**Term Project**

**Introduction**

With the age of big data, extracting valuable information from a large dataset becomes a requirement for making effective decisions in any industry. Data mining offers rigorous techniques such as classification, association, and clustering to uncover patterns and relationships hiding within data. The purpose of this project is to apply data mining techniques using WEKA software to examine an actual dataset and replicate and contrast results from a given research work. The objective is to gain better knowledge of crucial data mining techniques, classification, association, and clustering, by using them on an assigned dataset using WEKA and comparing the findings with those results presented in a research academic paper.

**Selected Option**

The selected option is Telco Customer Churn Dataset, with the following details:

* Used In: Customer Churn Prediction in Telecom Using Machine Learning
* Mining Type: Classification
* Paper Link: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-
* 019-0191-6
* Dataset Link: https://www.kaggle.com/datasets/blastchar/telco-customer-churn

The research paper is titled: Customer churn prediction in telecom using machine learning in big data platform, and below is the full summary of the paper.

**Research Paper Summary**

The research study is focused on developing a machine learning-based churn prediction model for SyriaTel, a telecommunication company, with the ultimate objective of identifying the customers who are most likely to churn so that the company may actively pursue retention efforts to retain them and improve overall revenue. The study makes extensive use of the various mobile network data features, including traditional statistical features and Social Network Analysis (SNA) features, for churn prediction.

The data is SyriaTel's prepaid customers' historical data. This data needs to be preprocessed and transformed to obtain meaningful features. Statistical features such as balance, usage patterns, and transactions are extracted on various time windows (e.g., 1-3 months, 4-6 months). SNA features such as cosine similarity, local clustering coefficients, and transaction percentages between operators are also computed. Feature engineering is necessary to improve the accuracy of churn prediction. The research explores the effectiveness of different past time windows (sliding windows) for front-end feature extraction. It is found that the most recent data performs best. Statistically derived features for the last six months showed significant improvements in model performance, while SNA features performed best when they were extracted from the last four months before the baseline.

Several classification algorithms are tried, including Decision Tree, Random Forest, GBM, and XGBoost. The most performing algorithm is the XGBoost algorithm, which achieves an AUC of 93.3% when using both statistical and SNA features. The second-best algorithm is the GBM algorithm, with Random Forest and Decision Trees performing worse. The models were trained using 10-fold cross-validation, and hyperparameter tuning was employed to optimize the models.

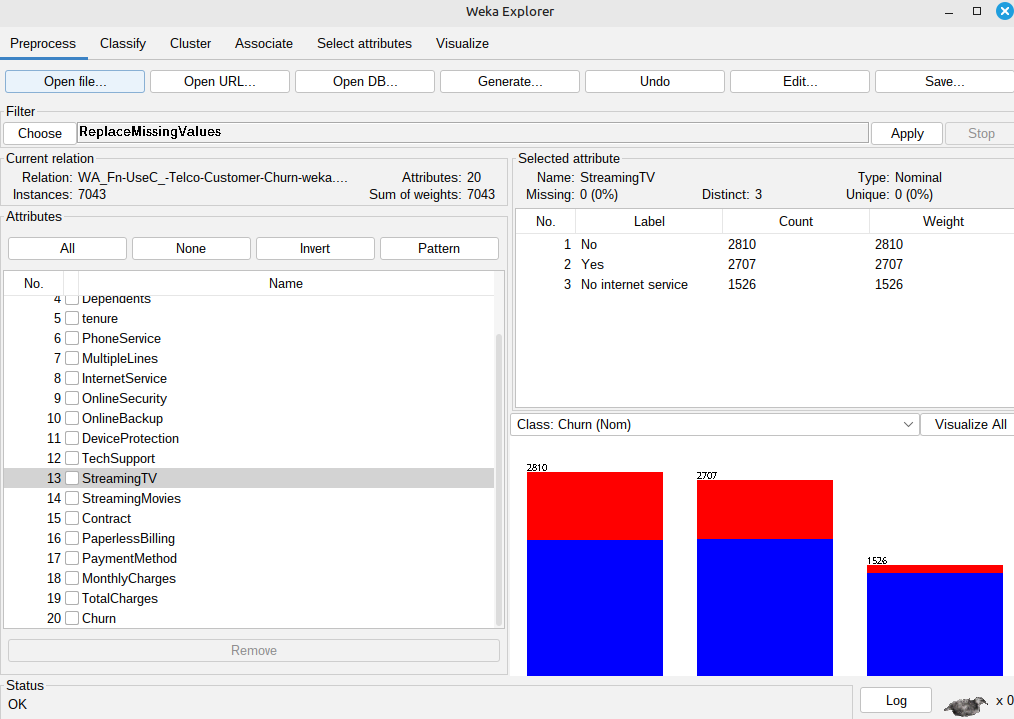
The proposed churn prediction system is deployed on new data of 7.5 million SyriaTel customers. The data is divided into two parts: the customers who are likely to churn and the customers who are not likely to churn. The marketing experts take early actions to retain the customers who are likely to churn, and a 47% retention is gained in the "Offered" dataset. This is equivalent to a reduction of 1.5% in the churn rate and a revenue increase for SyriaTel. The maximum AUC value on this test data for XGBoost is 89%, a decent score, though slightly less than the training phase.

