INF6028 - Data Mining

Titanic Survivors Classification: A Comparative Analysis of Logistic Regression and Decision Tree Classification using KNIME Workflow

Registration Number – 230140648

Word count – 2510

1. Abstract

This research presents a comprehensive comparative analysis of Logistic Regression (LR) and Decision Tree (DT) Classification models for predicting the survival of passengers aboard the Titanic. The analysis was conducted using the powerful KNIME Analytics Platform workflow, leveraging its robust data processing and machine learning capabilities. The dataset, obtained from Kaggle, consisted of 1204 passenger records, including 452 survivors, and featured attributes such as passenger's name, gender, age, number of siblings and parents aboard, port of embarkation, ticket number, fare, and cabin number. These attributes served as predicting variables, while the dependent variable was the passenger's survival status.

The analysis results revealed that both LR and DT Classification models demonstrated strong predictive performance for classifying Titanic survivors. However, a detailed evaluation based on measures such as accuracy, error rate, and AUC-ROC revealed the superiority of the Logistic Regression model. Specifically, the Logistic Regression model achieved an impressive accuracy of 86%, outperforming the Decision Tree Classification model's accuracy of 85% in predicting the survival outcomes.

Furthermore, the research delved into the feature importance analysis, shedding light on the most influential factors contributing to the models' predictions. Notably, the passenger's gender, age, and fare emerged as the top predictors, aligning with historical accounts of the Titanic.

2. Introduction

The Titanic Survivors prediction problem is a well-known binary classification task in data mining that aims to predict whether a passenger survived the Titanic disaster based on various features such as age, sex, ticket fare, and other characteristics. This problem serves as a valuable exercise in applying supervised machine learning techniques to real-world data. In this report, Logistic regression (LR) and Decision Tree (DT) classifiers have been extensively studied in the context of predicting Titanic survivors using KNIME Platform.

Durmuş & Güneri (2020) demonstrated that logistic regression and Decision Tree could predict Titanic survivors with higher accuracy. By leveraging the power of these data mining methods, this research delves into the data to gain insights and make accurate predictions. The KNIME platform is a comprehensive and user-friendly tool for data processing and machine learning. It offers a wide range of built-in nodes and functionalities, making it convenient for exploratory data analysis, preprocessing, model training, and evaluation.

The dataset used for this analysis is the Titanic Passenger Dataset acquired from Kaggle, which contains information on 1204 passengers aboard the Titanic, including their survival status, age, sex, ticket fare, and other relevant features.

This report walks through the entire machine learning pipeline where ‘Survived’ is the target variable for prediction with the below steps:

1. Exploratory data analysis to understand the dataset and identify missing values or outliers.
2. Data preprocessing, including handling missing values, encoding categorical variables, and scaling numerical features.
3. Building and training the Decision Tree model, tuning hyperparameters, and applying techniques like cross-validation to prevent overfitting.

4. Preprocessing data for Logistic Regression, feature selection, and fitting the Logistic Regression model.
5. Evaluating and comparing the performance of both models using metrics like accuracy, precision, recall, and F1 score.
6. Understanding the strengths and limitations of each algorithm and determining the most suitable approach for predicting Titanic survivorship.

3. Data mining theory

3.1 Decision Trees:

Decision Trees are a type of supervised learning algorithm used for both classification and regression tasks. They work by recursively partitioning the input space into smaller regions based on the feature values, creating a tree-like structure of decisions (Quinlan, J.R,1986). The internal nodes of the tree represent the features, and the branches represent the decision rules based on the feature values. The leaf nodes represent the final predictions or class labels.

DTs are appropriate for the Titanic Survivors prediction problem because they can handle both numerical and categorical data, and they are interpretable, making it easier to understand the decision-making process (Rokach, L., & Maimon, O., 2005). Additionally, DT can automatically handle feature interactions and non-linear relationships without the need for explicit feature engineering.

The performance of can be assessed using various metrics, such as accuracy, precision, recall, F1-score, and area under the Receiver Operating Characteristic curve (AUC-ROC) for classification tasks (Sokolova, M., & Lapalme, G., 2009). Cross-validation techniques, such as k-fold cross-validation, can be employed to obtain an unbiased estimate of the model's performance and to mitigate overfitting.

3.2 Logistic Regression:

Logistic Regression (LR) is a statistical method used for binary classification problems, where the goal is to predict the probability of an instance belonging to one of two classes (Hosmer JR., et.al., 2013). In the case of the Titanic Survivors prediction problem, the classes are "survived" and "not survived". LR models the relationship between the input features and the output probability using the logistic sigmoid function.

