Exploratory Data Analysis and Data Preprocessing using KNIME

AYMANE AACHA

KNIME Workflow and Visualizations

Importing Dataset

To import the data in KNIME is very simple; the **CSV Reader** node is used. Once imported, the **Column Renamer** node is used as the csv file is missing column headers. Finally, the **Color Manager** node is used to color records by origin attribute to improve readability. This workflow can be seen in *Figure 1*. Since there are only 16 columns, renaming them manually is trivial and can be done using the heart_descriptions.txt file provided. Opening the output table view shows a row count of 920 patients.

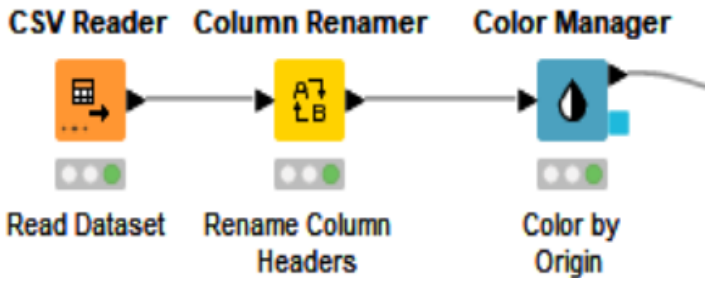

Figure 1:Dataset Import

Missing Values

The first step is to find out if there is missing data present within the dataset. The number of incomplete rows (rows with at least one missing attribute) can be found using a combination of the **Rule-Based Row Filter** and **Group By** nodes. Opening the output shows an incomplete row count of 621 patients; this can be observed in *Figure 2*.

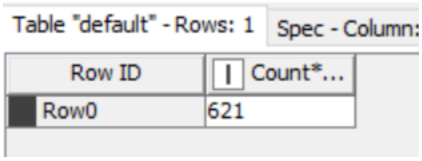

Figure 2: Incomplete Row Count

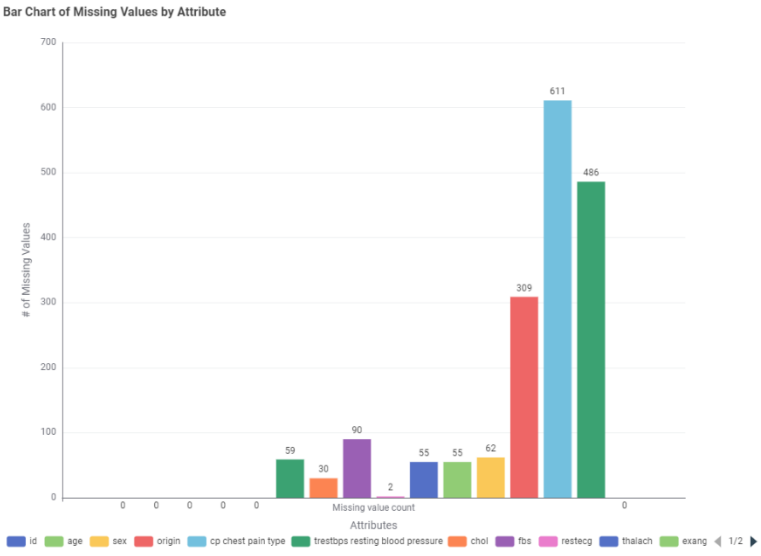

Figure 3: Missing Value Bar Chart

Finding the missing data percentage for each attribute is a slightly harder task but made easy using KNIME. The table is first transposed using the **Table Transposer** node, turning the rows into columns and vice-versa. The **Column Aggregator** node is then used with the aggregation method of 'missing value count,' adding a new column to the dataset with the missing value count of each attribute. After using the **Column Filter** and **Math Formula** nodes, a table with the missing value count and percentage for each attribute is created. The results show three attributes with a significant amount of missing data: slope with 33.6%, ca with 66.4% and thal with 52.8%. The visualization of these results can also be seen in a **Bar Chart** node as seen in *Figure 3*.

