Analysis of Decision Tree Classification Algorithm on Different Datasets using KNIME

Syeda Tahira Batool
Department of Computer Science
Riphah International University
Lahore, Pakistan
tahirabatool55@gmail.com

Anum Ghaffar
Department of Computer Science
COMSATS University Islamabad
Sahiwal campus, Pakistan
anumghaffar9@gmail.com

Arooj Fatima
Department of Computer Science
Riphah International University
Lahore, Pakistan
aroojkhalid122@gmail.com

Syed Mujtaba Hassan
Department of Computer Science
Riphah International University
Lahore, Pakistan
smhassan63@gmail.com

Imdad Hussain
Department of Computer Science
Riphah International University
Lahore, Pakistan
imdaadhussain@gmail.com

***Abstract* — Classification is a data mining techniques used for prediction group membership for data instances. There are numerous classification techniques that can be apply to achieve goals. In this study main concern is to implementing decision tree classification technique on two different datasets i.e. social networks ads dataset and gender classification dataset. The purpose of this paper is to present performance analysis of decision tree algorithm on different data sets. KNIME analytical tool is used to perform the purposed classification algorithm to collect the results. This study will be beneficial for institutes and novice in the domain of machine learning to further enhance the core of classification techniques.**

***Keywords*— social networks ads data sets, analysis, classification, Decision tree algorithm, KNIME tool**

I. INTRODUCTION

Machine Learning is the training of model to maximize the efficiency using data or prior experience. Some parameters have sets to define model, and system learning is the implementation of programming to maximize the variables through the training data or prior work. The model can be used for define purposes, it may be the predictions of future or detailed to acquire information from data.

Classification techniques have a vast scope of uses like artificial intelligence, fraud detection, churn prediction and credit card rating etc. Moreover, numerous classification algorithms open in previous work but purposed algorithm is generally employed on account of its ease of execution and simpler to comprehend by comparing to remaining algorithms of classification [1].

Decision Tree algorithm is a type of data mining technique that is employed to generate Classification models [9]. As like its name it basis tree-like structure to build model for classification. This kind of mining relates to supervised learning. Model is build using training data sets and performance is measured by providing test data sets.

In this study, we used decision tree classification algorithm on social media ads dataset and gender classification dataset. Performance or efficiency are used to measure of the purposed algorithm on taken datasets. For the implementation of decision tree classification model we used open source software for developing data science KNIME analytical tool [15].

The aim of this paper are as follow

a) To examine the performance of purposed algorithm on taken datasets in KNIME tool.

b) To help novice for better understanding the performance of algorithm.

The remaining section of paper is proceed as. (Part II) Literature review. (Part III) presents research questions. (Part IV) presents methodology (Part V) reviews purposed algorithm. (Part VI) reviews results and discussion and then last (Part VII) discuss the conclusions and future work of the paper.

II. LITERATURE REVIEW

Aized Amin Soofi et al [2] proposed a study where they discussed different application and issues using classification techniques in machine learning. The aim of their paper is present reviews of different machines learning classification techniques such as Decision Tree, SVM, K-NN, and Bayesian Network. Their result is to mention the solution of issues and applications.

Sam fletcher et al [3] presented a survey in which they were focused on one specific classification algorithm i.e. decision tree algorithm. They examine greedy decision trees techniques as well as random decision trees techniques, and disputes which appear when attempt to compensate privacy rules with the pattern's accuracy.

Srabanti Maji et al [4] proposed a paper to examine the performance of decision tree algorithm on heart disease patients' dataset. They used Weka tool for the implementation of purposed algorithm. Their results show that purposed algorithm gave best accuracy on taken dataset.

Nageswari S et al [5] presented a study in which they comparison different classification techniques including decision tree. They used educational dataset of student's performance to examine the accuracy of algorithms. According to their result decision tree and neural networks gave best accuracy.

