Cover sheet for submission of work for assessment

SWINBURNE UNIVERSITY OF TECHNOLOGY

UNIT DETAILS

Unit name	Data Science Principles			Class day/time	Fridays
Unit code	COS10022	Assignment no.	2	Due date	10.11.2024
Name of lecturer/teacher	Mr. Hoang Anh Minh				
Tutor/marker's name					

Office use only

Faculty or school date stamp

STUDENT(S)

	Family Name(s)	Given Name(s)	Student ID Number(s)
(1)	Tran	Thien Thao Vy	104991221
(2)			
(3)			
(4)			
(5)			
(6)			

DECLARATION AND STATEMENT OF AUTHORSHIP

1. I/we have not impersonated, or allowed myself/ourselves to be impersonated by any person for the purposes of this assessment.
2. This assessment is my/our original work and no part of it has been copied from any other source except where due acknowledgement is made.
3. No part of this assessment has been written for me/us by any other person except where such collaboration has been authorised by the lecturer/teacher concerned.
4. I/we have not previously submitted this work for this or any other course/unit.
5. I/we give permission for my/our assessment response to be reproduced, communicated, compared and archived for plagiarism detection, benchmarking or educational purposes.

I/we understand that:

6. Plagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a form of cheating and is a very serious academic offence that may lead to exclusion from the University. Plagiarised material can be drawn from, and presented in, written, graphic and visual form, including electronic data and oral presentations. Plagiarism occurs when the origin of the material used is not appropriately cited.

Student signature/s

I/we declare that I/we have read and understood the declaration and statement of authorship.

(1)		(4)	
(2)		(5)	
(3)		(6)	

Further information relating to the penalties for plagiarism, which range from a formal caution to expulsion from the University is contained on the Current Students website at **www.swin.edu.au/student/**

Copies of this form can be downloaded from the Student Forms web page at **www.swinburne.edu.au/studentforms/**

Swinburne University of Technology Hawthorn Campus
Dept. of Computing Technologies

COS10022 Data Science Principles

Assignment 2 - *Semester 1, 2024*

I. Introduction

This project focuses on analyzing a dataset with 100,000 records, each representing different financial credit score classes, categorized into three distinct groups. The dataset consists of 24 attributes. The initial goal is to clean and preprocess the data to prepare it for further analysis, addressing issues such as missing values, outliers, and data normalization. This step is crucial to ensure that the data is consistent, reliable, and suitable for model building. Once the data is cleaned, the next objective is to develop two predictive models aimed at forecasting the "Credit_Score" class. During the modeling phase, feature selection and evaluation of different algorithms will take place to choose the most appropriate models for accurate predictions. Throughout the project, some exploration of prebuilt tools and libraries will be carried out to streamline the process and ensure effective implementation of the models.

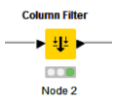
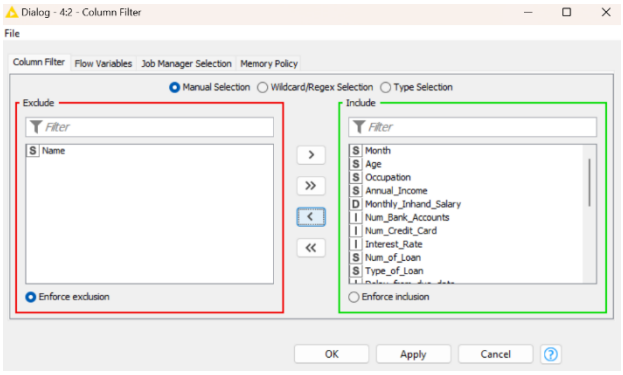
II. Assignment Task

1. Follow the instructions to clean the data and answer questions. If any of the nodes you used in the workflow has a random seed, set **9214** to the seed to fix the random state. **[65 marks in total]**

   1) Our goal is to predict the credit score from the given data. There is/are one (or multiple) attribute(s) which is/are significantly irrelevant to the goal. Pick the most irrelevant attribute and give a persuasive rationale for that. The excluded attribute(s) is _____, and the reason for removing it is _____. **[2.5 marks]**

   Ans:

   - **The excluded attribute is "Name". We can remove them by using Column Filter.**
   - **The reason for removing it is because that in the scenario of predicting dataset, employing name or any personal data can cause bias information, affect privacy also. Moreover, in the context of predicting credit score, there are more noteworthy attributes which can be payments, payments history or their paying behaviours, name is simply unnecessary.**

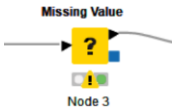
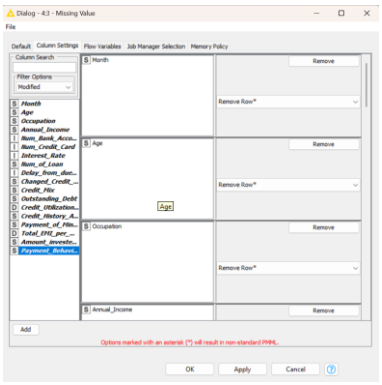
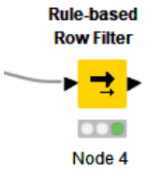
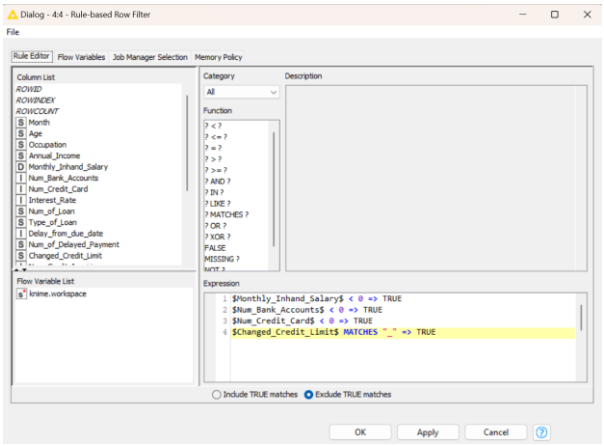
Sequence	Node	Command
1		

