Learning and Deep Learning Models for Churn Prediction,” *Information*, vol. 16, no. 7, art. 537, 2025, doi: 10.3390/info16070537.
- [4] University of Waikato, “WEKA Wiki: Use WEKA in your Java code,” WEKA Documentation. [Online]. Available: https://waikato.github.io/weka-wiki/use_weka_in_your_java_code/. [Accessed: Aug. 05, 2026].
- [5] M. Hall et al., “The WEKA Data Mining Software: An Update,” *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.
- [6] Salesforce, “Dashboards,” Tableau Desktop and Web Authoring Help. [Online]. Available: <https://help.tableau.com/current/pro/desktop/en-gb/dashboards.htm>. [Accessed: Aug. 10, 2026].
- [7] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA, USA: Morgan Kaufmann, 1993.
- [8] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [9] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, 1967, pp. 281–297.
- [10] Microsoft, “Create a chart from start to finish,” Microsoft Support. [Online]. Available: <https://support.microsoft.com/en-us/excel/get-started/create-a-chart-from-start-to-finish>. [Accessed: Aug. 07, 2026].

Appendix A – Additional Analytical Evidence

The appendix contains supporting visual evidence that is useful for auditability but is not required to interpret the main conclusions. The appendix provides additional analytical evidence supporting the main report. Figures A1–A3 compare key behavioural distributions by churn status, while Figure A4 compares churn rates across promotion-response categories. These visuals complement the descriptive statistics, model-comparison and cluster-profile evidence already presented in the main body.