It is appropriate for this research question as it can handle both numerical and categorical features, and it provides a probabilistic interpretation of the predictions (Kleinbaum, D. G., et.al., 2002). Additionally, it is relatively simple and easy to interpret, making it a good choice for exploratory data analysis and feature selection.

The performance of LR models can be evaluated using the same metrics as DTs, such as accuracy, precision, recall, F1-score, and AUC-ROC (Fawcett, T., 2006). These evaluation metrics provide quantitative measures of the logistic regression model's performance, allowing analysts to evaluate its effectiveness. Recall, also known as sensitivity, measures the model's ability to identify true positive cases. Precision, on the other hand, measures the model's ability to correctly identify positive cases out of all predicted positive cases. Accuracy represents the overall correctness of the model's predictions, while Error indicates the model's proportion of incorrect predictions.

In addition to these metrics, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is another valuable measure. It quantifies the model's ability to distinguish between true positive and false positive cases across different classification thresholds. A higher AUC-ROC value indicates better discrimination power and predictive performance of the logistic regression model.

Both Decision Trees and Logistic Regression offer unique advantages for predicting Titanic survivors, and their performance metrics provide a comprehensive evaluation of their predictive capabilities. By comparing these models, we can identify the most suitable approach for this specific problem, balancing interpretability and predictive power.

4. Data exploration and preparation

The data preprocessing steps aims to clean, transform, and prepare the data for the subsequent modelling stages, ensuring the data's quality and suitability for the predictive task. In the KNIME workflow, the following steps were taken:

4.1. Data Exploration: The data exploration step involved loading the 'titanic_ticket_data.csv' and 'titanic_personal_data.csv' datasets using the "CSV Reader" node in KNIME. The datasets were then concatenated on the 'PassengerId' column using the "Joiner" node. The "***Statistics***" node was used to explore the data distribution, identify potential outliers, and visualize different features.

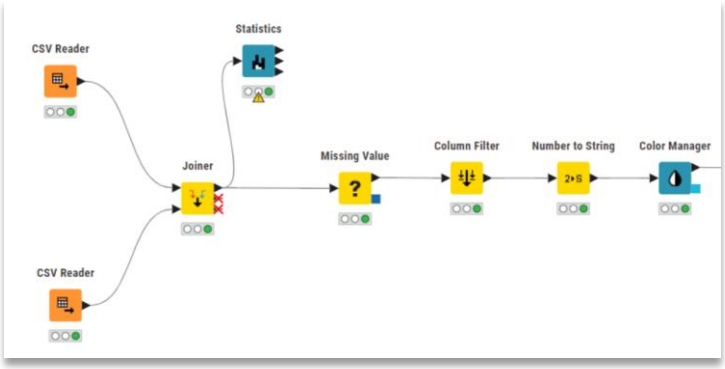

Figure 1: Data Exploration and Preparation Setup

4.2. Feature Selection:

- The "Column Filter" node was employed to remove the 'Cabin' column, which had 942 missing values, and other columns that were identified as not significantly associated with the 'Survived' column based on the Coefficient statistics and Decision Tree branch nodes.
- The accuracy and F-measure were increased by selecting the features that contributed to the classification.

Table 1: Statistics Analysis of the Dataset

Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Overall sum	No. missing
Age	0.17	76	29.81497	14.40176	207.4106	0.388189	0.059744	28682	242
SibSp	0	8	0.501661	1.035928	1.073148	3.808586	19.83152	604	0
Parch	0	9	0.384551	0.874005	0.763884	3.756856	22.31838	463	0
Salary	2.03	4962.37	1222.591	1433.959	2056237	1.181256	0.045128	1471999	0
Survived	0	1	0.375415	0.484431	0.234674	0.515211	-1.73745	452	0
Fare	0	512.3292	33.8085	52.68236	2775.431	4.359255	26.81311	40671.62	1

- The "Missing Value" node was implemented to replace missing values with the 'most frequent value' for string columns and the 'mean' for number columns.

- The "Decision Tree Learner" node was used to automatically rank the features based on their importance in predicting the target variable, allowing for the selection of the most relevant features.

4.3. Feature Transformation and Normalization:

- The 'Survived' column was converted from a number type to a string type and color-coded using the 'Number to String' and 'Colour Manager' nodes.
- Numerical features were normalized using the "Normalizer" node to ensure that all features were on the same scale, preventing features with larger ranges from dominating the Logistic Regression model.