For further insights into where the missing values come from, a visualization can be created to illustrate the missing value count by origin value. To do this, a combination of the **Rule Engine**, **Group By**, **Pivot**, and **Bar Chart** nodes were used. The resulting bar chart can be seen in *Figure 4*. 'Cleveland's' records are almost all complete (no missing values), while almost all of 'Hungary's' and 'VA Long Beach's' records have at least one missing attribute. All of 'Switzerland's' records are missing at least one attribute. The workflow with all the previously mentioned nodes regarding missing value insights can be found in the MissingCount.knar file; a screenshot of the workflow can be seen in *Figure 5*.

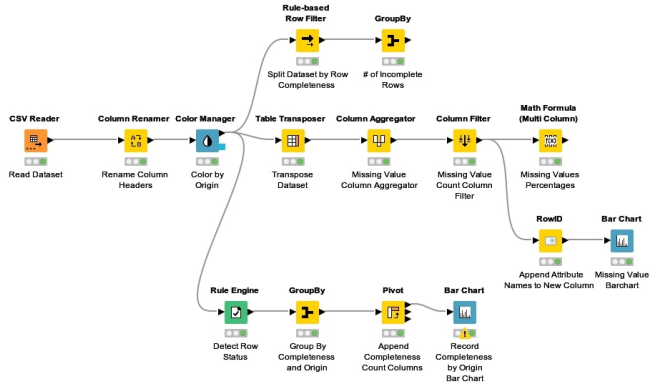

Figure 5: MissingCount.knar Workflow

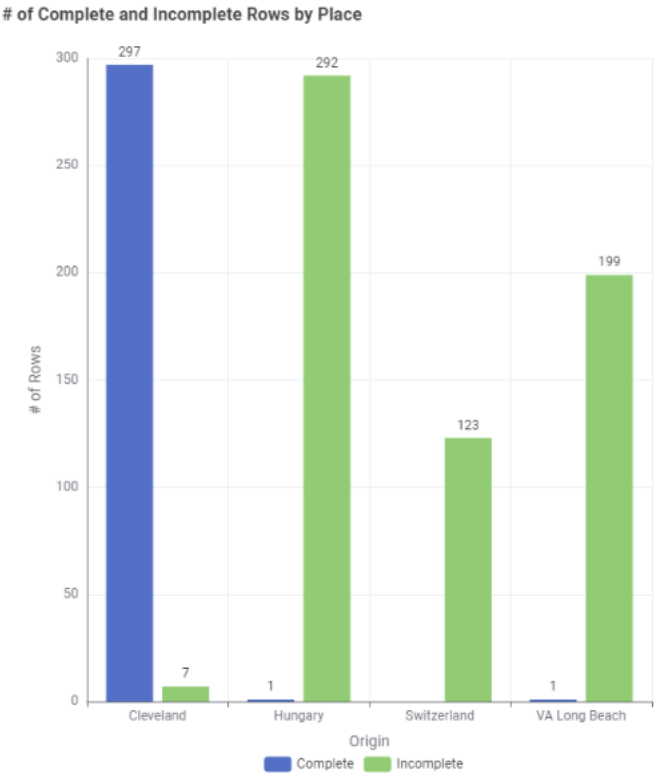

Figure 4: Row Completeness by Place

Due to the medical nature of this dataset, replacing the missing data with another value such as mean, mode or another meaningful statistical value (imputation) could be considered unethical; if this dataset is used to train a heart disease prediction model and the replacement values are wrong, the model could make an incorrect medical prediction. According to Sterne *et al*. (2009), single imputation of missing data "can lead to serious bias." As such, it would be better to remove missing values from the dataset to conserve data integrity. The disadvantage of this approach is that the dataset is shrunk and as such there is less data for analysis, training or any other future application of the dataset.

Next, while an analysis of erroneous/outlier values present could be conducted before removing missing values, it would be redundant as some outlier records in the analysis could be removed during the missing value removal process. Consequently, it would be better to first remove the missing values and conduct an erroneous/outlier value analysis after.

