BUSA3020: Advanced Analytics Techniques

ASSIGNMENT 2: PREDICTIVE ANALYTICS

45345473

Amritha Jeyarathan

TABLE OF CONTENTS

I. INTRODUCTION

I A. PURPOSE

To utilise various prediction tools on the *Titanic* dataset from two alternative software to recommend which software is most beneficial to a client.

II. METHODOLOGY

II A.THE DATA CLEANING PROCESS

III A. STEP 1

Variables were assigned an appropriate data type (*figure 1*). *Orange Data Miner* was utilised for the data cleaning process and followed the workflow shown in *figure 2. Figure 3* illustrates the data, prior to and following, the data cleaning process.

III A. STEP 2

The *Impute* function was employed to remove missing values by setting those in variables *Age* and *Passenger Fare* to their average, and those in *Port of Embarkation* to 'Cherbourg'.

III A. STEP 3

The *Feature Constructor* function was used to create new variables *has Cabin* and *Family Size* (*figure 4*).

Variable	Data type
Survived	Categorical
Passenger	Categorical
Sex	Categorical
Cabin	Categorical
Port of Embarkation	Categorical
Age	Numerical
No. of Siblings	Numerical
No. of Parents	Numerical
Passenger Fare	Numerical
Name	Text
Ticket Number	Nil (variable ignored)
Lifeboat	Nil (variable ignored)

Figure 1. Each variable was assigned an appropriate data type, variables *Ticket Number* and *Lifeboat* were ignored as their effect on whether a passenger survives was considered insignificant.

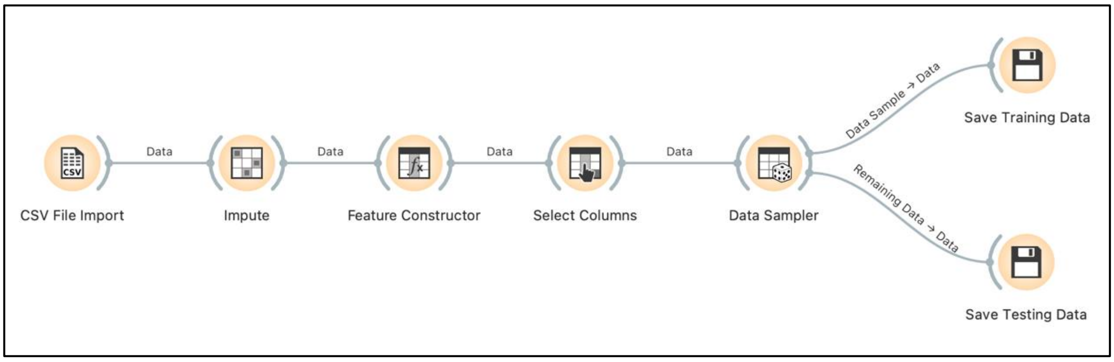

Figure 2. The overall workflow on *Orange Data Miner*. The functions *Impute*, *Feature Constructor*, *Select Columns* and *Data Sampler* were used to clean the data and to create both a testing data subset and a training data subset.

Variable Name	Missing values (Prior to cleaning)	Missing values (Following *Impute*)	Missing values (Following *Feature Construction*)
Passenger Fare	1 (0%)	0 (0%)	0 (0%)
No. of Parents or Children on Board	0 (0%)	0 (0%)	N/A (removed)
No. of Siblings or Spouses on Board	0 (0%)	0 (0%)	N/A (removed)
Age	263 (20%)	0 (0%)	0 (0%)
Port of Embarkation	2 (0%)	0 (0%)	0 (0%)
Cabin	1,014 (77%)	1,014 (77%)	N/A (removed)
Sex	0 (0%)	0 (0%)	0 (0%)
Passenger Class	0 (0%)	0 (0%)	0 (0%)
Survived	0 (0%)	0 (0%)	0 (0%)
Family Size	N/A	N/A	0 (0%)
Has Cabin	N/A	N/A	0 (0%)

Figure 3. Missing values obtained from the feature statistics of the dataset prior to cleaning, following imputing the missing values and following the utilisation of the *Feature Constructure* function to create the variables *has Cabin* and *Family Size*. The variables *No. of Siblings or Spouses on Board* and *No. of parents or children on Board* were removed from the dataset as they are reflected in *Family Size*. *Cabin* was also removed as its values are reflected in *has Cabin*. Additionally, there are now no missing values in any of the variables post-cleaning.

