Module Code: CSMDS
Assignment report Title: Data Exploratory Analysis for Heart Disease Dataset
Student Number: 32824514
Actual hrs spent for the assignment: 12
Which Artificial Intelligence tools used: None

1. Discussions of Data Assessment and Preprocessing

1.1 Data Characteristics and Quality

UCI heart disease dataset is an open source, which were utilized by many researches as data source. It contains of 920 rows and of 14 attributes (TABLE 1), including ratio attributes such as age, trestbps resting blood pressure, chol (serum cholesterol in mg/dl), thalach (maximum heart rate achieved), and oldpeak (ST depression induced by exercise relative to rest), nominal attributes such as sex, origin, fbs (if fasting blood sugar > 120 mg/dl), and exang (exercise-induced angina), and ordinal attributes such as cp chest pain type, restecg (resting electrocardiographic results), slope (the slope of the peak exercise ST segment) and thal (normal; fixed defect; reversible defect).

This dataset works well for Support Vector Machine (SVM) model with an accuracy level of 85% [1] and even better for Hybrid model (Hybrid of random forest and decision tree with an accuracy level of 88.7%) [2].

Name	Type	Min	Max	Missing Value	Data Type	Methods for Missing Value
age	Integer	28	77	0	Ratio	
sex	String			0	Nominal	
origin	String			0	Nominal	
cp	String			0	Ordinal	
trestbps	Integer	0	200	59	Ratio	Rounded Mean
chol	Integer	0	603	30	Ratio	Rounded Mean
fbs	String			90	Binary	Most frequent
restecg	Integer			2	Ordinal	Most frequent
thalach	Integer	60	202	55	Ratio	Rounded Mean
exang	String			55	Binary	Most frequent
oldpeak	Double	-2.6	6.2	62	Ratio	Mean
slope	String			309	Ordinal	marked as Unknown
ca	Integer	0	3	611	Ratio	Remove
thal	String			486	Ordinal	marked as Unknown
num	Integer	0	4	0	Ratio	

TABLE 1

As shown in TABLE 1, there are some attributes containing missing values, especially attributes slop, ca and thal. Take attribute CA as an example, 299/304 cases in Cleveland got ca records, but only 10/616 cases in other 3 places of study got ca records. For this reason, ca attribute is removed as the highly rate of missing values will bring loud noise. For less bias, missing values in integer attributes such as trestbps, chol and thalach were replaced by the rounded mean values. Missing values in slope and thal attributes were marked as Unknown [3], considering the highly missing values rate could significantly affects performance of other values. Other missing values were replaced by most frequent values.

In details, for the classification column num, which means no heart disease or the stages of heart disease, it has five categories and the percentages of categories are vary from nearly 45% to 3% (FIGURE 1 and FIGURE 3). If you consider using binary classification models, healthy and non-healthy, or Logical Regression, this distribution is fine. But if you use it for multi-classification models, the data of serials stages of 2, 3, 4 should be incremental processed. Additionally, ages are close to standard distribution for different stages of the disease (FIGURE 2), and the figure show that as age increases, the risk of heart disease increases and the degree of verification becomes more serious. This could be true, but since the data is not statistically significant, it may not be suitable globally.

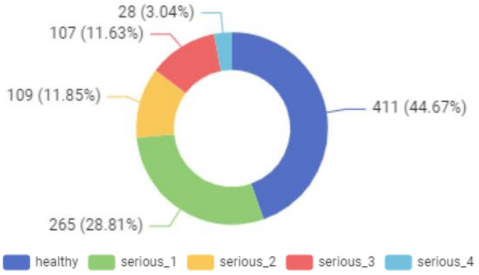


FIGURE 1

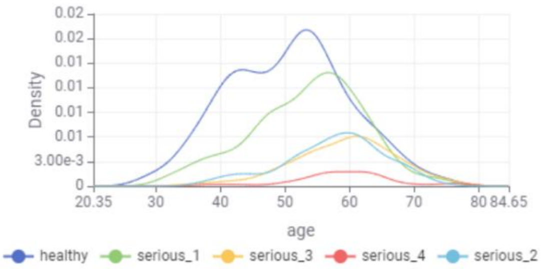
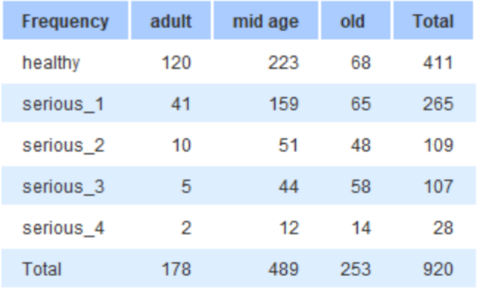