The addition of SNA features to traditional statistical features greatly improves the performance of the churn prediction model. The model using both feature types achieved the highest AUC of 93.3%. Moreover, XGBoost outperforms other algorithms in every scenario and is selected as the optimal model for churn prediction. When the model was applied to new real-time data, it still achieved an 89% AUC, which is excellent given the non-stationary nature of real-world data. The research demonstrates the significant role of both statistical and Social Network Analysis features for customer churn prediction in the telecommunications industry.

**Dataset and Preprocessing**

The dataset that will be used in this project is also provided and can be accessed using the following link: <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>. It will be used to run experiments in weka.

The first step in this section of the project is loading the data into Weka. This was done without as issues, as evidenced in the screenshot below:

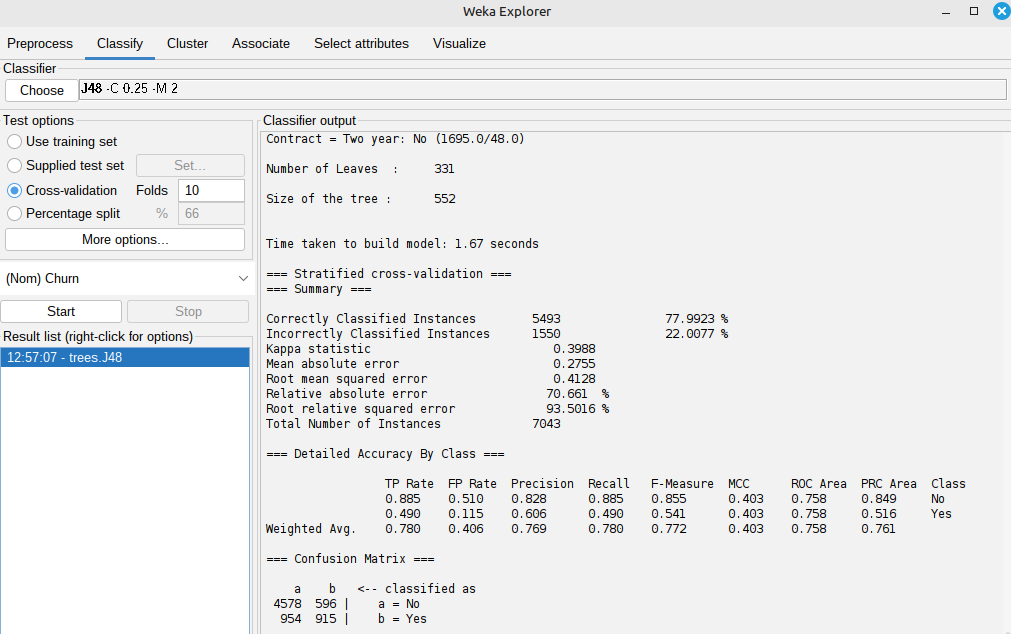


*Figure 1: Loading data into Weka*

Before building the models, the data had to be cleaned and prepared in WEKA. The customerID attribute was dropped first, since it is not predictive and could introduce noise into the models. Missing values across the dataset were then addressed by using the ReplaceMissingValues filter to ensure consistency and avoid errors during model training. If there were any columns found to be string data types, the StringToNominal filter was used to transform them into a useful nominal format for classification purposes. Having finished the cleaning process, the second step was to set the target attribute. In WEKA's Classify tab, the Churn attribute was chosen as the class label, establishing the prediction task as identifying whether a customer would churn (Yes/No).