   2) After removing the selected attribute(s), let's start to remove tuples containing missing values. Remove tuples only if any of the attributes listed below have missing values: "Month," "Age," "Occupation," "Annual_Income," "Num_Bank_Accounts," "Num_Credit_Card," "Interest_Rate," "Num_of_Loan," "Delay_from_due_date," "Changed_Credit_Limit," "Credit_Mix," "Outstanding_debt," "Credit_Utilization_Ratio," "Credit_History_Age," "Payment_of_Min_Amount," "Total_EMI_per_month," "Amount_invested_monthly," and "Payment_Behaviour." Moreover, some tuples with infeasible values in the attributes, such as "Monthly_Inhand_Salary" < 0,

"Num_Bank_Accounts" < 0, "Num_Credit_Card" < 0, and "Changed_Credit_Limit" contains "_", should also be removed. List the node(s) (in sequence) and the corresponding command(s) used in this process. **[5 marks]**

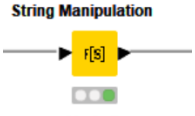
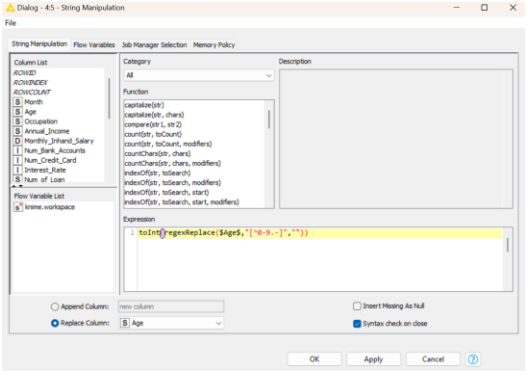
Ans:

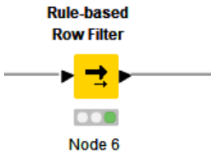
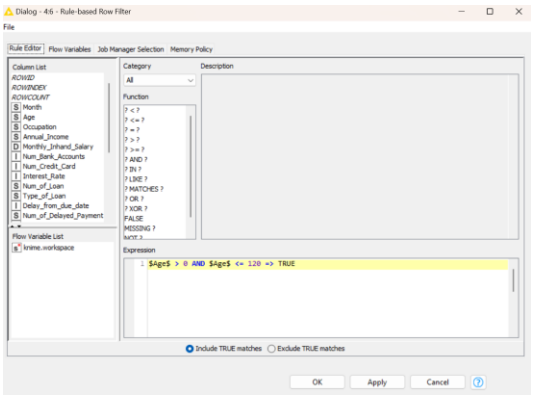
- **(1) List of tuples to remove are: "Month", "Age", "Occupation", "Annual_Income", "Num_Bank_Accounts", "Num_Credit_Card", "Interest_Rate", "Num_of_Loan", "Delay_from_due_date", "Changed_Credit_Limit", "Credit_Mix", "Outstanding_debt", "Credit_Utilization_Ratio", "Credit_History_Age", "Payment_of_Min_Amount", by using Missing Value node.**
- **(2) After that, we can use node "Rule-based Row Filter" to narrow the rules for attributes to prevent infeasible values. The command expression is listed in sequence 2.**

Sequence	Node	Command
1		
2		

3) Check for the "Age" attribute to eliminate symbols that are not numbers to recover the data into the usual number format. Moreover, drop the tuples whose "Age" value is lower than or equal to 0 or greater than 120. List the node(s) (in sequence) and the corresponding command(s) used in this process. **[5 marks]**

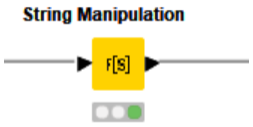
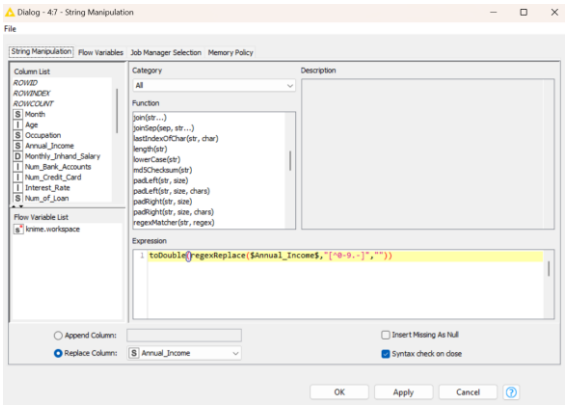
Ans:

Sequence	Node	Command
1		

2		

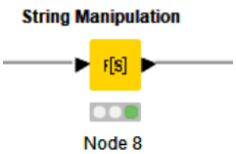
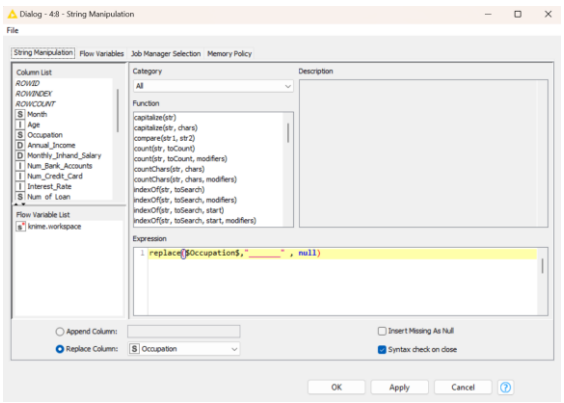
4) Remove the non-numerical symbol in the "Annual_Income" column and convert it to the double format. List the node(s) (in sequence) and the corresponding command(s) used in this process. **[5 marks]**