FIGURE 2



Frequency	adult	mid age	old	Total
healthy	120	223	68	411
serious_1	41	159	65	265
serious_2	10	51	48	109
serious_3	5	44	58	107
serious_4	2	12	14	28
Total	178	489	253	920

FIGURE 3

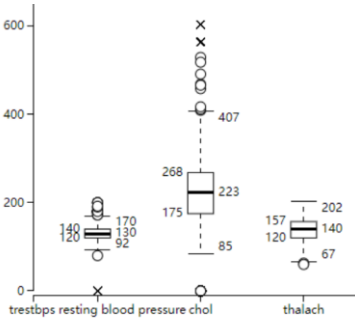


FIGURE 4

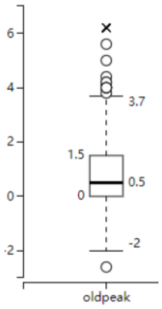


FIGURE 5

For four measurement fields, chol, thalach, oldpeak and trestbps resting blood pressure, FIGURE 4 and FIGURE 5 show the outliers of them and these outliers should be replaced by closest permitted values for less bias. Additionally, FIGURE 6 (all samples)

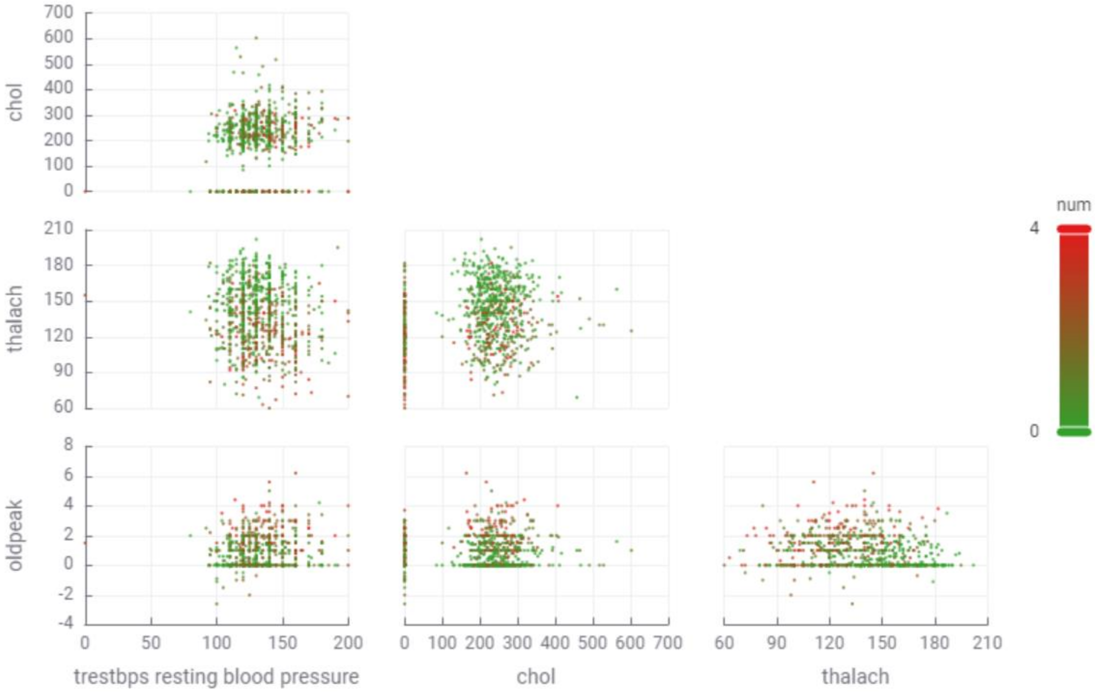


FIGURE 6

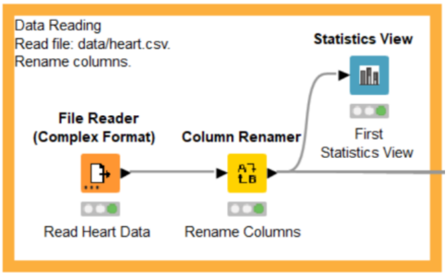
shows the relevance matrix of them, coloring by the severity of heart disease. It can be seen from the figure that there seems to be a positive correlation between oldpeak and the severity of heart disease. Contrarily, thalach shows a negative correlation to the severity of heart disease. Besides, there are not obvious correlations of trestbps resting blood pressure and chol to the severity of heart disease.

1.2 Data Preprocessing

There are 7 main steps for data preprocessing, including data reading, missing values, outliers, category attributes to numbers, normalization, classification number one to many and data output. Considering the number of samples is small (920), to keep the process simple, all operations run serially, not in parallel.

1.2.1 Missing values

As listed in TABLE 1, there are some missing values in several attributes. Attribute CA is removed as there are over 60% missing values. Missing values in other attributes also are processed by the rules discussed in 1.1. Overall, no rows are removed.