**Decision Tree**

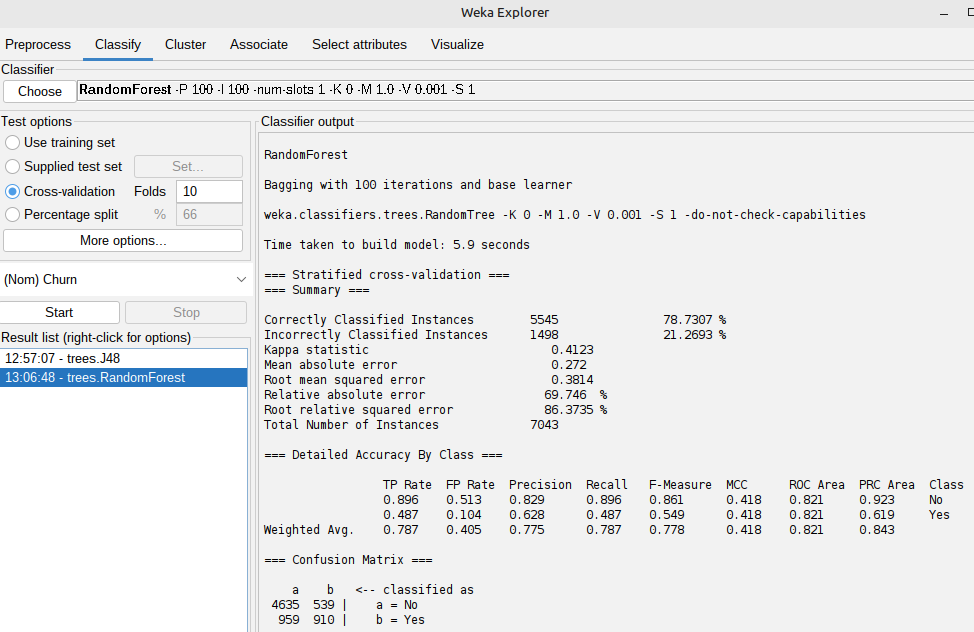
The first model to be created is the decision tree (J48). To achieve this, we go to the classification tab, choose > trees > J48, confidence factor = 0.25, and MinNumObj = 2, as shown in the screenshot below with the configurations and results.



*Figure 2: Decision tree model results*

**Random Forest**

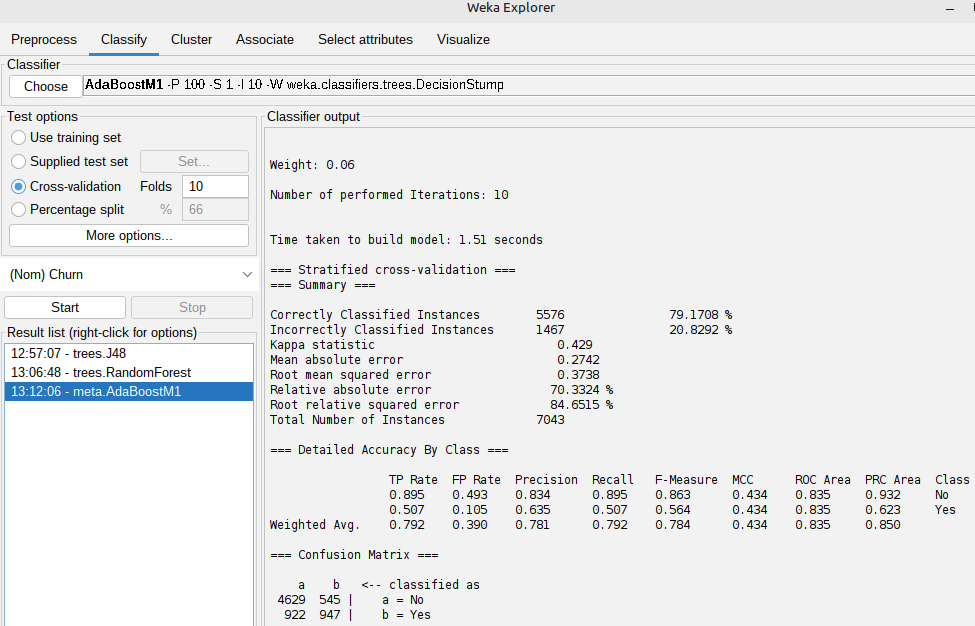
The next model is random forest. For this, we choose trees and then random forest. The following results were obtained for this model.