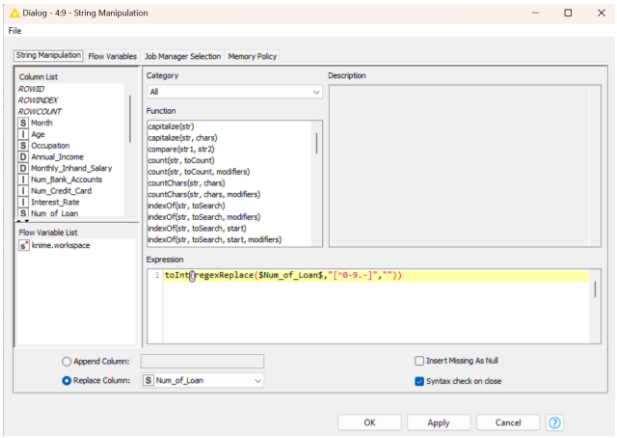
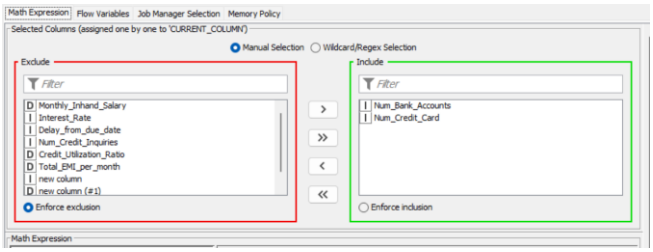
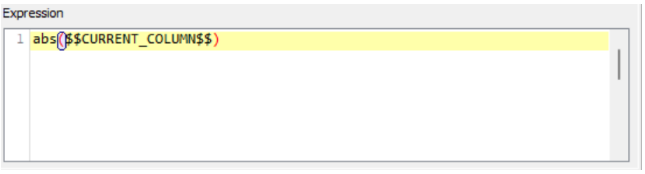
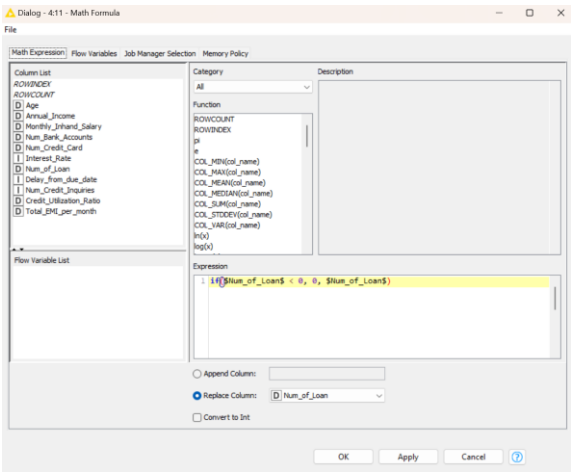
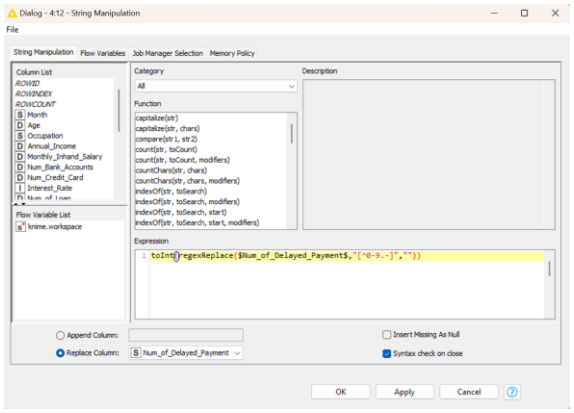
Ans: We use regular expressions, listed in the command, in order to form a pattern for non-numerical data. Pattern brackets with caret are used to find arange of characters that is not in range.

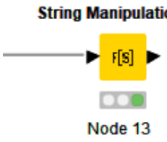
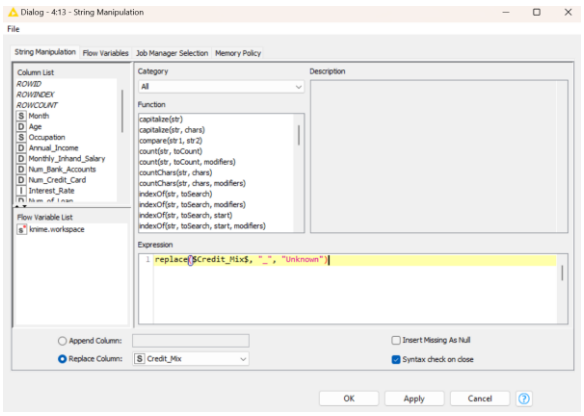
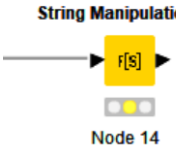
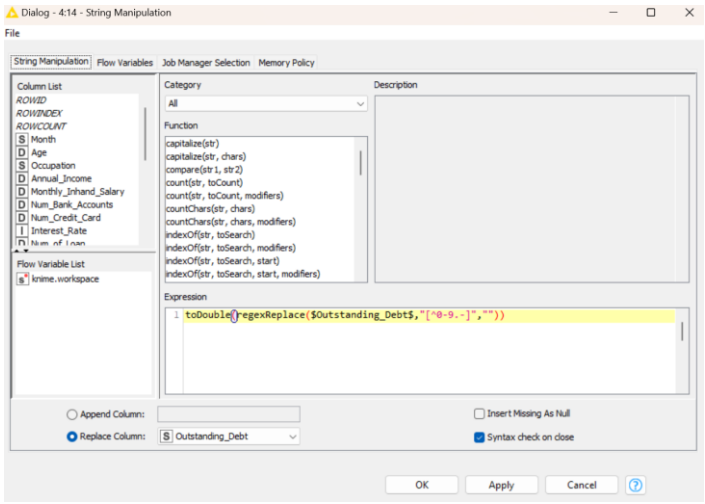
Sequence	Node	Command
1		

5) Convert the "_____" in the "Occupation" attribute to Null. Please note that Null is different from an empty string. Remove the non-numerical symbol in "Num_of_Loan" and convert it to integer data type. Take absolute values of attributes "Num_Bank_Accounts" and "Num_Credit_Card." Set values to 0 for the "Num_of_Loan" attribute if the original values are negative. Remove the non-numerical symbol in "Num_of_Delayed_payment" and convert it into integer format. Set the "Credit_Mix" value to "Unknow" if the original value is "_".Remove the non-numerical symbol in "Outstanding_Debt" and convert it into the double format. List the node(s) (in sequence) and the corresponding command(s) used in this process. **[10 marks]**