Majeed, Fiaz et al [6] proposed a paper for analysis of 28k tweets of healthcare firms collected from 13 news channels. Comparative analysis of different algorithm were performed on it to analyze the accuracy, recall, precision and f-measure values.

III. RESEARCH QUESTIONS

In this study we discussed decision tree classification algorithm and this algorithms are applied on social network ads data set and gender classification data set, after applying we get their accuracy and conclude result on the given data set. This research cover the following question;

1. Is this algorithm best for taken type of datasets?
2. What accuracy result we get from these type of datasets using proposed algorithm?

IV. METHODOLOGY

For analysis of social networks ads dataset and gender classification dataset, below figure of workflow shown from start to end level using advanced KNIME analytical tool.

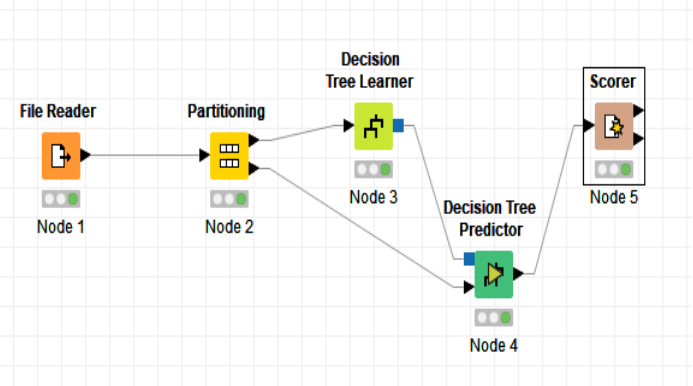


Fig 1. Workflow in KNIME

Decision algorithm performed on two different datasets to compare their performance. KNIME analytical tool version 4.1.2 used for Implementation. Computed system installation based on Windows 10, 64-bit operating system, CPU Intel ® Core i3 having 6 GB random access memory.

A. Data Access

Initially, file reader node is used in workflow to access the dataset for further work. Social networks ads dataset contains 5 attributes and 400 instance has taken from kaggle. Below figure show the table of taken datasets