Variable	Equation
Has Cabin	= isnan(Cabin)
Family Size	= No_of_Siblings_or_Spouses_on_Board + No_of_Parents_or_Children_on_Board

Figure 4. The equations formed to create: the new variable *has Cabin*, which only holds a value of either '0' or '1' based on whether the corresponding value in the variable *Cabin* exists or not i.e. whether or not the individual in question has a cabin, where 0 = no and 1 = yes; and *Family Size*, which is simply the sum of the corresponding values in the variables *No. of Siblings or Spouses on Board* and *No. of parents or children on Board*. This variable further simplifies the data and allows the user to easily identify the number of family members a passenger may have.

II B. CREATING TRAINING AND TESTING DATASETS

Training and testing datasets were then created using the *Data Sampler* function in *Orange* and using the *Partitioning* function in *KNIME* (see evaluation on IV A. inoperability), with 70% of the data being allocated to a training subset and the remaining 30% into a testing subset.

II C. TESTING ALTERNATIVE PREDICTION ALGORITHMS

The table below illustrates the various prediction algorithms tested on both *Orange Data Miner* and *KNIME* and the reasoning behind selecting these algorithms.

Algorithm Tested	Reasoning
Decision Tree	Decision trees are the simplest of the algorithms and can be easily understood and interpreted by users with very little knowledge on predictive analytics. When visualised, a user can easily distinguish variables that contribute to a passenger having a higher survival rate and thus, can predict whether a certain individual will survive.
Random Forest	Random forests aggregate multiple decision trees and are particularly useful for large datasets. The algorithm was included in the testing as it is more powerful and accurate than using a decision tree. Unlike decision

	trees, random forests choose features randomly i.e., they do not give more importance to certain features. They also control any effects of overfitting whilst minimising bias.
Logistic Regression	Logistic regressions are particularly useful in predictive analytics when the dependent variable is binary and not continuous i.e., can only take on two values. In this case it is appropriate, as the variable *Survived* takes on the values *yes* or *no*. As such, logistic regressions are more fitting compared to linear regressions, as they can show the probability that a passenger survives or does not survive.
Neural Network (pNN in *KNIME*)	Neural networks are incredibly powerful and accurate at predictive analytics and are particularly beneficial for large datasets. It outperforms decision trees, random forests and logistic regressions as these algorithms rely on the input and output nodes to predict whether an individual survives. Neural networks use the 'hidden layer/s', as they have the ability to learn in a way that is similar to the processing of the human brain. A probabilistic neural network (pNN) was utilised in *KNIME*. pNN is a four-layered neural network.

III. ANALYSIS

III A. CRITERIA FOR ASSESSING SOFTWARE

Criterion	Description
Accuracy	How well does the software provide accurate results? That is, results that are produced with little bias or error.
Suitability	How suitable is the software for predictive analytics? Does the software provide a wide range of predictive algorithms?
Interoperability	How well does the software integrate with outside sources? How easily does it accept data?
Learnability	How easily can a user with no experience in predictive analytics learn to use the software? If learning materials are provided, are they clear and understandable?
Attractiveness	How visually appealing is the software?
Functionality	How practical is the software? Can it be used in an efficient and effective manner by the user?
Interpretation of Output	How effective is the software in interpreting and visualising the data output?
Installability	How easily can the software be installed?
Accessibility	Can the software be used on a range of operating systems? Is the software accessible to everyone?

Figure 5. The criteria on which each software program, *Orange Data Miner* and *KNIME*, will be assessed against. Following the assessment, an appropriate mark will be awarded for each criterion (see below).

III B. MARKING CRITERIA

Mark	Description
5	Far Exceeds Expectations
4	Exceeds Expectations
3	Meets Expectations
2	Approaches Expectations
1	Below Expectations

Figure 6. The marking criteria on which an appropriate mark will be awarded for each criterion listed in figure 10 for each software program.

IV. RESULTS

IV A. EVALUATION OF SOFTWARE ON CRITERIA

The following tables expand on an evaluation on each software program based on the given criterion.

IV A. CRITERION: ACCURACY

Criterion	Mark Assigned		Justification
	Orange	KNIME	
Accuracy	4	3	*Orange Data Miner* proved to have slightly greater accuracy rates than *KNIME* (as seen in the confusion matrices and accuracy statistics tables below).

CONFUSION MATRIX: DECISION TREE (PROPORTION OF PREDICTED)

		Predicted in *Orange*			Predicted in *KNIME*		
Actual		**No**	**Yes**	**Sum**	**No**	**Yes**	**Sum**
	No	88.5%	8.4%	**242**	74.1%	19.3%	**162**
	Yes	11.5%	91.6%	**150**	25.9%	80.7%	**229**
	Sum	**261**	**131**	**392**	**158**	**233**	**391**

Figure 7. When employing a decision tree, the predictions made in *Orange Data Miner* were more accurate to the actual values compared to those produced in *KNIME*. For example, 91.6% of the actual values 'YES' in the variable *Survived* in *Orange* were predicted correctly, compared to 80.7% in *KNIME*. Note that bias may be present, as the total sum does not match between the two programs, Nor does the total in *KNIME* agree to its corresponding totals in the following three tables, this is likely due to an error in *KNIME* where two values were not considered.