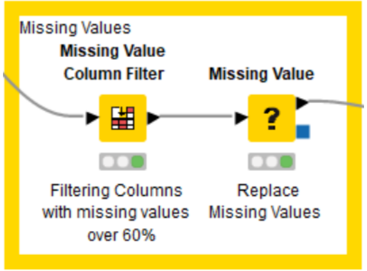


FIGURE 7

FIGURE 8

1.2.2 Category attributes to Numbers

There are some attributes of string format. If we use models such as Decision Tree, string attributes are fine. But if we consider other models such as Logistic Regression and SVM, category attributes should be converted to numbers. It is not matter for different order for nominal attributes, but a customized mapping is more robust. For example, SEX attribute is converted by rules: Female -> 0, Male -> 1. Ordinal attributes are transformed according to specific business scenarios, for instance, SLOPE attribute is converted by rules: down sloping -> 0, flat -> 1, Unknown -> 2 and up sloping -> 3. Intuitively, there is a natural order of down sloping, flat and up sloping. Unknown is at the middle, as it won't be too far away from the others.

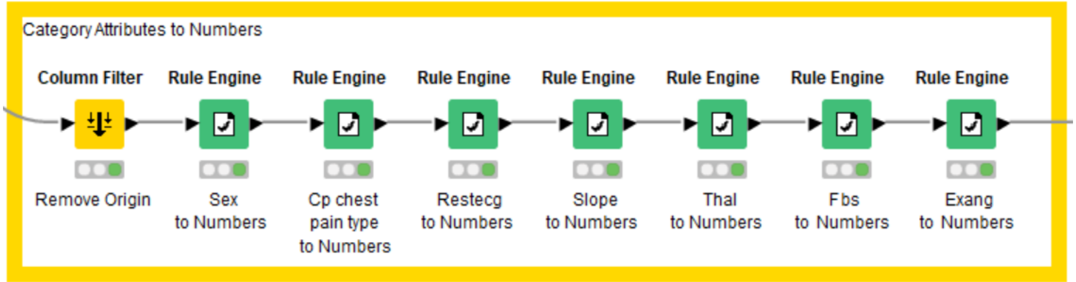


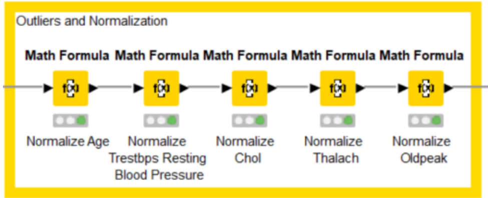
FIGURE 9

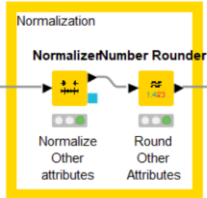
1.2.3 Outliers and Normalization for Ratio Attributes

Outliers could be true or caused by different reasons, but correcting them can improve accuracy and speed up training. Ratio attributes, such as asge, trestbps, chol, thalach and oldpeak, were checked outliers (FIGURE 4 and FIGURE 5) and outliers were replaced by closest permitted values (TABLE 2). Although there are Numeric Outliers nodes in KNIME can provide outlier function, a customized and fixed math formula provided by Math Formula (FIGURE 10) node could make the process more certain and universal for future datasets to be predicted. For this reason, normalizations of these 5 attributes are applied within the Math Formula nodes.