5. Experimental setup

This section describes the experimental design in the workflow,

***5.1. Data Partitioning*:** The "Partitioning" node was used to split the dataset into training and testing sets of 80:20 ratio. A stratified sampling approach was employed to ensure that the class distribution in the training and testing sets was representative of the overall dataset.

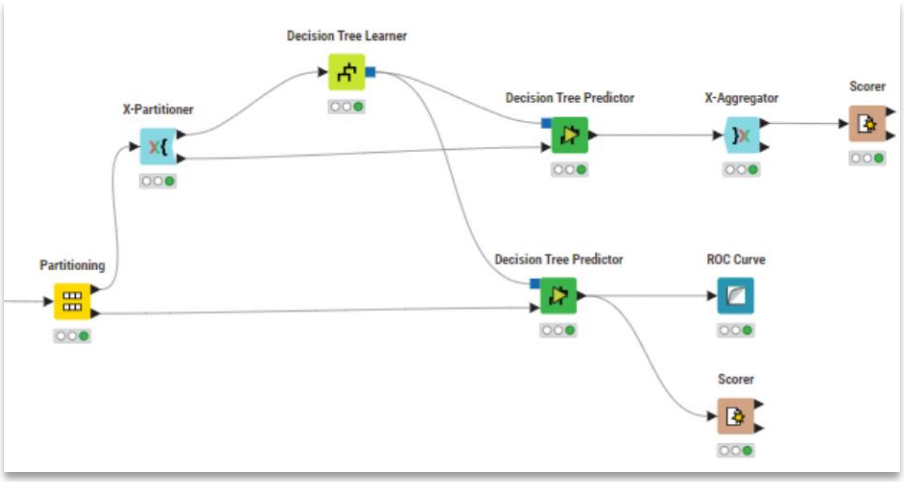

Figure 2: Decision Tree Environmental Setup

***5.2. Model Training and Evaluation*:** The "Decision Tree Learner" and "Logistic Regression Learner" nodes were used to train the respective models on the training data. The “Decision Tree Predictor” and the “Logistic Regression Predictor” nodes were employed to predict the survival. The "Scorer" and the “ROC Curve” nodes were used to evaluate the performance of the models on the testing data using various metrics, such as accuracy, precision, recall, F1-score, and AUC-ROC.

5.3. Cross-Validation: To mitigate overfitting in the “Decision Tree” algorithm and obtain a more reliable estimate of the models' performance, k-fold cross-validation was performed using the "X-Partitioner" and "X-Aggregator" nodes. The process involved 10 validations with stratified sampling. The average performance metrics across the folds were reported.

However, the Decision Tree model initially showed signs of overfitting, with a high training accuracy (95%) and a low training error (5%), which differed significantly from the test accuracy. To address this issue, k-fold cross-validation was introduced for the Decision Tree method. However, the Logistic

Regression method did not perform well with k-fold cross-validation. To improve the performance of the Logistic Regression model, normalization was implemented.

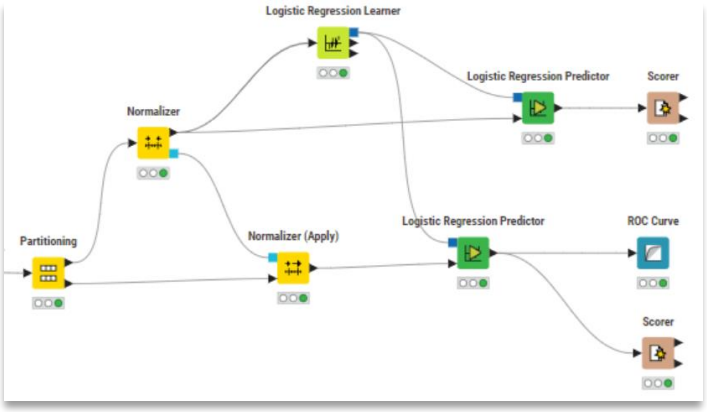

Figure 3: Logistic Regression Environmental Setup

5.4. Hyperparameter Tuning: The "Decision Tree Learner" and "Logistic Regression Learner" nodes in KNIME offer various hyperparameters that can be tuned to optimize the models' performance. For both models, the default options were selected. Additionally, the "Number of threads" was set to 8, and the "Binary Nominal Splits" option was checked for splitting purposes in the Decision Tree Learner.