The columns with a high missing value ratio (ca, slope, and thal) can be removed from the dataset using the **Missing Value Column Filter**. After applying this filter, there is still an incomplete row count of 180, obtained using the previous combination of **Rule-Based Row Filter** and **Group By** nodes; while this is a significant number of rows, this is still much less than the previous count of 621. If these rows were removed, it would mean the number of records would decrease by about 20%; a necessary sacrifice to ensure clean data. Subsequently, these incomplete records are deleted

using the **Rule-Based Row Filter** node. After this, there are no more missing values in the dataset and outliers can be searched for and treated as necessary.

Outliers

Using the **Statistics** node provides histogram views for each attribute; chol and trestbps resting blood pressure are two attributes with clear lower bound outliers. The histograms for these attributes, made using the **Column Splitter** and **Histogram** nodes, can be seen in *Figures 6* and *7* respectively.

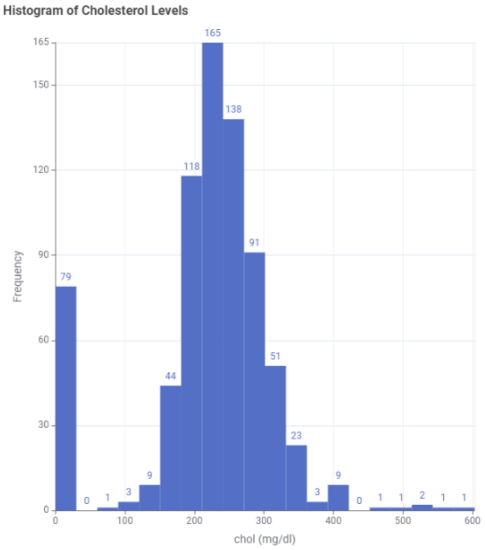

Figure 5: Cholesterol Levels Histogram

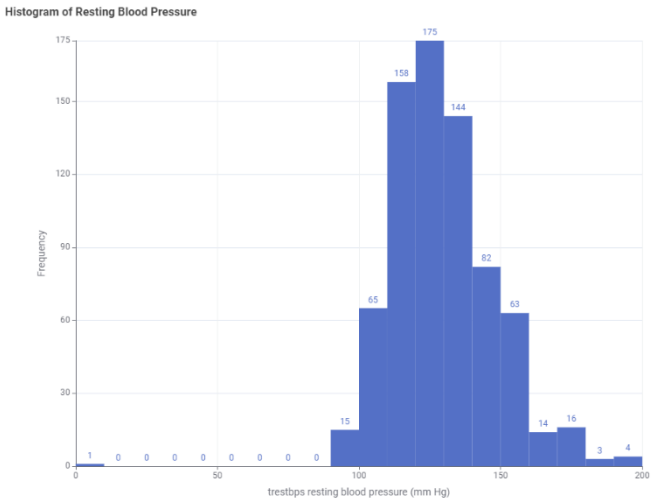

Figure 7: Resting Blood Pressure Histogram

According to an online publication by Johns Hopkins Medicine (no date), "Cholesterol is a natural component in everyone's blood, and supports functions within the body." (Johns Hopkins Medicine, no date) This implies that a cholesterol reading of 0mg/dl is impossible as the body needs cholesterol to function. Furthermore, a blood pressure level of 0 is also impossible as it would imply that blood is not flowing through the patient's blood vessels. Consequently, these zero values can be removed through the **Rule-based Row Filter** node. Interestingly, there are no more 'Switzerland' records after removing these erroneous values, meaning that all the cholesterol readings collected in Switzerland were either measured incorrectly or cholesterol data was lost while being digitally transcribed. This, combined with the previous finding that all of 'Switzerland's' records were incomplete before value cleaning implies Switzerland has issues in its data gathering process.

After removing these lower-bound outliers, the dataset is now composed of 661 records—a 28% decrease from the original 920 records. The **Box Plot** node was then used to easily spot visual outliers. While there are numerical outliers present, they should not be removed; heart disease patients will naturally have unusual medical readings and hence outliers should be expected. According to Gress, Denvir and Shapiro (2018), "Data editing with elimination of outliers [in medical research]… can have significant effects on the occurrence of type 1 error," which is a "statistically significant difference discovered when in reality it does not exist." To avoid type 1 error, the outlier values will not be removed.

