

Introduction to Data Analytics

Assignment 3

Name:

Student id:

Data mining Problem:

Briefly discuss, what is data mining problem of the given task? Basically what is the business problem do you want to solve?

Input:

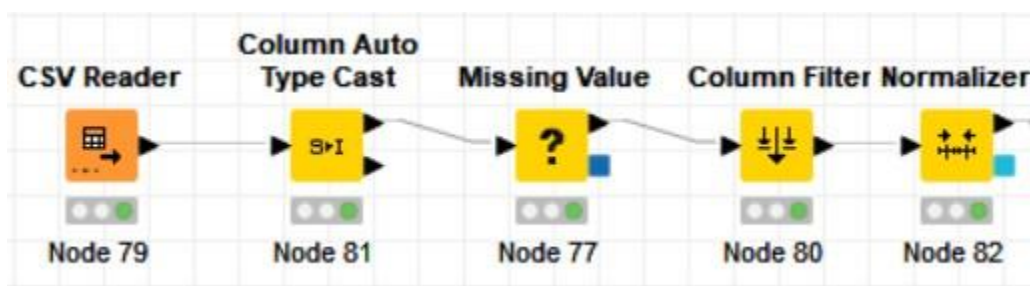
Briefly discuss, What kind of information do you have to solve the data mining problem?

Output:

Briefly discuss, What is the outcome of the problem or what do you have as output?

Data Preparation (e.g: Data Pre-processing and Data Exploration)

Mention about the data pre-processing and data exploration if you have done any of these and keep the screenshot of the used node (KNIME) or function/library (python) such as below (just an example):



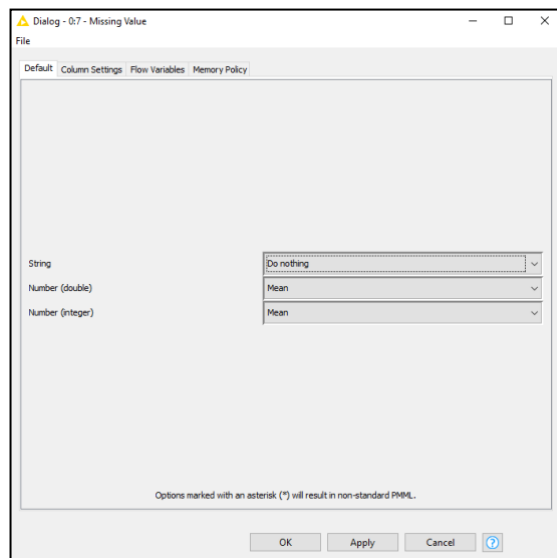
Missing values: (Just an example)

Table X shows the percentage missing values greater than 35% that exist for 4 attributes, namely Cloud9am, Cloud3pm, Evaporation and Sunshine.

Table X: Attributes missing large amounts of data (>35%)

Attribute	% of rows without data
Evaporation	42.7%

Sunshine	47.5%
Cloud9am	37.8%
Cloud3pm	40.2%

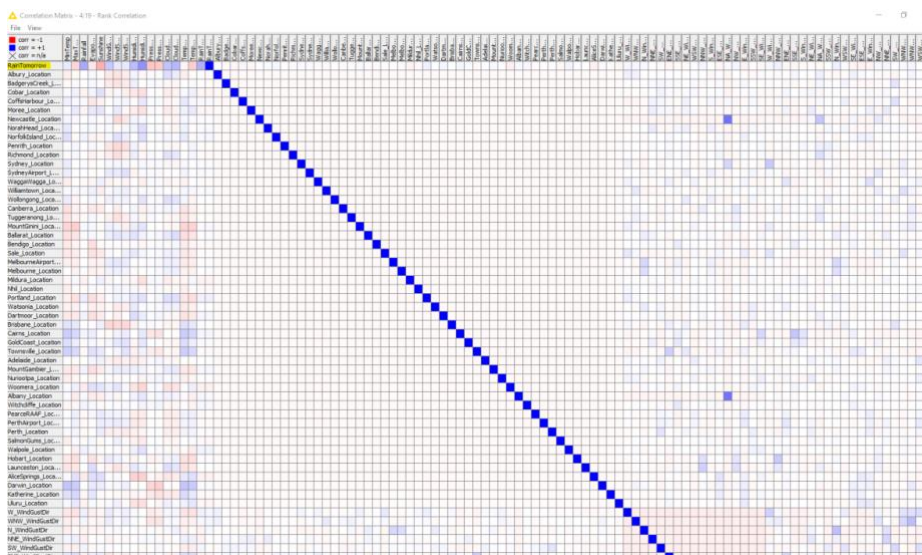


Normalization:

Data transformation:

Linear/rank correlation:

Briefly discuss about it and how did you apply it for building the models.