*Figure 3: Random forest results*

**ADABOOST**

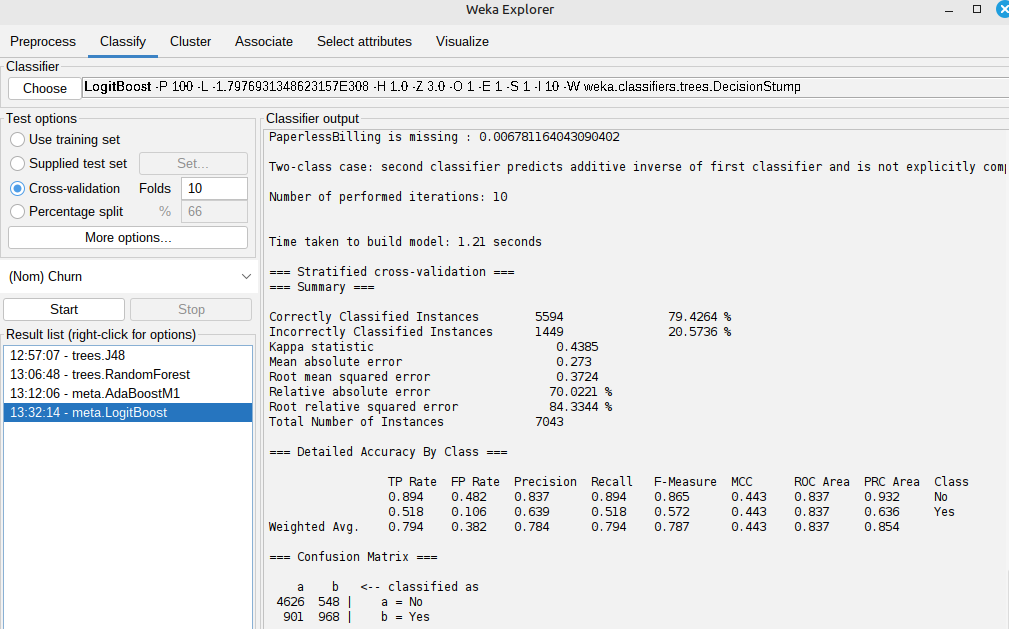
The next model is ADABOOST (Gradient Boosting). For this, we select meta then AdaBoostM1. The following results were obtained from the model:



*Figure 4: ADABOOST results*

**LogitBoost Model (Alternative to XGBoost)**

Due to the unavailability of XGBoost in WEKA, the LogitBoost classifier was used as an alternative. LogitBoost builds additive logistic regression models and also serves as a powerful boosting algorithm, making it best for classification tasks. The results are shown in the screenshot below:



*Figure 5: Logiboost results*

**Analysis and Comparison**

### This section contrasts the result of our churn prediction model using the Telco Customer Churn dataset in WEKA with the result given in the mentioned research paper, which used a massive SyriaTel telecom dataset and big data tools.

| **Model** | **Research Paper** | **Project (WEKA)** | **AUC (Paper)** | **AUC (WEKA)** | **Accuracy (WEKA)** | **Notes** |
| --- | --- | --- | --- | --- | --- | --- |
| XGBoost | Yes | No | 93.3% | — | — | Best performer in the paper using both statistical and SNA features. |
| LogitBoost | No | Yes | — | 83.7% | 79.4% | Used as an alternative to XGBoost; performed best in WEKA. |
| Gradient Boosting | Yes (GBM) | Yes (AdaBoostM1) | ~85.5% | ~83.5% | 79.2% | Good performance in both studies. |
| Random Forest | Yes | Yes | 83.4% | ~82.1% | 78.7% | Consistently strong performer. |
| Decision Tree (J48) | Yes | Yes | 79.1% | ~75.8% | 78.0% | Lower AUC but reasonable accuracy. |

Comparing our project with the research paper, we observed that both studies established the effectiveness of ensemble learning methods, particularly boosting algorithms, in customer churn prediction modeling. While the first study employed the general-purpose XGBoost algorithm and achieved an AUC of 93.3% using advanced feature engineering (which included Social Network Analysis), our project used LogitBoost in WEKA as a replacement due to tool limitations. Even without real-time big data and features at a high level, LogitBoost alone was able to achieve an AUC of 83.7% and accuracy of 79.4%, demonstrating its capability for being a classifier. Random Forest and Gradient Boosting were also very robust, with Decision Trees returning relatively lower values of AUC.

**Conclusion**

In this project, we observe ensemble learning algorithms or boosting approaches like LogitBoost and AdaBoost being effectively applied in tasks like customer churn prediction, although we couldn't specifically replicate identical conditions and feature applications from that research paper, we did, e.g., employing the WEKA system's incorporated facilities. LogitBoost emerged as the strongest replacement for XGBoost, indicating that similar algorithmic structures can have powerful prediction results. The comparison also shows again that boosting models outperform single-tree classifiers consistently, emphasizing the importance of model selection in data mining. In general, this exercise improved our appreciation of the major classification methods and illustrated how software packages such as WEKA can be used to reproduce scholarly research in real-world data, despite certain tool-based constraints.