6. Results

The models achieved reasonably good performance, with the Logistic Regression model slightly outperforming the Decision Tree model across all metrics. The LR model had higher accuracy, F-measure and AUC-ROC values, indicating its superiority in predicting passenger survival on the Titanic.

Table 2 and Table 3 summarizes the performance of the models on the Titanic Survivors prediction problem:

Table 2: Confusion Matrix and Accuracy Statistics – Decision Tree

RowID	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	F-measure	Accuracy	Cohen's kappa
0	136	19	71	15	0.900662	0.877419	0.888889	0.858921	0.695797
1	71	15	136	19	0.788889	0.825581	0.806818		

Table 3: Confusion Matrix and Accuracy Statistics – Logistic Regression

RowID	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	F-measure	Accuracy	Cohen's kappa
0	137	19	71	14	0.907285	0.878205	0.892508	0.863071	0.704074
1	71	14	137	19	0.788889	0.835294	0.811429		

In the Decision Tree model, the feature importance analysis highlighted that the passenger's gender was the most significant predictor of survival. This is visually represented in the tree's structure, where branches and nodes prominently feature splits based on the passenger's gender. Further analysis revealed that the most influential features for predicting survival were sex, age and fare (representing socioeconomic status) and the significant scores indicated in Table 4 and Figure 4 below. This aligns with the historical accounts of the Titanic disaster, where women and children from higher socioeconomic classes were given priority during the evacuation process.

Table 4: Confusion Matrix and Accuracy Statistics – Logistic Regression

RowID	Logit	Variable	Coeff.	Std. Err.	z-score	P>|z|
Row1	1	Sex=male	-3.71343	0.208934	-17.7732	0
Row2	1	Age	-2.2143	0.616128	-3.59389	0.000326
Row3	1	Salary	2.670792	0.398881	6.695713	2.15E-11
Row4	1	Fare	1.019811	1.149481	0.887192	0.374975
Row5	1	Constant	1.793295	0.268915	6.66862	2.58E-11

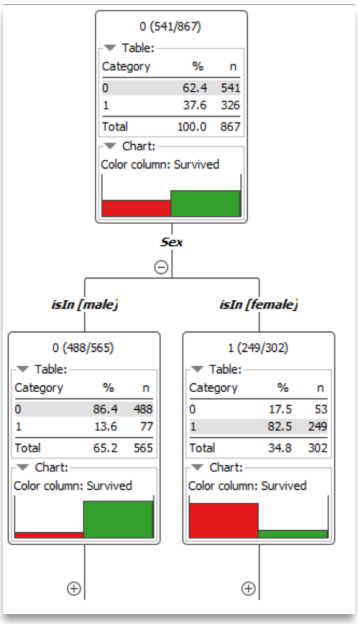

Figure 4: Decision Tree Map

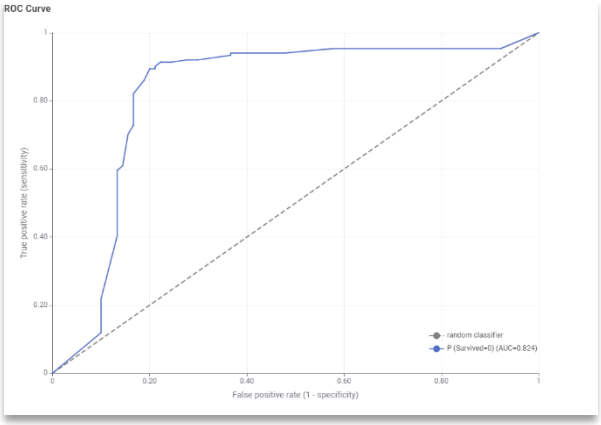

Figure 5: ROC Curve Decision Tree

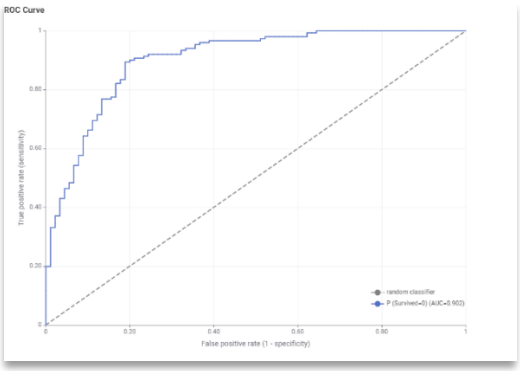

Figure 6: ROC Curve Logistic Regression

As evident from the Figure 4, top features of the decision tree map are the passenger's gender as a primary predictor of survival. The AUC-ROC value for Logistic Regression is 0.902, demonstrating its robustness in distinguishing between survived and non-survived passengers This high AUC-ROC value reflects the model's strong ability to predict the probability of a passenger's survival correctly. The Decision Tree model, while slightly less performant, still achieved a commendable AUC-ROC value of 0.824. This indicates that although the Decision Tree model is effective, it is slightly less accurate in its predictions compared to the LR model.

