Exploratory Analysis of Serious Injuries caused by Accidents using KNIME Analytical Tool

*Ms*. *Sushmitha Suresh*

*Department of computing*

*Letterkenny Institute of Technology*

*Letterkenny, Ireland*

[L00144354@student.lyit.ie](mailto:L00144354@student.lyit.ie)

***Abstract*- Examination of hazard factors that add to the serious injuries in vehicle mishaps has ended up being an interesting and testing issue. The consequences of such examination can enable better to comprehend and conceivably moderate the extreme dangers associated with vehicle crashes and in this way advance the prosperity of individuals engaged with these car crashes. For this purpose, there are several analytical platforms that can be adapted to predict the values to be analysed. Having said that, one such tool called ‘KNIME’ is being used in this paper for predicting the serious injuries from the given dataset. Whilst its counterparts, the core reason behind using this tool specifically is because of the concept of pipelining that has been adopted to create visual data flows to execute the data and also displays the accuracy percentage obtained. Additionally, even the non-experts can visually analyse the data from a given large dataset.**

***Keywords*- *KNIME model design, Serious injuries, Big data***

# INTRODUCTION

Enormous information has turned into a prevailing term in portraying the exponential development, openness, accessibility, and across the board utilization of data in an organized, semi-organized and unstructured configuration in an assortment of business setting. Paying little respect to the dimension of volume, assortment, or speed when dealing with such huge information, is useless except if analysts accomplish something with it that conveys esteem. That is the place “Big data” analytics comes into consideration. One territory where Big data analytics has the best potential to have a huge effect is in the basic examination of automobile crashes and resultant serious injuries. According to WHO 2018 report its estimated that nearly 1.35 million people are dying each year as a result of road traffic crashes(Delen *et al.* 2017). Be that as it may, this pattern can change later on as it is difficult to anticipate the rate at which street auto collisions happen as it can happen in any circumstance. In this manner, an examination can be done that impacts the auto collision seriousness levels utilizing data mining procedures (Atnafu and Kaur n.d.)

In order to encourage the passage to the learning extraction models for people not related to programming building, various business, and non-business programming suites have been made available. Early programming suites include SPSS Clementine, Oracle Data Mining. Yet there is yet a few well understood open source programming examples. Having said that, a specific subset is addressed by workflow situations, in which approximately coupled, singular preparing nodes can be 'darted together' to allow complex computational tasks. (Naik and Samant 2016)

One such platform is KNIME. It is a domain which empowers simple calculations, information control and perception techniques as models. The interface is configurable by choosing among a few distinct techniques. The requirement for secluded data analysis conditions has expanded significantly over the previous years. So as to make utilization of the immense assortment of data analyzation techniques around, it is fundamental that such a situation is simple and natural to utilize, takes into consideration brisk and intelligent changes to the analysing procedure and empowers the client to outwardly investigate the outcomes. To address these difficulties data pipelining situations have accumulated inconceivable force over the previous years. KNIME, the Konstanz Information Miner provides such a pipelining environment.

.

# COMAPRITIVE STUDY

There are many demerits while working on traditional analytical platforms like image processing, mass spectrometry where it involves complex codes to be written for the purpose of controlling the flow and for the change of data (Naik and Samant 2016). Such codes are inclined to be stage subordinate, as well as will in general develop as the analysis advances and are only occasionally very much archived, a reality that blocks the reproducibility of the test. Workﬂow frameworks, for example, KNIME Analytics Platform being a unified data model plan to take care of these issues by giving a stage to interfacing apparatuses graphically and ensuring similar outcomes on diﬀerent working frameworks and its flexible in providing one’s extension if needed. Although there are many data mining platforms like WEKA, Rapid miner, Tanagra, Orange but apparently, KNIME stood first in obtaining the better accuracy when compared to its peers according to Correlation review conference where all the tools mentioned along with KNIME were compared with each other based on classification algorithm namely KNN, Decision tree, Naïve Bayes as shown in the Figure 1 below.