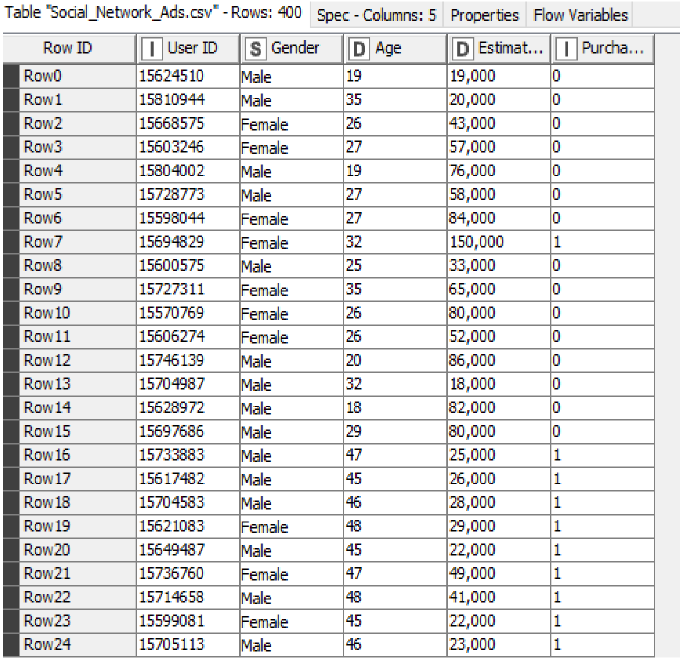


Table "Social_Network_Ads.csv" - Rows: 400 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	I User ID	S Gender	D Age	D Estimat...	I Purcha...
Row0	15624510	Male	19	19,000	0
Row1	15810944	Male	35	20,000	0
Row2	15668575	Female	26	43,000	0
Row3	15603246	Female	27	57,000	0
Row4	15804002	Male	19	76,000	0
Row5	15728773	Male	27	58,000	0
Row6	15598044	Female	27	84,000	0
Row7	15694829	Female	32	150,000	1
Row8	15600575	Male	25	33,000	0
Row9	15727311	Female	35	65,000	0
Row10	15570769	Female	26	80,000	0
Row11	15606274	Female	26	52,000	0
Row12	15746139	Male	20	86,000	0
Row13	15704987	Male	32	18,000	0
Row14	15628972	Male	18	82,000	0
Row15	15697686	Male	29	80,000	0
Row16	15733883	Male	47	25,000	1
Row17	15617482	Male	45	26,000	1
Row18	15704583	Male	46	28,000	1
Row19	15621083	Female	48	29,000	1
Row20	15649487	Male	45	22,000	1
Row21	15736760	Female	47	49,000	1
Row22	15714658	Male	48	41,000	1
Row23	15599081	Female	45	22,000	1
Row24	15705113	Male	46	23,000	1

Fig 2. Social network ads dataset

Gender classification dataset consists of 5 attributes and 67 instances. It consists of data based on personal preferences. This dataset has taken from kaggle.

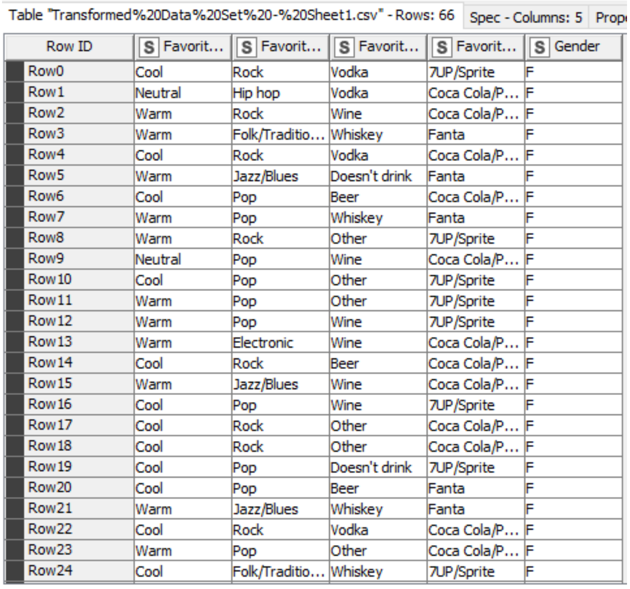


Table "Transformed%20Data%20Set%20-%20Sheet1.csv" - Rows: 66 | Spec - Columns: 5 | Prop

Row ID	S Favorit...	S Favorit...	S Favorit...	S Favorit...	S Gender
Row0	Cool	Rock	Vodka	7UP/Sprite	F
Row1	Neutral	Hip hop	Vodka	Coca Cola/P...	F
Row2	Warm	Rock	Wine	Coca Cola/P...	F
Row3	Warm	Folk/Traditio...	Whiskey	Fanta	F
Row4	Cool	Rock	Vodka	Coca Cola/P...	F
Row5	Warm	Jazz/Blues	Doesn't drink	Fanta	F
Row6	Cool	Pop	Beer	Coca Cola/P...	F
Row7	Warm	Pop	Whiskey	Fanta	F
Row8	Warm	Rock	Other	7UP/Sprite	F
Row9	Neutral	Pop	Wine	Coca Cola/P...	F
Row10	Cool	Pop	Other	7UP/Sprite	F
Row11	Warm	Pop	Other	7UP/Sprite	F
Row12	Warm	Pop	Wine	7UP/Sprite	F
Row13	Warm	Electronic	Wine	Coca Cola/P...	F
Row14	Cool	Rock	Beer	Coca Cola/P...	F
Row15	Warm	Jazz/Blues	Wine	Coca Cola/P...	F
Row16	Cool	Pop	Wine	7UP/Sprite	F
Row17	Cool	Rock	Other	Coca Cola/P...	F
Row18	Cool	Rock	Other	Coca Cola/P...	F
Row19	Cool	Pop	Doesn't drink	7UP/Sprite	F
Row20	Cool	Pop	Beer	Fanta	F
Row21	Warm	Jazz/Blues	Whiskey	Fanta	F
Row22	Cool	Rock	Vodka	Coca Cola/P...	F
Row23	Warm	Pop	Other	Coca Cola/P...	F
Row24	Cool	Folk/Traditio...	Whiskey	7UP/Sprite	F