Ans:

Sequence	Node	Command
1		

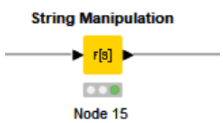
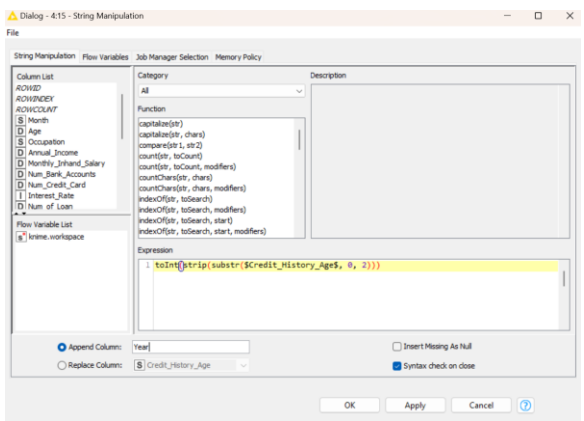
2	String Manipulation Node 9	
3	Math Formula (Multi Column) Node 10	 
4	Math Formula Node 11	
5	String Manipulation Node 12	

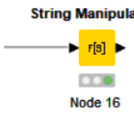
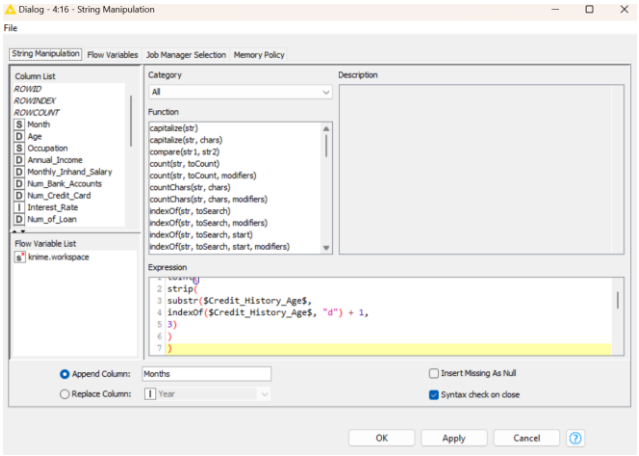
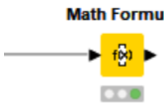
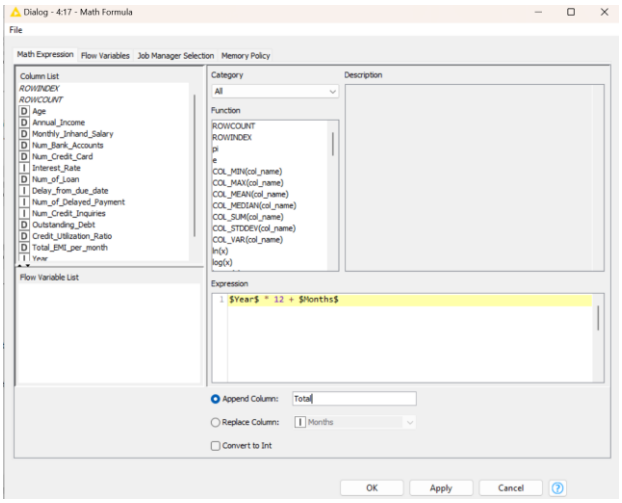
6		
7		

(1): Convert the "_____" in the "Occupation" attribute to Null.
(2): Remove the non-numerical symbol in "Num_of_Loan" and convert it to integer data type.
(3): Take absolute values of attributes "Num_Bank_Accounts" and "Num_Credit_Card."
(4): Set values to 0 for the "Num_of_Loan" attribute if the original values are negative.
(5): Remove the non-numerical symbol in "Num_of_Delayed_payment" and convert it into integer format.
(6): Set the "Credit_Mix" value to "Unknow" if the original value is "_".
(7): Remove the non-numerical symbol in "Outstanding_Debt" and convert it into the double format.

6) Convert the "Credit_History_Age" to the count of months and store it in the integer format. For example, if the original value from a tuple is "22 Years and 1 Months", the value will be 265 after the conversion (22 * 12 + 1 = 265). Store the converted result in a new attribute called "Total_CHA." List the node(s) (in sequence) and the corresponding command(s) used in this process. **[10 marks]**

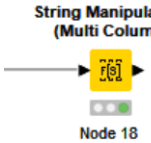
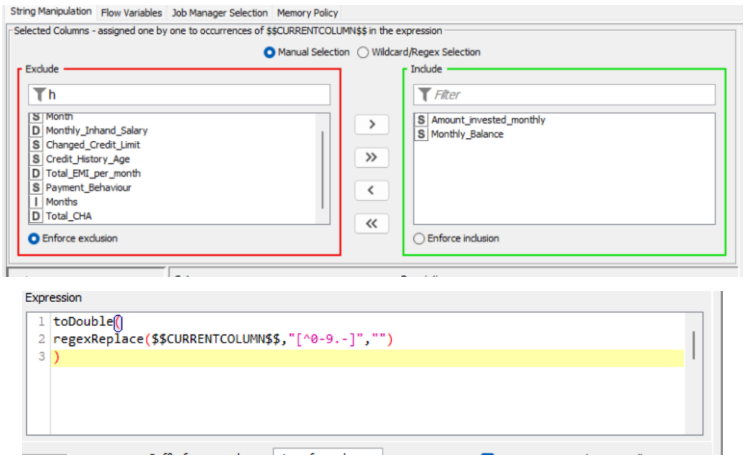
Ans:

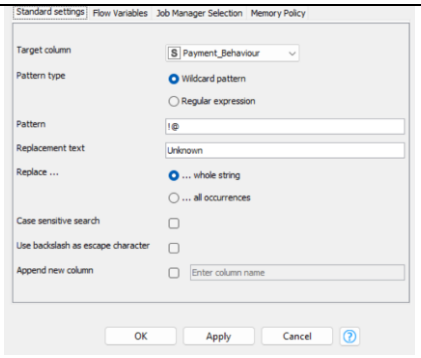
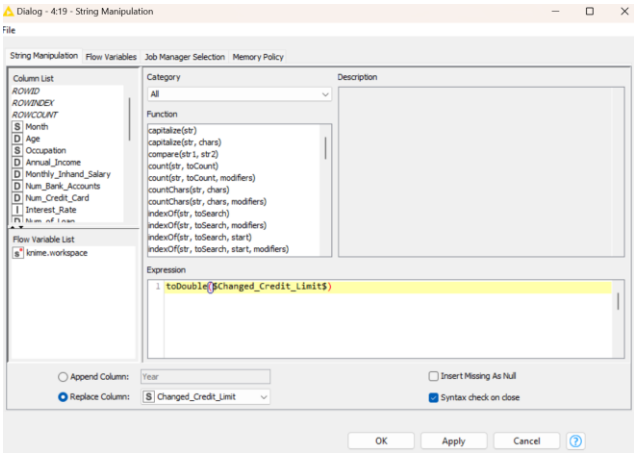
Sequence	Node	Command
1		

2		
3		

7) Remove the non-numerical symbol in “Amount_invested_monthly” and convert it to the double format. Set the value to “Unknow” if the original value in “Payment_Behaviour” attribute starts with “!@”. Remove the non-numerical symbol in “Monthly_Balance” and convert it to the double format. Convert “Changed_Credit_Limit” into the double format. List the node(s) (in sequence) and the corresponding command(s) used in this process. **[5 marks]**

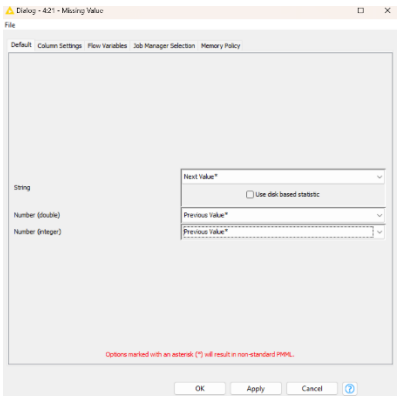
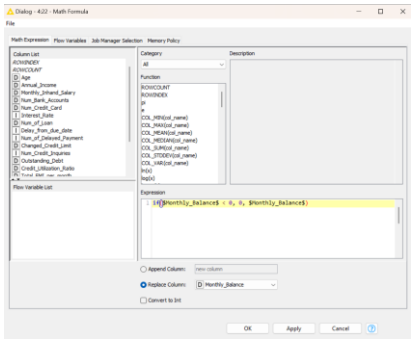
Ans:

Sequence	Node	Command
1		
2		

	String Replacer Node 20	
3	String Manipulation Node 19	

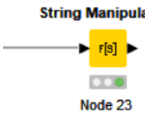
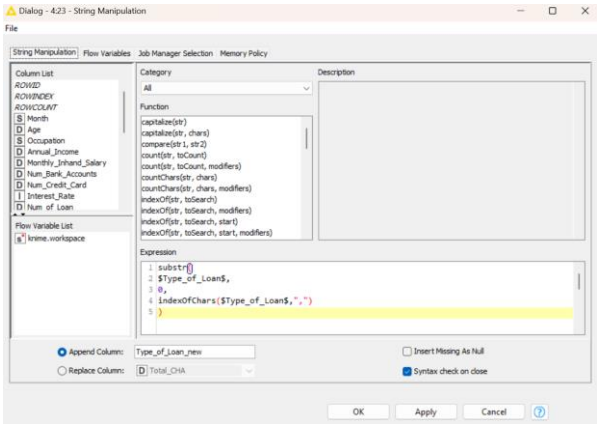
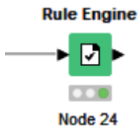
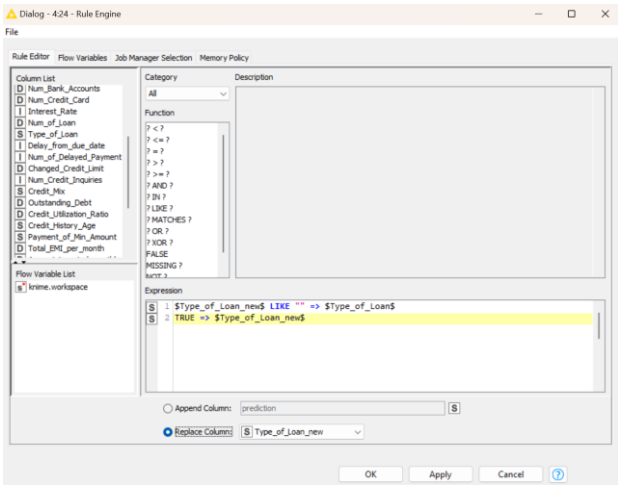
8) Use the “Missing Value” node and use the “Next Value*” to replace missing values in all string type attributes. Use the “Previous Value*” in the same node to replace missing values in any numerical format. If the value of “Monthly_Balance” is negative, replace the value with 0. Screenshot the pop-up window with the correct settings. **[5 marks]**

Ans:

Sequence	Node	Command
1	Missing Value Node 21	
2	Math Formula Node 22	

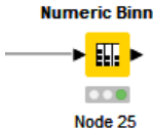
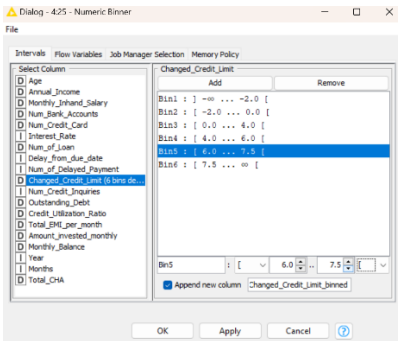
9) Simplify the “Type_of_Loan” attribute. If the original content has more than one type separated by a comma, keep only the first part. Otherwise, keep the full description if there is no comma included. For example, “Auto Loan, Credit-Builder Loan, Personal Loan, and Home Equity Loan” will become “Auto Loan”, “Credit-Builder Loan” will still be “Credit-Builder Loan”, and “Not Specified, Auto Loan, and Student Loan” will become “Not Specified” after the process. List the node(s) (in sequence) and the corresponding command(s) used in this process. **[10 marks]**

