DATA ANALYTICS ASSESSMENT: ANALYSE A DATASET

By

Aaron Ward

Supervisor(s):

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
B.SC IN COMPUTING AND INFORMATION TECHNOLOGY
AT
INSTITUTE OF TECHNOLOGY BLANCHARDSTOWN
DUBLIN, IRELAND
2017

Declaration

I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, except where otherwise stated. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references.

I/We understand that plagiarism, collusion, and copying are grave and serious offences and accept the penalties that would be imposed should I/we engage in plagiarism, collusion or copying. I acknowledge that copying someone elses assignment, or part of it, is wrong, and that submitting identical work to others constitutes a form of plagiarism. I/We have read and understood the colleges plagiarism policy 3AS08.

This material, or any part of it, has not been previously submitted for assessment for an academic purpose at this or any other academic institution. I have not allowed anyone to copy my work with the intention of passing it off as their own work.

Dated: 2017

Author: Aaron Ward

Table of Contents

Table of Contents ii

List of Tables iii

Business Understanding 1

Data Understanding 2

Data Preparation 11

Modeling 15

Evaluation 19

List of Figures 1

1 Numeric Attribute Description. 4

2 Categorical Attribute Description. 5

3 Cross Validation - Decision Tree 16

4 Artificial Neural Network Test Results 18

Business Understanding

Business Understanding

This dataset provides data on subjects with and without meningitis. It contains information such as age, gender, location, sum of health problems such as headaches, fevers and seizures. Additionally, is provides an attribute that indicate if the subject does or doesn't have meningitis, with a negative or positive value.

Business Objective

- Predict if someone is at risk of getting meningitis.

Data Mining objective

- The main objective is to create a model to predict the risk of a person getting meningitis
- This model will use the attributes provided in the dataset such as age, gender, seizure history etc to assess the prediction accuracy.
- The model shall test multiple data mining algorithms to obtain a prediction.

Data Understanding

Describing Data

In this section the allocated dataset is explained in terms of informational content, data quality and usability. The data set itself consists of attributes in relation to meningitis, a neurological infectious disease that can cause brain inflammation due to bacteria or viruses infecting that brain. As seen in tables below, the attributes have been segregated from numeric and categorical data. The numeric data given a description and a data type. Additionally, the mean, minimum, maximum and standard deviation values are given. Please refer to table 1 for information of the numeric data.

Numeric Attributes						
Name	**Description**	**Data type**	**Mean**	**Min**	**Max**	**SD**
AGE	List the age of each person	Numeric	37.6285	10.0	84.0	15.3853
COLD	Number of days since last cold	Numeric	2.6642	0.0	35.0	4.8273
HEADACHE	Days since last headache	Numeric	7.1857	0.0	63.0	9.1278
FEVER	Days since last fevers	Numeric	6.3428	0.0	63.0	8.0294
NAUSEA	Start of nausea	Numeric	2.4857	0.0	32.0	4.5856
LOC	When loss of consciousness occurs	Numeric	0.7428	0.0	26.0	2.6481

SEIZURE	When convulsions are observed	Numeric	0.1857	0.0	6.0	0.8780
BT	Body temperature	Numeric	37.625	35.5	40.2	1.3041
STIFF	Neck stiffness	Numeric	1.9571	0.0	5.0	1.4033
KERNIG	Kernig sign	Numeric	0.2142	0.0	1.0	0.4117
LASEGUE	Lasegue sign	Numeric	0.0785	0.0	1.0	0.2700
GCS	Glasgow coma scale	Numeric	14.7071	9.0	15.0	1.1536
WBC	White blood cell count	Numeric	8743.42	1070	90009	7795.80
CRP	C-Reactive protein	Numeric	1.6878	0.0	31.0	4.1317
ESR	Blood sedimentation test	Numeric	5.9285	0.0	60.0	11.880
CSF_CELL	Cell Count in Cerebulospinal Fluid	Numeric	1505.4	0.0	63350	5708.83
Cell_Poly	Polynuclear cell in CSF	Numeric	1025.85	0.0	61520	5402.38
Cell_Mono	Mononuclear cell in CSF	Numeric	465.08	0.0	7840	816.98
CSF_PRO	Protein in CSF	Numeric	99.414	0.0	474.0	96.307
CSF_GLU	Glucose in CSF	Numeric	56.578	0.0	520	44.3412
CSF_CELL3	Cell Count CSF 3 days after the treatment	Numeric	385.18	8	4860	1038.37

CSF_CELL7	Cell Count of CSF 7 days after treatment	Numeric	205.61	0.0	7840	816.98

Table 1: Numeric Attribute Description.

The categorical dataset is given a description to the attribute labels and given a data type. Most of the attributes consist of only 2 values, but does of whom that are multivalued are displayed with the highest and lowest values in the table. See table 2 for further insight to the dataset.