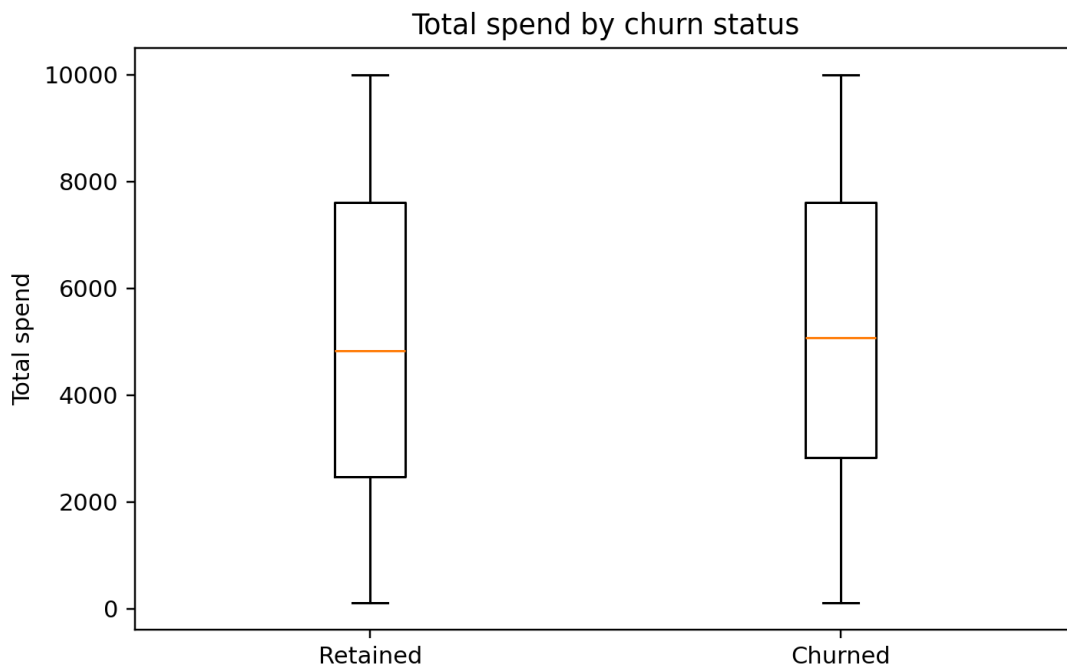


Figure A1. Total Spend distribution by churn status

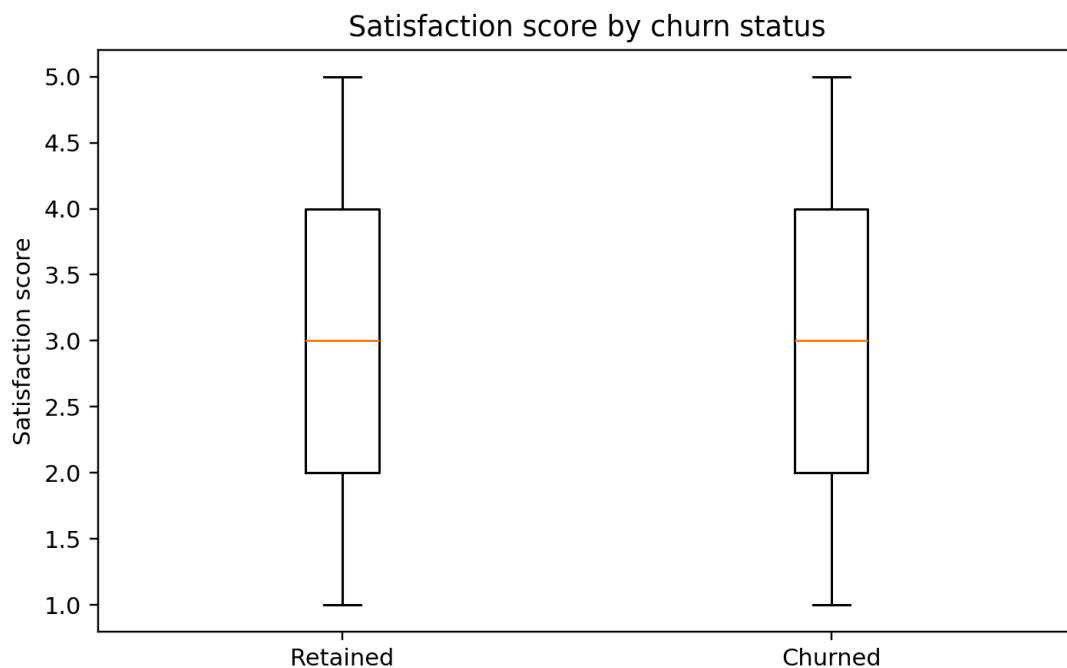


Figure A2. Satisfaction Score distribution by churn status

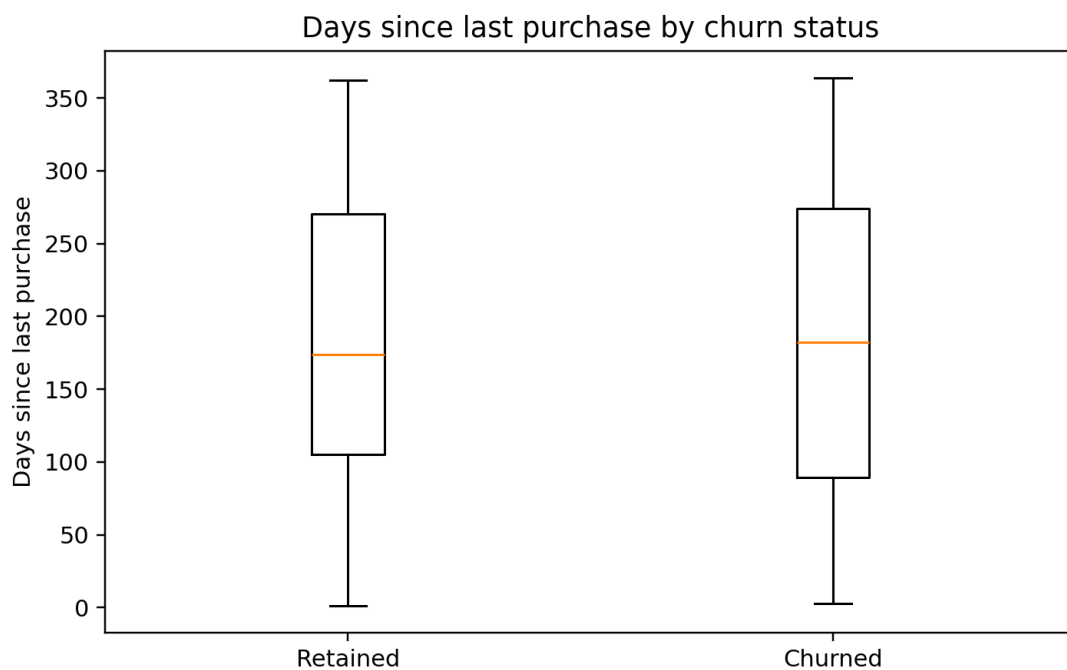


Figure A3. Days since last purchase by churn status

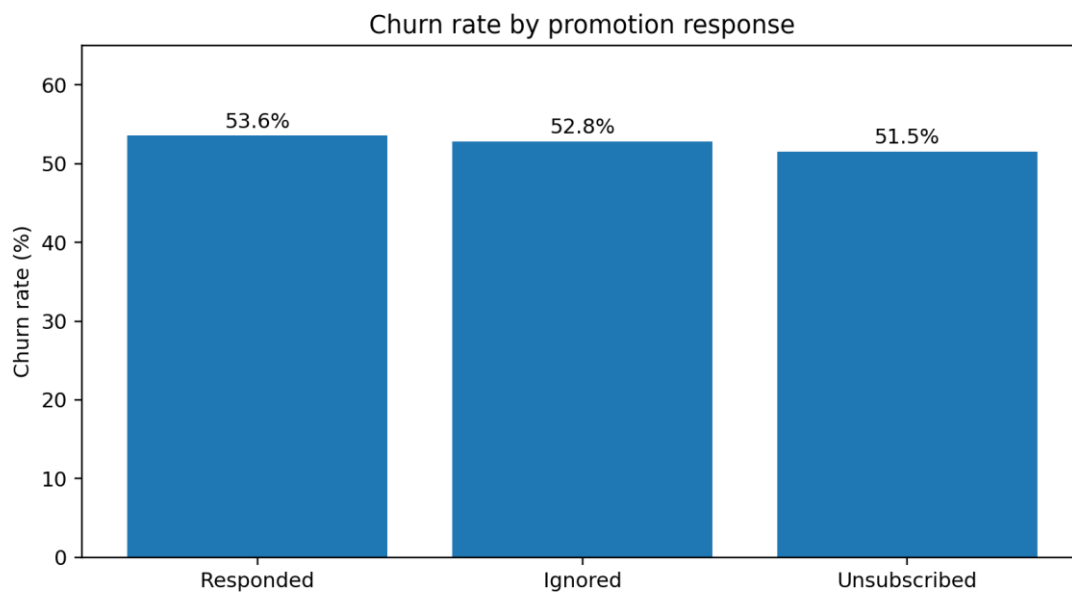


Figure A4. Churn rate by promotion-response category