Ans:

Sequence	Node	Command
1		
2		

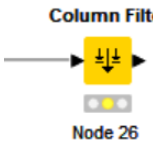
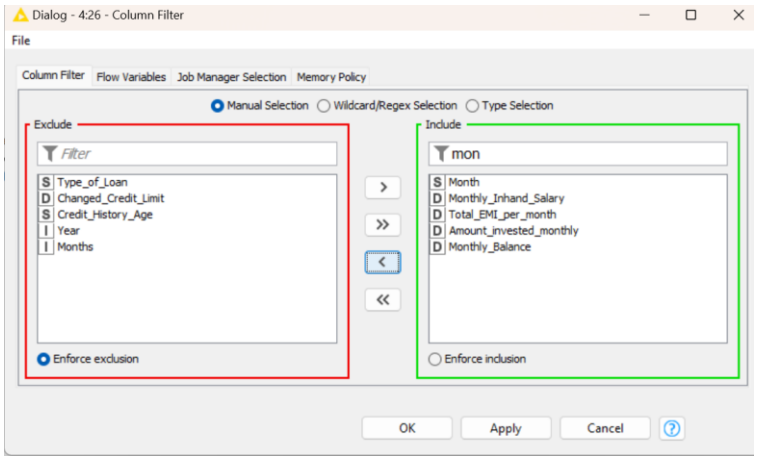
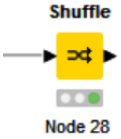
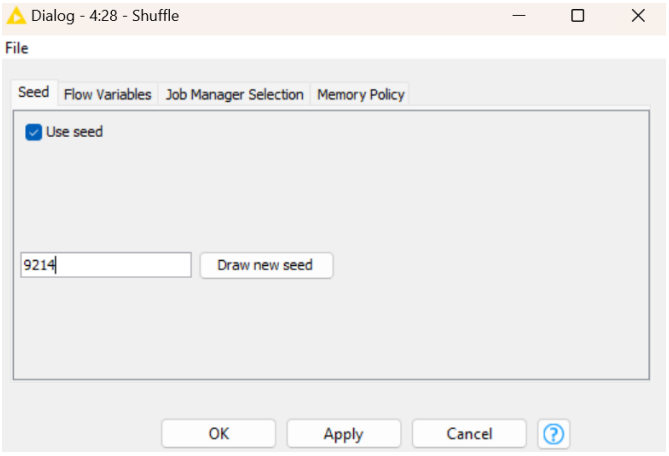
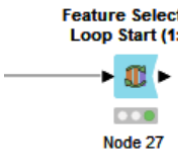
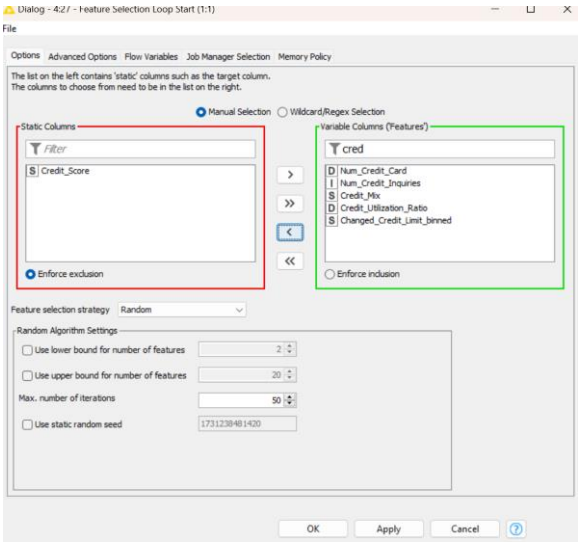
10) Bin the “Changed_Credit_Limit” attribute with six bins of ranges: $[-\infty, -2.0)$, $[-2.0, 0)$, $[0, 4.0)$, $[4.0, 6.0)$, $[6.0, 7.5)$, and $[7.5, \infty)$ and put the result into a new attribute called “Changed_Credit_Limit_binned”. Screenshot the pop-up window with the correct settings of your binner. **[5 marks]**

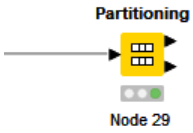
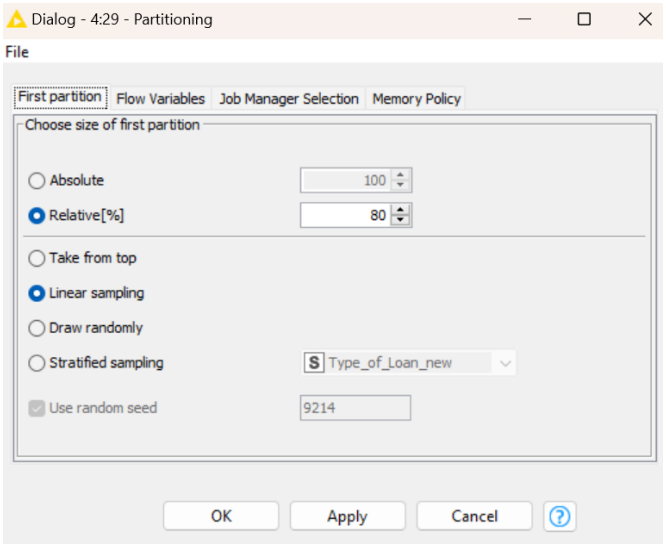
Ans:

Sequence	Node	Command
1		

11) Remove all temporarily created or useless attributes. Use the "Feature Selection Loop Start (1:1)" node to select the feature. The class label should be excluded from the features in the feature selection node. The Genetic Algorithm is specified to be the feature selection strategy with default population size and the maximum number of generations. Again, **9214** should be used as the static random seed. After selecting features, shuffle the data with seed **9214**. The data should be partitioned by "Linear sampling", with 80% data in the training set and 20% in the test set. How many tuples and attributes (excluding the class label) are in the training set at the end? **[5 marks]**

Ans:

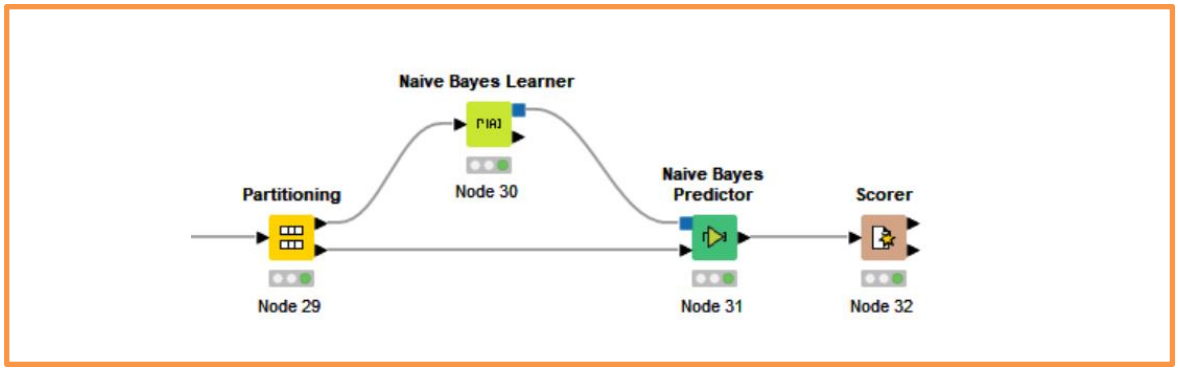
Sequence	Node	Command
1		
3		
2		

4		

2. Build a Naïve Bayes classifier using the training and test sets created in the previous task. Answer the following questions after completing the model training and test. **[15 marks in total]**

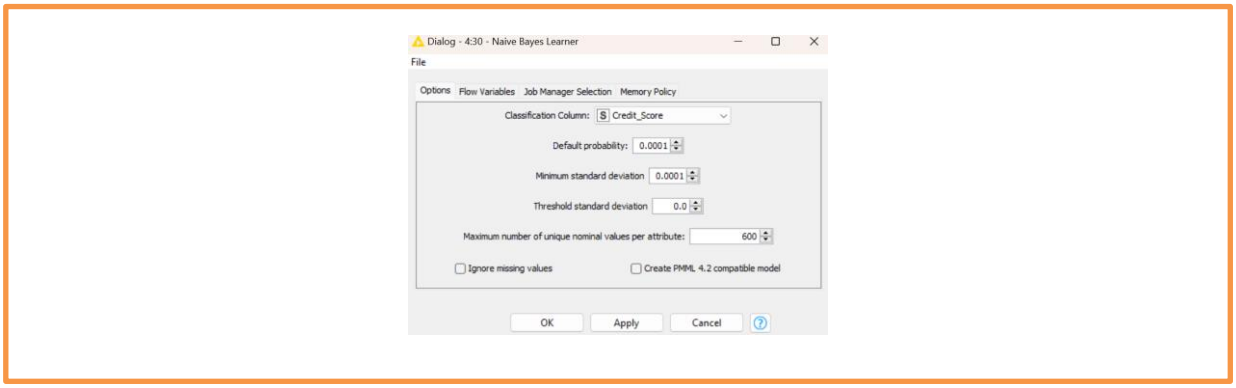
1) Give a screenshot of the Naïve Bayes classifier in the KNIME workflow. You can take the screenshot starting from the portioning node output to the end of the Naïve Bayes classifier part scorer. **[2.5 marks]**

Ans:



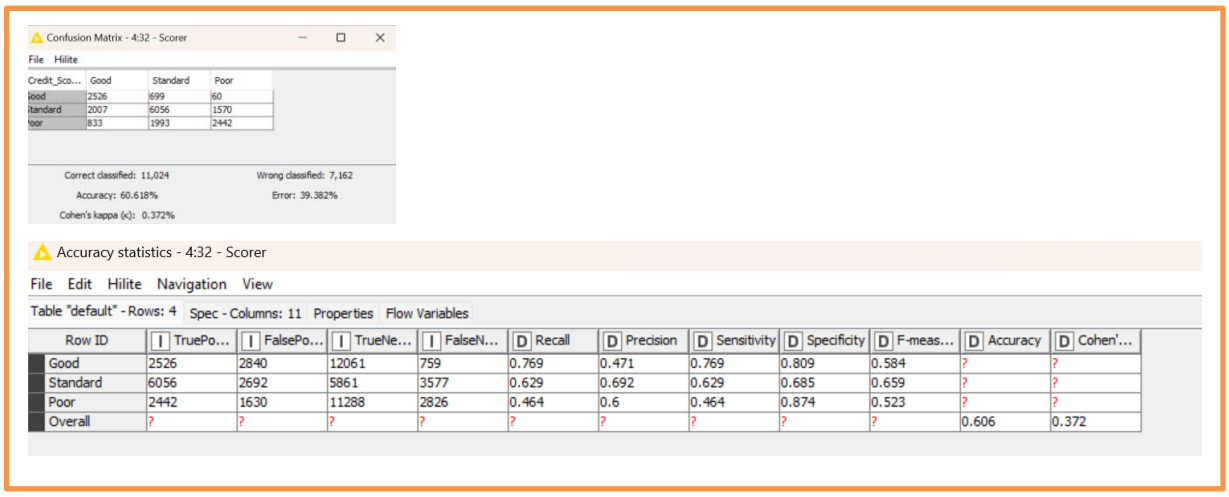
2) The default probability should be 0.0001, the minimum standard deviation is 0.0001, the threshold standard deviation is 0, and the maximum number of unique nominal values per attribute should be set to 600 in the classifier. Screenshot the setting dialogue of your Naïve Bayes Learner. **[2.5 marks]**