Fig. 3 Gender Classification Dataset

B. *Data partition*

For training purpose, we used 80% data and for testing purpose, used 20% data. For this purpose we used Partitioning node in workflow to partition data into test and train data.

V. PURPOSED ALGORITHM

We selected decision tree algorithms for analyze their performance. For the implementation of this techniques we use KNIME analytical tool.

A. Decision Tree

It generate classification or regression models in terms of a tree structure. It is a supervised learning techniques that use both continuous and discrete parameters for work. It divide the data into sub instances on the basic of the most important parameters in the data.

How the algorithm determines this parameters and how this division is made definite by decision tree. It will remain working unless a quit conditions for instance the lowest figure of observations etc. is come. Decision nodes and leaf nodes is the conclusive forms of tree.

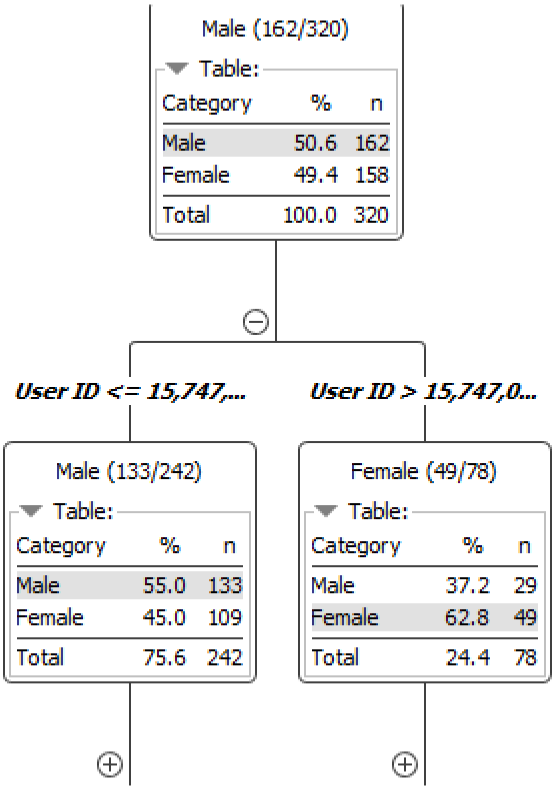
VI. RESULTS AND DISCUSSION

By compare the result of both datasets, gender classification type dataset has high accuracy as compare to social networks ads dataset.

TABLE I COMPARISON

DATASETS	ACCURACY
Social Network ads	47.5%
Gender Classification	64.286%

Below diagram show the decision tree view of predicator after providing social networks ads train and test data for modeling.



Male (162/320)

Table:

Category	%	n
Male	50.6	162
Female	49.4	158
Total	100.0	320

Male (133/242)

Table:

Category	%	n
Male	55.0	133
Female	45.0	109
Total	75.6	242

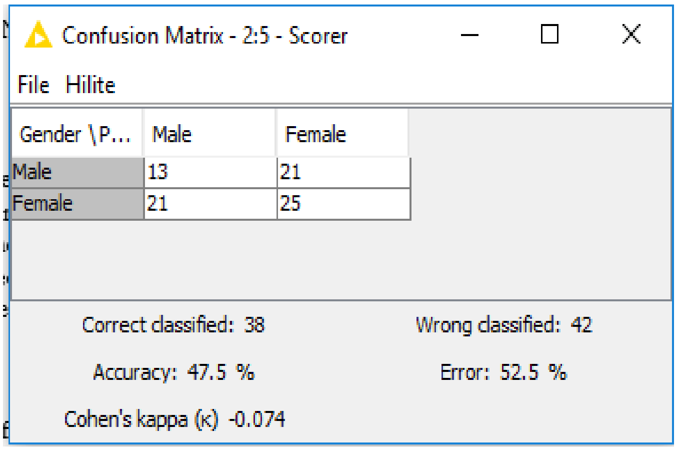
Female (49/78)