Others:

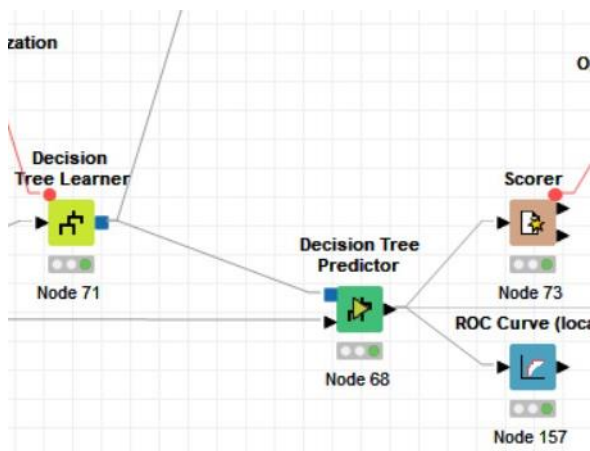
How you went about solving the problem

Talk a bit of parameter optimization and feature selection if you have applied and how did it help to build the better classifiers. Discuss, how you handle the unbalanced data issues, whether it helped you to solved the problem if you faced during building the classifiers. Also, talk on your partitioning strategy, how did you split your dataset whether it is normal partition or use cross validation. keep the screenshots the method as node (KNIME) or function/library (python).

Building Models

Decision Trees (DT)


Briefly discuss about your approach for building DT classifier and keep the screenshot the DT workflow (KNIME) or function/library (Python). (just an example)



Result table (just an example)

	Accuracy	Recall

Or (just an example)



Confusion Matrix - 0:196 - Scorer

File

Hilite

RainTomor...	0	1	
0	32376	6176	
1	3866	7340	

Correct classified: 39,716

Wrong classified: 10,042

Accuracy: 79.818 %

Error: 20.182 %

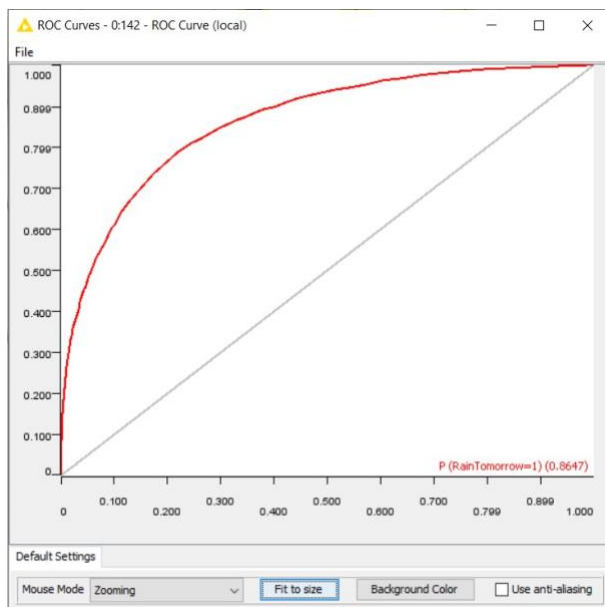
Cohen's kappa (κ) 0.461

Figure:



Table:

ROC (Just an example)



Describe briefly about the table and plot. Also talk about the parameter optimization if you have done any and put the results in above table. Keep the screenshot of parameter setting.....

KNN

Follow the same structure as above (DT)

Random Forest (RF)

Follow the same structure as above

Ensemble methods

SVM

NN

Other classifiers.....

Best classifier

It's type, its performance, how it solved the problem (if it makes sense for that type of classifier), and reasons for selecting it;. You may considering to **Keep the results of all classifiers in a table and briefly discuss the results.** (just an example)

	Accuracy	Recall	F1 Score	
DT							
KNN							
RF							
....							
....							

Kaggle submission

Keep the Kaggle score of your best model with brief description. Also, put the screenshot of the Kaggle score along with the position in Kaggle competition.

Reference

Keep the reference if there is any.

Appendix

Keep the screenshot of the KNIME workflow or Python code