Categorical Attributes				
Name	**Description**	**Data type**	**Value 1**	**Value 2**
SEX	Gender of people	Nominal	M (82)	F (58)
Diag2	Diagnoses	Nominal	VIRUS (98)	BACTERIA (42)
ONSET	Inception	Nominal	CHRONIC (1)	ACUTE (130)
LOC_DAT	Loss of consciousness	Nominal	- (98)	+ (42)
FOCAL	Focal Sign	Nominal	- (105)	+ (35)
CT_FIND	CT findings	Nominal	normal (101)	abnormal (39)
EEG_WAVE	Electroencephalography Wave Findings	Nominal	abnormal (117)	normal (23)
EEG_FOCUS	Focal sign in EEG	Nominal	-(104)	+(36)
CULT_FIND	If bacteria or virus found	Nominal	F (107)	T(33)
CULTURE	Name of bacteria/virus found	Nominal	Tb (1)	- (107)
THERAPY2	Therapy	Nominal	PIPC+CTX (1)	no_therapy (58)
C_COURSE	Clinical course at discharge	Nominal	negative (117)	paralysis (1)
COURSE(Grouped)	Grouped attribute of C_COURSE	Nominal	n (117)	p (23)

RISK(Grouped)	Class label - at risk	Nominal	n (121)	p (19)

Table 2: Categorical Attribute Description.

The data above could be divided into a number of sections. Attributes such as AGE and SEX can be categorized as *personal information*. COLD, HEADACHE, NAUSEA LOC etc. can be described as *subject history* as they provide some information on the commencement of the symptoms. BT, STIFF, KERNIG, GCS can be assigned to a category of *physical examination* as they attribute values obtained during investigation. Further more, *laboratory investigation* used to describe the attributes such as WBC, EEG_WAVE, CULT_FIND, ESR etc. These are values collected during further investigate of the bodily anomalies. Lastly, *postliminary treatment* if used to describe THERAPY2, CSF_CELL3 and CSF_CELL7 as they are attributes describing values after a subject has been treated for meningitis. The dimensionality of the dataset is a total of 36 attributes and there are 141 instances.

Data Exploration

The following section will utilize the Exploratory Data Analysis (EDA) techniques to further analyze the data. This will give insight on how to improve the quality of the data and reduce dimensionality among the dataset. It shall also provide a number of charts to visually represent the data in regards to attributes that may prove to be predictive, show indication of correlations, and data with low variability.

As seen in Figure1 a histogram is used to plot out the main symptoms of meningitis. This consists of the attributes COLD, HEADACHE, FEVER, NAUSEA. The values are attributed to the amount of days since these sensations have been felt. Therefore they are applicable to the mining objectives as they appear to be very predictive.

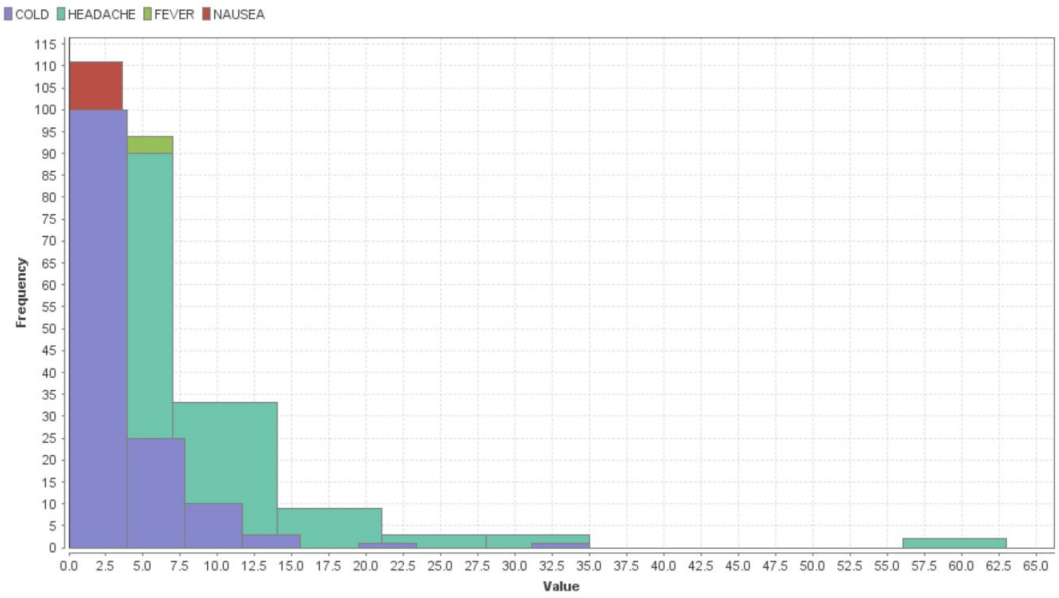


Figure 1: Main symptoms for meningitis

Furthermore, a high correlation between the attributes ONSET and CSF_CELL3 can be seen in Figure 2. Due to this high correlation one of these attributes can be removed from the dataset as they are redundant. Furthermore. They are scarce in values so they may not prove beneficial to the mining objectives.