Table:

Category	%	n
Male	37.2	29
Female	62.8	49
Total	24.4	78

Fig 4. Decision tree view of Social network ads dataset

Below figure show confusion matrix of social networks ads dataset as well as accuracy 47.5%, correct classified 38 and wrong classified 42.



Gender \P...	Male	Female
Male	13	21
Female	21	25

Correct classified: 38 Wrong classified: 42

Accuracy: 47.5 % Error: 52.5 %

Cohen's kappa (κ) -0.074

Fig 5. Confusion matrix of social networks ads dataset

Below diagram show the decision tree view of predicator after providing gender classification train and test data for modeling.

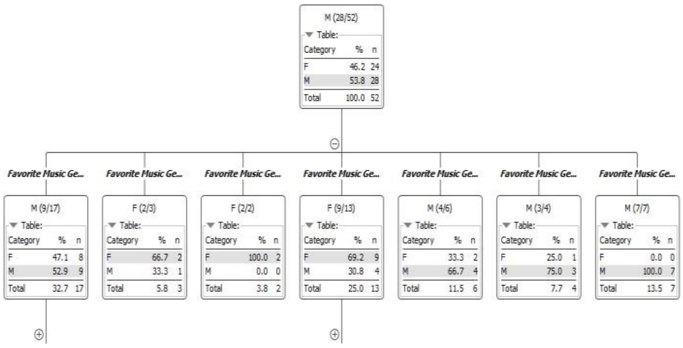
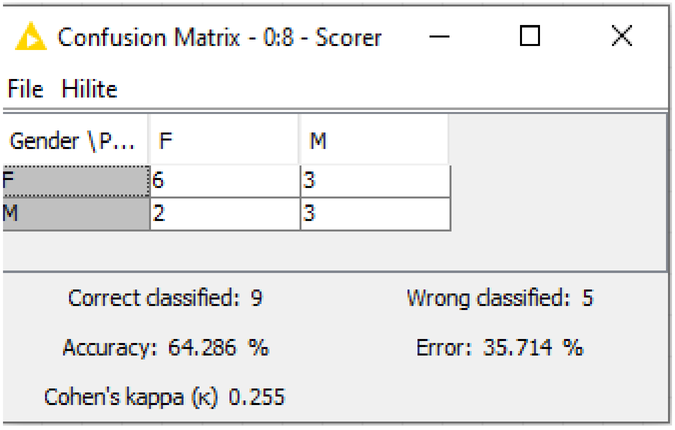


Fig 6. Decision tree view of gender classification dataset

Below figure show confusion matrix of gender classification dataset as well as accuracy 64.286%, correct classified 9 and wrong classified 5.



Gender \P...	F	M
F	6	3
M	2	3

Correct classified: 9 Wrong classified: 5

Accuracy: 64.286 % Error: 35.714 %

Cohen's kappa (κ) 0.255

Fig 7. Confusion matrix of gender classification

VII. CONCLUSION

There are numerous classification techniques that can be apply for classification to achieve goals. In this study main concern is to implementing decision tree classification technique on two different datasets i.e. social networks ads dataset and gender classification dataset.

The purpose of this paper is to present performance analysis of decision tree algorithm on different datasets. KNIME analytical tool is used to perform the purposed classification algorithm to collect the results. In this proposed study, we used decision tree classification algorithm on two different types of datasets i.e. social networks ads dataset and gender classification datasets. Purposed Classification algorithms applied to analyze their performance. By compare the result of both datasets, gender classification type dataset has high accuracy i.e. 64.287% as compare to social networks ads dataset accuracy i.e. 47.5%.