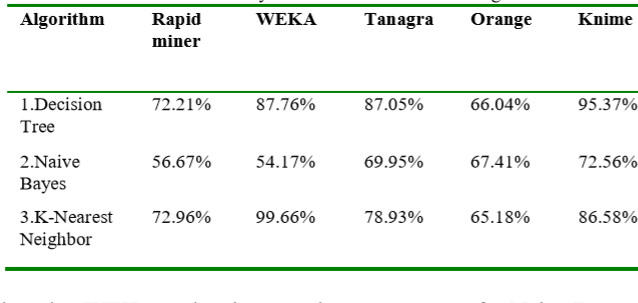


Figure 1: Accuracy measure of classification algorithm (Naik and Samant 2016)

# EXPERIMENT

Analysing dataset using KNIME analytical platform:

Barcelona accident dataset 2017 is downloaded from Kaggle to analyse the serious injuries caused in an accident which has 10340 records in it. Since analysing injuries is the major concept in this paper, the serious injuries column is selected as a nominal value for colour selection in the colour manager node after the accident dataset is fed into the file reader node. When this is directed to decision tree learner node after passing through partitioning node in the experiment, in order to increase the prediction and to reduce the tree size the post pruning option must be adjusted by selecting the configuration option on the respective node. Further the scorer node connected from the partitioning and the decision tree learner through the decision tree predictor is used in the platform to generate the confusion matrix and the accuracy statistics being the result of the experiment. The overall connections between nodes can be seen in the Figure 2 and the final accuracy obtained from the experiment is shown in the figure 3.

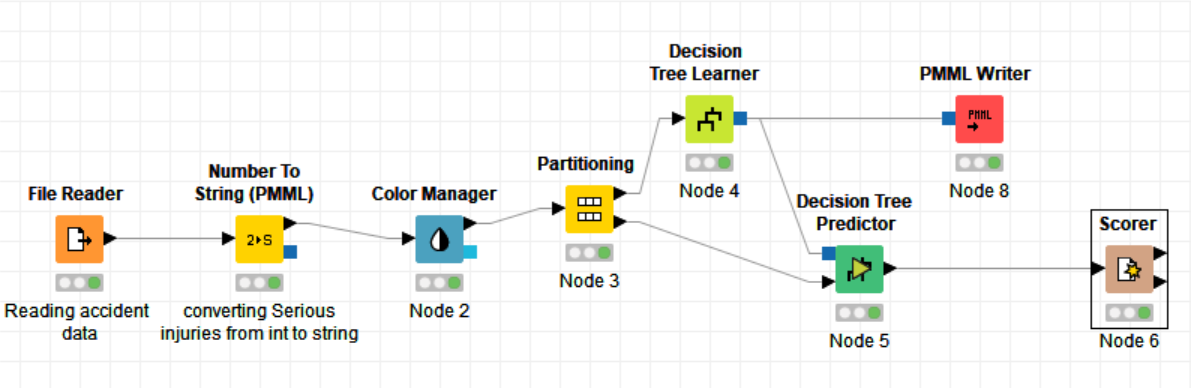


Figure 2: Connection between nodes in KNIME

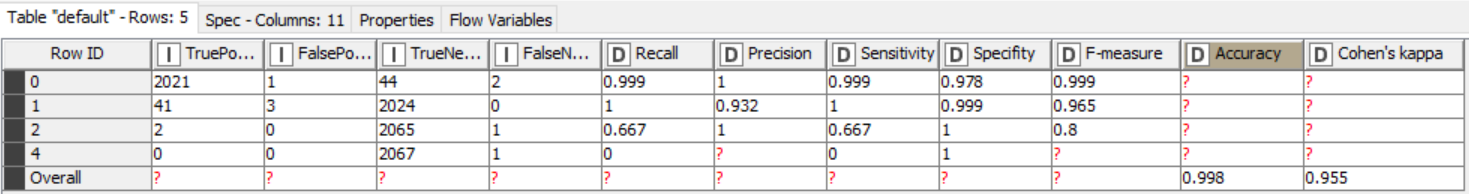


Figure 3: Accuracy statistics

# FUNDAMENTAL VIRTUE OF USING KNIME

KNIME being an analytical platform gives the users the opportunity to explore, visualize and combine the data through pipelines. Some of the features of KNIME that benefit the users are as follows

1. It accepts data from various formats of file and even from URL location since its configured to read various formats.
2. Assigns colours using a colour manager node based on the possible attribute values of the selected column individually or via predefined palettes. However, colours can also be applied to new values present at execution time if the user encounters the selected column is not likely to be relevant.
3. The **Decision tree learner** and the **Decision tree predictor** nodes adds a major advantage to effectively analyse the data in KNIME analytical platform as it has the following fundamental features:

* **Pruning method** that lessens the tree size and abstains from overfitting which really increases the overall performance, and in this way, the forecast quality. "Minimal Description Length" (MDL) pruning is available in KNIME which can be used to obtain a steady exactness for every one of the tests in the decision tree based on persistent traits (Wax 1990)
* **Reduced error pruning** is a straightforward pruning strategy used to cut the tree in a post-handling step: Starting at the leaves, every hub is replaced with its most well-known class, yet just if the forecast precision doesn't diminish. Reduced error pruning has the benefit of effortlessness and speed.
* **No true child strategy** option that has two strategies namely **return null prediction** and **return last prediction**, which predicts the missing value and returns the majority class of the last node respectively whenever it encounters unknown attribute value.
* Another strategy option called **missing value strategy** which also has two strategy options namely **last prediction** and **default child**, where the last known class is updated, and the path is traversed continuously by the use of default child respectively.

1. The input data is partitioned into **train** and **test** (Gavankar and Sawarkar 2015) by the use of **partition node** for the reason being to reduce the contention and to improve the overall performance.
2. The **Scorer node** being the last node to display the prediction accuracy as a final output of the experiment, has **Confusion matrix** (Simon and Simon 2010) and **Accuracy statistics** where the number of matches is known in the confusion matrix table as shown in the Figure 4 and accuracy is obtained as shown in the Figure3.

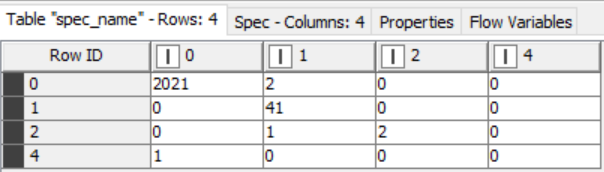


Figure 4: Confusion matrix

# CONCLUSION

In this paper, the goal was to predict serious injuries caused by an accident from a given dataset using KNIME analytical tool.

As part of this experiment, prediction accuracy was established to be in line with the original value thus KNIME is deemed to be a good platform for prediction.

Since KNIME is mainly based on the concept of pipelining, it was found to be easy to implement as all nodes took a single clock cycle to process when each node was connected by a previous node and as a result, I found no buffering between the stages from the beginning till the end of the experiment. Moreover, each node has three indicators which helps users to know if they are ready to connect on to the next node only when the current node is in the executable state. The platform also has a node description on the workspace which paves the way even for the non-experts too to understand the functionality of each node and to effectively use them in carrying out the experiment.

A major advantage found in using this tool was its ‘no true child’ and ‘missing value’ strategies which updates the last prediction when there are missing, or unknown attribute values encountered in the input dataset. On the other hand, a column having only nominal values are accepted to analyse and therefore PMML node was used to convert, this would be a demerit when expecting fast and optimistic performance. Overall, the accuracy predicted was as expected and as a result the number of serious injuries were determined.

REFERENCES

1. ‘Atnafu and Kaur - Survey on Analysis and Prediction of Road Traffic .pdf’ (n.d.).
2. Atnafu, B., Kaur, G. (n.d.) ‘Survey on Analysis and Prediction of Road Traffic Accident Severity Levels using Data Mining Techniques in Maharashtra, India’, 6.
3. Delen, D., Tomak, L., Topuz, K., Eryarsoy, E. (2017) ‘Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods’, *Journal of Transport & Health*, 4, 118–131.
4. Gavankar, S., Sawarkar, S. (2015) ‘Decision Tree: Review of Techniques for Missing Values at Training, Testing and Compatibility’, in *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)*, Presented at the 2015 3rd International Conference on Artificial Intelligence, Modelling & Simulation (AIMS), IEEE: Kota Kinabalu, Malaysia, 122–126, available: http://ieeexplore.ieee.org/document/7604563/ [accessed 22 Feb 2019].
5. Naik, A., Samant, L. (2016) ‘Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime’, *Procedia Computer Science*, 85, 662–668.
6. Simon, D., Simon, D.L. (2010) ‘Analytic Confusion Matrix Bounds for Fault Detection and Isolation Using a Sum-of-Squared-Residuals Approach’, *IEEE Transactions on Reliability*, 59(2), 287–296.
7. https://www.kaggle.com/xvivancos/barcelona-data-sets/version/2#accidents\_2017.csv