CONFUSION MATRIX: RANDOM FOREST (PROPORTION OF PREDICTED)

		Predicted in *Orange*			Predicted in *KNIME*		
Actual		**No**	**Yes**	**Sum**	**No**	**Yes**	**Sum**
	No	93.4%	10.1%	**242**	79.0%	20.8%	**162**
	Yes	6.6%	89.9%	**150**	21.0%	79.2%	**231**
	Sum	**243**	**149**	**392**	**138**	**255**	**393**

Figure 8. When utilising a random forest, the predictions made in *Orange Data Miner* were more accurate to the actual values compared to those produced in *KNIME*. For example, 89.9% of the actual values 'YES' in the variable *Survived* in *Orange* were predicted correctly, compared to 79.2% in *KNIME*. Note that bias may be present, as the total sum does not match between the two programs.

CONFUSION MATRIX: LOGISTIC REGRESSION (PROPORTION OF PREDICTED)

		Predicted in *Orange*			Predicted in *KNIME*		
Actual		**No**	**Yes**	**Sum**	**No**	**Yes**	**Sum**
	No	80.5%	25.9%	**242**	69.1%	17.5%	**162**
	Yes	19.5%	74.1%	**150**	30.9%	82.5%	**231**
	Sum	**257**	**135**	**392**	**181**	**212**	**393**

Figure 9. When using a logistic regression, the predictions made in *Orange Data Miner* were more accurate to the actual values compared to those produced in *KNIME* for the value 'NO', but less accurate for the values 'YES'. For example, 80.5% of the actual values 'NO' in the variable *Survived* in *Orange* were predicted correctly, compared to 69.1% in *KNIME*; and 74.1% of the actual values 'YES' was correct in *Orange*, while 82.5% was correct in *KNIME*. Note that bias may be present, as the total sum does not match between the two programs.

CONFUSION MATRIX: NEURAL NETWORK (PROPORTION OF PREDICTED)

		Predicted in *Orange*			**Predicted in *KNIME***		
		No	**Yes**	**Sum**	**No**	**Yes**	**Sum**
Actual	**No**	83.5%	18.9%	**242**	60.0%	32.5%	**162**
	Yes	16.5%	81.1%	**150**	40.0%	67.5%	**231**
	Sum	**260**	**132**	**392**	**125**	**268**	**393**

Figure 10. When employing a neural network (pNN was used in *KNIME*), the predictions made in *Orange Data Miner* were more accurate to the actual values compared to those produced in *KNIME*. For example, 81.1% of the actual values 'YES' in the variable *Survived* in *Orange* were predicted correctly, compared to 67.5% in *KNIME*. Note that bias may be present, as the total sum does not match between the two programs.

ACCURACY STATISTICS (TARGET CLASS – YES)

	F1		**Precision**		**Recall**	
Model	***Orange***	***KNIME***	***Orange***	***KNIME***	***Orange***	***KNIME***
Tree	0.702	0.731	0.741	0.741	0.667	0.722
Random Forest	0.759	0.727	0.811	0.790	0.713	0.673
Neural Network	0.854	0.523	0.916	0.600	0.800	0.463
Logistic Regression	0.886	0.732	0.921	0.642	0.853	0.852

Figure 11. The statistics assessed in *Orange Data Miner* and *KNIME* based on instances where *Survived* was predicted as 'YES'. The *F1* statistic is a weighted mean of *Precision* and *Recall*; it is evident that *Orange* has the higher F1 statistic among the models on average. *Precision* informs the user of the proportion of those survived correctly identified as 'survived'; on average, *Orange* has the higher precision statistic. *Recall* is the proportion of passengers who have survived among all those classified as 'survived'; on average, *Orange* distinctly has the higher recall statistic.

ACCURACY STATISTICS (TARGET CLASS – OVERALL)

	Classification Accuracy	
Model	***Orange***	***KNIME***
Tree	0.783	0.780
Random Forest	0.827	0.791
Neural Network	0.895	0.651
Logistic Regression	0.916	0.743

Figure 12. The statistics assessed in *Orange Data Miner* and *KNIME* based on the average of the instances where *Survived* was predicted as 'YES' and 'NO'. *Classification Accuracy* is the proportion of instances correctly classified. Evidently, *Orange* has the higher classification accuracy for all models.