Ans:



3) Screenshot the confusion matrix and the Accuracy statistics of the test result. If the bank wants to minimise the risk of lending money to customers, the “Good” in “Credit_Score” should be the major target. Based on the current result, does the classifier perform satisfactorily? **[5 marks]**

Ans:

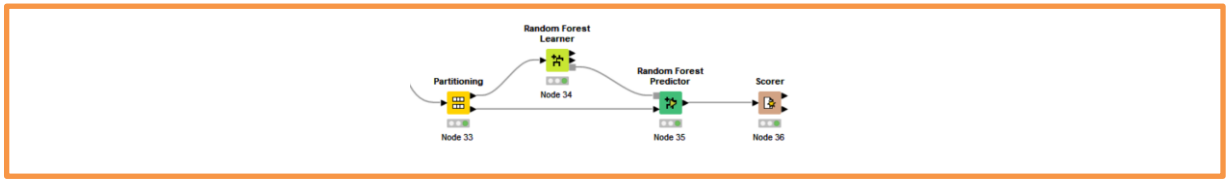


Explanation: From the precision of “Good” in Accuracy Statistics, which is only 0.471, we can see that the Naïve Bayes model does not perform satisfactorily.

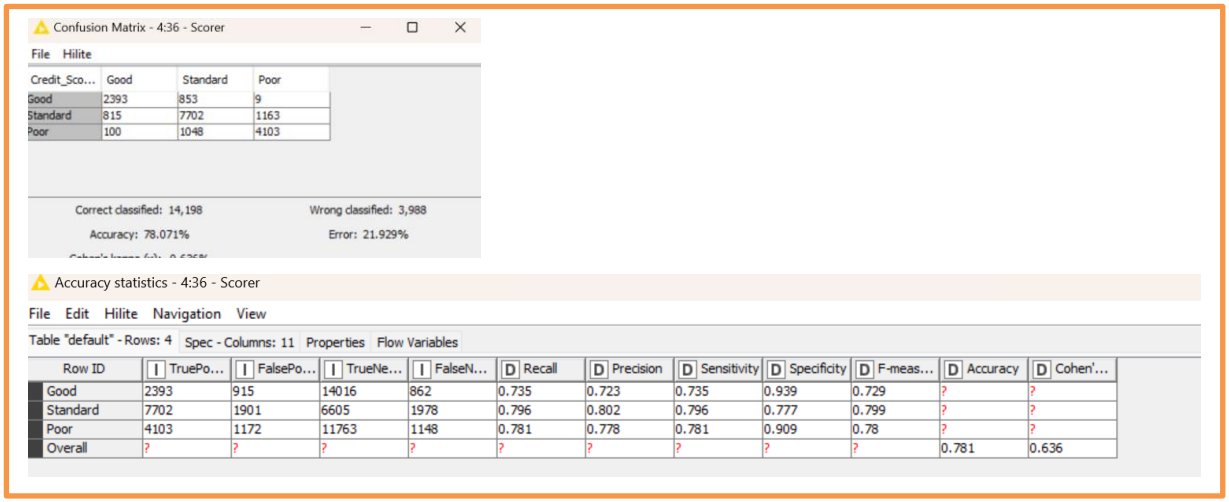
4) Which measurement should we look at to interpret your conclusion in this case? **[5 marks]**
Ans: The best metric to evaluate the model is the precision of the "Good", also known as positive predictive value as precision measures the percentage of correct positive predictions out of all the positive predictions made by the model. In this case, the bank wants to minimize the risk of lending money to customers who might default on their payments. Therefore, it is crucial for the bank to accurately identify "Good" clients who are less likely to default.

3. Build a random forest classifier using the training and test sets created in the previous task. Answer the following questions after completing the model training and test. Use the information gain ratio as the split criterion and **9214** as the static random seed to build the random forest model. **[15 marks in total]**

1) Give a screenshot of the random forest classifier in the KNIME workflow. You can take the screenshot starting from the portioning node output to the end of the Naïve Bayes classifier part scorer. **[2.5 marks]**
Ans:



2) Screenshot the confusion matrix and the Accuracy statistics of the test result. **[2.5 marks]**
Ans:



3) If the bank wants to minimise the risk of lending money to customers, the “Good” in “Credit_Score” should be the major target. Compare the measurements between random forest results and Naïve Bayes results. Which model presents a more suitable result? Which measure should be used to make the comparison? **[5 marks]**

Ans:

As mentioned earlier, we can use the precision of the "Good" class to compare the two models in this case. By focusing on precision for the "Good" class, the bank can reduce the risk of defaults, ensuring that customers identified as "Good" are truly reliable borrowers.

The Accuracy Statistics of two models show that the precision of “Good” in Naïve Bayes is 0.471, and in Random Forest is 0.723. It also shows that the overall Accuracy of Naïve Bayes is 0.606, while the overall Accuracy of Random Forest model is 0.781.

Giving the 2 models’outputs, Random Forest is better and more suitable regarding the large dataset like this.

4) Which class does the built random forest model perform the best? What measurement(s) should we look at to find the answer? **[5 marks]**

Ans:

From the Accuracy Statistics of Random Forest model, we can see that in terms of different metrics like Precision, Recall or F-mesure, Standard class performs the best.

As mentioned, we should look at different mesurements like Precision, Recall or F-measure because these metrics show the percentage of rightness within the dataset. For example, precision tells us the proportion of correct positive predictions made for each class. For each class, it measures how many of the predicted positive cases (e.g., predicting a "Good" borrower) are truly correct. If precision is highest for a particular class, it indicates that the model is best at correctly identifying that class. Recall also shows how many actual positive cases of each class were correctly identified by the model. A high recall means the model is good at detecting all the instances of that class. And F-measure is the harmonic mean of precision and recall, offering a balanced view of both metrics.

---------- End of Submission ----------