In future, we can use different classification techniques on different datasets for their comparison and to get accuracy.

REFERENCES

[1] Anyanwu, Matthew & Shiva, S.. (2009). Comparative Analysis of Serial Decision Tree Classification Algorithms. International Journal of Computer Science and Security. 3(3).

[2] Soofi, A.A., & Awan, A. (2017). Classification Techniques in Machine Learning: Applications and Issues. *Journal of Basic and Applied Sciences, 13*, 459-465.

[3] Sam Fletcher and Md. Zahidul Islam. 2019. Decision Tree Classification with Differential Privacy: A Survey. ACM Comput. Surv. 52, 4, Article 83 (September 2019), 33 pages. DOI:https://doi.org/10.1145/3337064

[4] Maji S., Arora S. (2019) Decision Tree Algorithms for Prediction of Heart Disease. In: Fong S., Akashe S., Mahalle P. (eds) Information and Communication Technology for Competitive Strategies. Lecture Notes in Networks and Systems, vol 40. Springer, Singapore

[5] S, Nageswari and Goel, Pallavi M., Comparison of Classification Techniques on Data Mining (April 19, 2019). International Journal of Emerging Technology and Innovative Engineering, Volume 5, Issue 5, May 2019 . Available at SSRN: https://ssrn.com/abstract=3375191

[6] Majeed, F., Asif, M.W., Hassan, M.A., Abbas, S.A., & Lali, M.I. (2019). Social Media News Classification in Healthcare Communication. *Journal of Medical Imaging and Health Informatics, 9*, 1215-1223.

[7] Ozbay, Feyza & Alatas, Bilal. (2019). Fake news detection within online social media using supervised artificial intelligence algorithms. Physica A: Statistical Mechanics and its Applications. 540. 123174. 10.1016/j.physa.2019.123174.

[8] Mohd Fauzi bin Othman, "Comparison of Different Classification Techniques Using WEKA for Breast Cance., 2007

[9] Rafet Duriqi, Vigan Raca, "Comparative Analysis of Classification Algorithms on Three Different Datasets using WEKA", 5th Mediterranean Conference on Embedded Computing2016.

[10] Rohit Arora, Suman," Comparative Analysis of Classification Algorithms on Different Datasets using WEKA", International Journal of Computer Application September 2012
.

[11] SAGAR S. NIkAM" A Comparative Study of Classification Techniques in Data Mining Algorithms", Computers & Education, 2015

[12] Chong Sun1, Narasimhan Rampalli1, Frank Yang," Chimera: Large-Scale Classification using Machine Learning, Rules, and Crowdsourcing".

[13] Chitra Jalota, Rashmi Agrawal," Analysis of Educational Data Mining using Classification", International Conference on Machine Learning, Big Data, Cloud and Parallel Computing Feb 2019.

[14] SAGAR S. NIkAM," A Comparative Study of Classification Techniques in Data Mining Algorithms", oriental journal of computer science & technology April 2015.

[15] Open for Innovation KNIME. Retrieved from https://www.knime.com/

[16] Jun Lee, S. and Siau, K. (2001), "A review of data mining techniques", *Industrial Management & Data Systems*, Vol. 101 No. 1, pp. 41-46. https://doi.org/10.1108/02635570110365989

[17] Maulana, Mohamad & Defriani, Meriska. (2020). Logistic Model Tree and Decision Tree J48 Algorithms for Predicting the Length of Study Period. PIKSEL : Penelitian Ilmu Komputer Sistem Embedded and Logic. 8. 39-48. 10.33558/piksel.v8i1.2018.

[18] Amos Pah, Clarissa & Utama, Ditdit. (2020). Decision Support Model for Employee Recruitment Using Data Mining Classification. 8. 1511-1516. 10.30534/ijeter/2020/06852020.