IV A. CRITERION: SUITABILITY

Criterion	**Mark**		**Justification**
	Orange	**KNIME**	
Suitability	3	4	*Orange Data Miner* provides a reasonable number of predictive functions. Although this is suitable for a beginner, their range is comparably limited and does not include highly advanced or complex models. *KNIME* however, provides a more comprehensive range of algorithms.

IV A. CRITERION: INTEROPERABILITY

Criterion	Mark		Justification
	Orange	KNIME	
Interoperability	2	3	Both programs worked satisfactorily in extracting and interpreting data from a CSV file. However, problems were encountered when using *Orange* in detecting the testing data subset. This prevented the testing data to be included when using the prediction algorithms in this analysis. The testing data initially created in *Orange* did not integrate well in *KNIME, as such a separate testing data subset was created.* Numerous errors were also detected when using the data in *KNIME*, most of which were ambiguous and unclear thus, it was challenging finding a solution. However, unlike *Orange*, reasons are given as to why an algorithm is not detected. More analysis is needed to ascertain whether these programs integrate well (if at all) with other external sources.

IV A. CRITERION: LEARNABILITY

Criterion	Mark		Justification
	Orange	KNIME	
Learnability	4	3	*Orange Data Miner* is most suitable for beginners as it is very easy to learn. The program provides numerous learning materials including videos and example workflows, it also provides a description of each function when requested. *KNIME* however takes a lot of time to become familiarised with, likely due to the larger range of functions in its repository. KNIME also provides a description of each function and offers a self-paced program, however, the examples shown online are not comprehensive.

IV A. CRITERION: ATTRACTIVENESS

Criterion	Mark		Justification
	Orange	KNIME	
Attractiveness	5	4	The physical layout in *Orange Data Miner* is clean, interactive and colourful; the icons of each function and the description on each node is particularly helpful. Each function is categorised into *Data*, *Visualise*, *Model*, *Evaluate* and *Unsupervised*, which makes it

			easy for users to find everything. While the layout in KNIME is also colourful and interactive, it is overcrowded. Users also have to search for a function is the node repository.

IV A. CRITERION: FUNCTIONALITY

Criterion	Mark		Justification
	Orange	KNIME	
Functionality	4	3	Both programs utilise a drag and drop graphical user interface (GUI). However, *KNIME* proved to take more time in performing the same tasks, as unlike *Orange Data Miner*, it does not combine the scoring function and allow tasks to be performed simultaneously.

IV A. CRITERION: INTEPRETATION OF OUTPUT

Criterion	Mark		Justification
	Orange	KNIME	
Interpretation of Output	3.5	3	Both provides provide the user with confusion matrices and accuracy scores, which are satisfactory. However, a limitation in KNIME is that it does not present its confusion matrices in percentages, which reduces the efficiency of communicating its results. As such the percentage confusion matrix for *KNIME* was calculated manually based on the number of instances. Unlike *Orange*, *KNIME* does not provide its results in a single function – an individual scorer has to be employed for each model. This makes it particularly difficult for visualisation.

IV A. CRITERION: INSTALLABILITY

Criterion	Mark		Justification
	Orange	KNIME	
Installability	4	4	Both software programs install easily on iOS. No additional programs are required to be downloaded for the program to run effectively.

IV A. CRITERION: ACCESSIBILITY

Criterion	Mark		Justification
	Orange	KNIME	
Accessibility	3	3	*Orange Data Miner* runs on Microsoft Windows, macOS and Linux, while *KNIME*

			runs on Microsoft Windows, iOS and Linux. Both are free and open-source programs.

V. RECOMMENDATION

It is advised that the preferred software program for the client to adopt in future is *Orange Data Miner* with a total score of 32.5 out of 45 based on the criteria outlined in this report. The program only just outperforms *KNIME* with a score of 30 out of 45, particularly due to its strengths in:

- Accuracy – particularly with a dataset size comparable to that of titanic
- Learnability – the abundance of learning resources free and available
- Attractiveness – clean, simple and colourful layout; not overcrowded
- Functionality – drag and click GUI allows multiple functions to be executed at once
- Interpretation of output – thorough and results from all models shown in one place
-

However, a significant limitation is that the analyses were not performed on a range of data of varying sizes and as mentioned, certain algorithms work best with data of considerable size. Hence, it is possible, KNIME may be more beneficial for larger datasets. It is also important to note that these results may have bias, as it is evident from the confusion matrices that values may not be considered and that the actual value of 'YES' and 'NO' in *Survived* is not equal i.e., are skewed.