Duplicates

Finally, the **Duplicate Row Filter** can be used to check for any duplicate records present within the dataset. A new column named 'Duplicate Status' is appended which shows whether each record is unique or a duplicate; sorting the output table by 'Duplicate Status' shows that all the records are unique.

The workflow with all the previously mentioned nodes regarding data cleaning can be found in the DataCleaning.knar file; a screenshot of the workflow can be found in *Figure 8*.

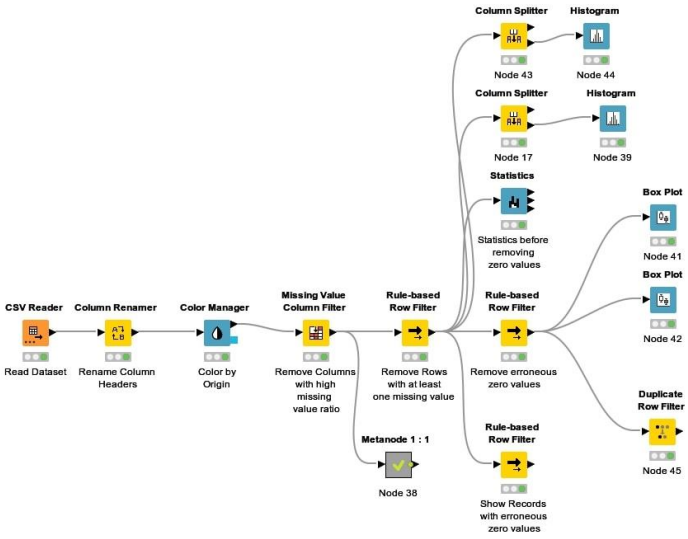

Figure 8: DataCleaning.knar Workflow

Conclusion

A strong framework for getting the dataset ready for further analysis was offered by the exploratory data analysis (EDA) and data preprocessing carried out in KNIME. The dataset was carefully examined and cleaned, guaranteeing its dependability and integrity, using KNIME's flexible nodes and workflows.

Significant missing data in three important attributes (slope, ca, and thal) were discovered during the initial missing value exploratory analysis. These columns were omitted, and the remaining incomplete rows were deleted because of the possible ethical ramifications of imputing missing values in a medical dataset. This method resulted in a smaller sample size but maintained the integrity of the dataset.

Additionally, implausible values, such as blood pressure or cholesterol readings of 0, were identified and eliminated. Potential problems with the data gathering procedures were brought to light by the findings, especially regarding records from Switzerland. This brings to light how important it is to collect data accurately in medical research to prevent biases or mistakes.

In accordance with accepted best practices for medical research, outlier values were kept during outlier analysis to prevent data distortion and type 1 errors. Lastly, duplicate record checks ensured that the analysis was original and free of redundancies.

Overall, KNIME's preparation procedures made sure the dataset was clear, consistent, and ready for additional analysis, such as statistical analysis or model training. These initiatives call attention to how decisive careful EDA and preprocessing are to obtaining trustworthy and consequential comprehension, especially in delicate fields like healthcare.

References

Gress, T., Denvir, J. and Shapiro, J. (2018) ‘Effect of removing outliers on statistical inference: implications to interpretation of experimental data in medical research’, *Marshall J Med.*, 4(2), pp. 1-2. Available at: https://doi.org/10.18590/mjm.2018.vol4.iss2.9

Johns Hopkins Medicine (no date) *Cholesterol: 5 Truths to Know*. Available at: https://www.hopkinsmedicine.org/health/wellness-and-prevention/cholesterol-5-truths-to-know (Accessed: 21 November 2024).

Sterne, J. *et al*. (2009) ' Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls', *BMJ*, 338:b2393, Available at https://doi.org/10.1136/bmj.b2393