In this analysis, the Logistic Regression model produced the best performance, with higher accuracy, F-measure, and AUC-ROC values compared to the Decision Tree model. The Logistic Regression model's ability to handle multicollinearity and provide probability estimates for the outcome, coupled with careful feature selection and preprocessing, likely contributed to its superior performance. However, it is important to note that the Decision Tree model still achieved good results and can be valuable in situations where interpretability and capturing non-linear relationships are prioritized.

7. Conclusion and Reflections

In this analysis, we explored the application of Logistic Regression and Decision Tree models for predicting the survival of passengers aboard the Titanic. The results demonstrated that both models achieved reasonably good performance, with the Logistic Regression model slightly outperforming the Decision Tree model across various evaluation metrics

While the Logistic Regression model produced the best performance in this analysis, it is essential to consider the advantages and disadvantages of each method. Logistic Regression models are simple, interpretable, and handle multicollinearity well, but they assume linearity between the independent variables and the log odds, and can be sensitive to outliers. On the other hand, Decision Tree models can handle both categorical and numerical data, capture non-linear relationships, and provide interpretable visualizations, but they are prone to overfitting and can be unstable with small variations in the data.

Reflecting on the choice of methods for this problem, the Logistic Regression model's performance could potentially be improved by exploring more advanced feature selection techniques, such as recursive feature elimination or regularization methods (El-Koka, A., et.al., 2013). These techniques can help identify the most relevant features and mitigate the impact of multicollinearity, potentially improving the model's accuracy and generalization capabilities.

Additionally, ensemble methods like Random Forests or Gradient Boosting could be explored to improve the performance of the Decision Tree model (A. Singh, et.al., 2017, Tabbakh, A., et.al., 2021). These methods combine multiple Decision Trees, reducing the impact of individual tree biases and potentially achieving better predictive performance.

In conclusion, while both Logistic Regression and Decision Tree models demonstrated good performance in predicting the survival of Titanic passengers, there is room for improvement by exploring advanced feature selection techniques, ensemble methods, and incorporating domain knowledge. According to D. Collaris and J. J. van Wijk (2023), the choice of method ultimately depends on the specific problem requirements, such as the need for interpretability, handling non-linear relationships, or the availability of expert knowledge.

References

[1] Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1*(1), 81-106. https://doi.org/10.1007/BF00116251

[2] Rokach, L., & Maimon, O. (2005). Decision trees. In Maimon, O., & Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA. https://doi.org/10.1007/0-387-25465-X_9

[3] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45*(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

[4] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

[5] Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002).Logistic regression. Springer.

[6] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

[7] Durmuş, Burcu & Isci Guneri, Oznur. (2020). Analysis and detection of Titanic survivors using generalized linear models and decision tree algorithm. International Journal of Applied Mathematics Electronics and Computers. https://doi.org/10.18100/ijamec.785297

[8] Collaris, D., & van Wijk, J. J. (2023). StrategyAtlas: Strategy analysis for machine learning interpretability. IEEE Transactions on Visualization and Computer Graphics, 29(6), 2996-3008. https://doi.org/10.1109/TVCG.2022.3146806

[9] Singh, A., Saraswat, S., & Faujdar, N. (2017). Analyzing Titanic disaster using machine learning algorithms. 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 406-411. https://doi.org/10.1109/CCAA.2017.8229835

[10] Tabbakh, A., Rout, J. K., & Rout, M. (2021). Analysis and prediction of the survival of Titanic passengers using machine learning. In Tripathy, A., Sarkar, M., Sahoo, J., Li, K. C., & Chinara, S. (Eds.), Advances in Distributed Computing and Machine Learning (Vol. 127). Springer, Singapore. https://doi.org/10.1007/978-981-15-4218-3_29

[11] El-Koka, A., Cha, K.-H., & Kang, D.-K. (2013). Regularization parameter tuning optimization approach in logistic regression. 2013 15th International Conference on Advanced Communications Technology (ICACT), PyeongChang, Korea (South), 13-18.