Name	Min	Max	New Range	Math Formula
age	28	77	[20, 80]	max(min((val-20) / 60, 1), 0)
trestbps	0	200	[92, 170]	max(min((val-92) / 78, 1), 0)
chol	0	603	[85, 407]	max(min((val–85) / 322, 1), 0)
thalach	60	202	[67, 202]	max(min((val-67) / 135, 1), 0)
oldpeak	-2.6	6.2	[-2, 3.7]	max(min((val+2) / 5.7, 1), 0)

TABLE 2





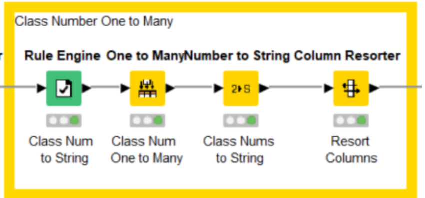


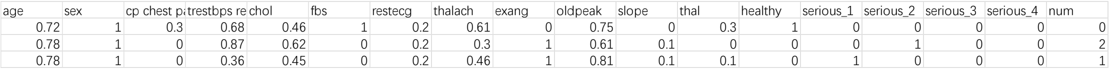
FIGURE 10 FIGURE 11 FIGURE 12

1.2.4 Normalization for other Attributes

For some models such as Logical Regression and SVM, to increase learning speed and to improve the accuracy, all other numerical attributes should be normalized to 0-1 (FIGURE 11). Additionally, all attributes are rounded to 2 decimal places.

1.2.5 Classification Number One to Many

Though the predicted column num is orderly presentation of healthy or the severity of the disease and can be accepted by some of the models, the prediction accuracy of the binary classification model is much higher than that of the milti-classification model. For this reason, column num is applied to be converted to many columns (FIGURE 12). Additionally, columns are resorted (TABLE 3).



age	sex	cp chest pa	trestbps re	chol	fbs	restecg	thalach	exang	oldpeak	slope	thal	healthy	serious_1	serious_2	serious_3	serious_4	num
0.72	1	0.3	0.68	0.46	1	0.2	0.61	0	0.75	0	0.3	1	0	0	0	0	0
0.78	1	0	0.87	0.62	0	0.2	0.3	1	0.61	0.1	0	0	0	1	0	0	2
0.78	1	0	0.36	0.45	0	0.2	0.46	1	0.81	0.1	0.1	0	1	0	0	0	1

TABLE 3

1.2.6 Data Output

Finally, the preprocessed data are written to file data/heart-fine.csv.

2. Conclusion

Overall, the dataset is well collected and labelled based on real cases from 4 places of study. Though data from Cleveland are more complete, through specifically missing value processing, data from other places also work well. The numbers of healthy and non-healthy cases are similar which means it is suitable for a binary classification model. If considering using multi-classes models, more cases of serious stages of heart disease should be collected or generated by incremental technology as serious sample bias affects the accuracy of model training. The data exploratory analysis above shows that some attributes such as age, oldpeak and thalach have an obvious relationship with the severity of heart disease. Additionally, all attributes are normalized and ready for further training. Classification columns could be used by different models.

References

[1] Anderies, A., Tchin, J. A. R. W., Putro, P. H., Darmawan, Y. P., & Gunawan, A. A. S. (2022). Prediction of Heart Disease UCI Dataset Using Machine Learning Algorithms. *Engineering, MAthematics and Computer Science (EMACS) Journal*, *4*(3), 87–93. https://doi.org/10.21512/emacsjournal.v4i3.8683

[2] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021). *Heart Disease Prediction using Hybrid machine Learning Model*. IEEE Xplore. https://doi.org/10.1109/ICICT50816.2021.9358597

[3] Mohammad Alfadli, K., & Omran Almagrabi, A. (2023). Feature-Limited Prediction on the UCI Heart Disease Dataset. *Computers, Materials & Continua*, *74*(3), 5871–5883. https://doi.org/10.32604/cmc.2023.033603