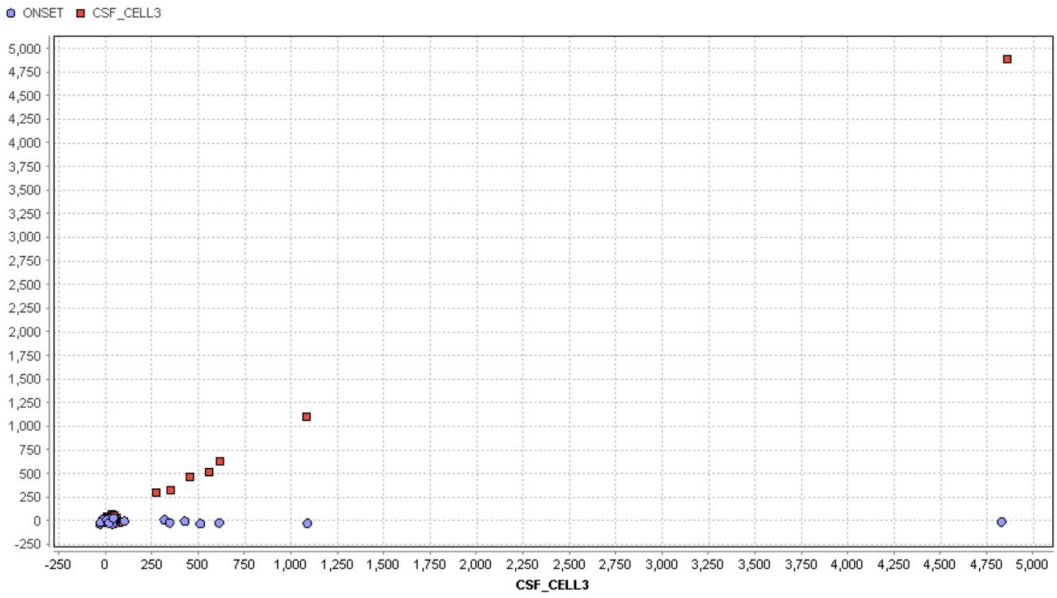


Figure 2: Scatter plot for ONSET and CSF_CELL3

The following histogram displays the values for positive (P) and negative (N) in the class label RISK(Grouped).This unbalance in inequitable group of values may prove to cause some difficulties as it may falsely predict everyone to be negative(n) for risk of having meningitis. See Figure 3 below.

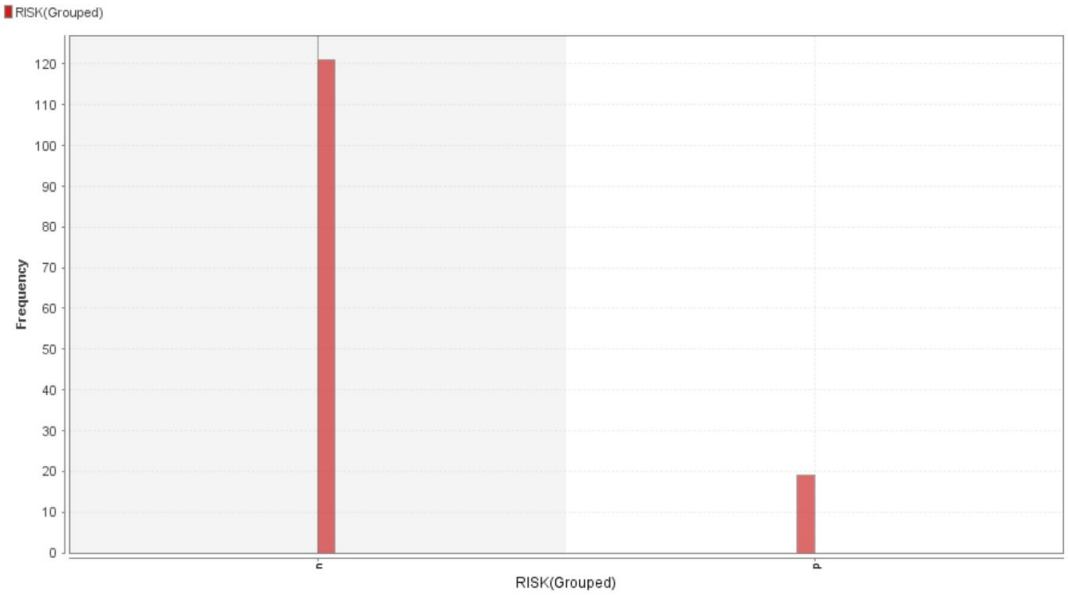


Figure 3: Histogram for class label RISK(Grouped)

Figure 4 shows a box plot of CSF_PRO the distribution of data. Although these values may prove beneficial in distinguishing between at risk and not at risk for meningitis, there are some outlier values seen which can skew the data.

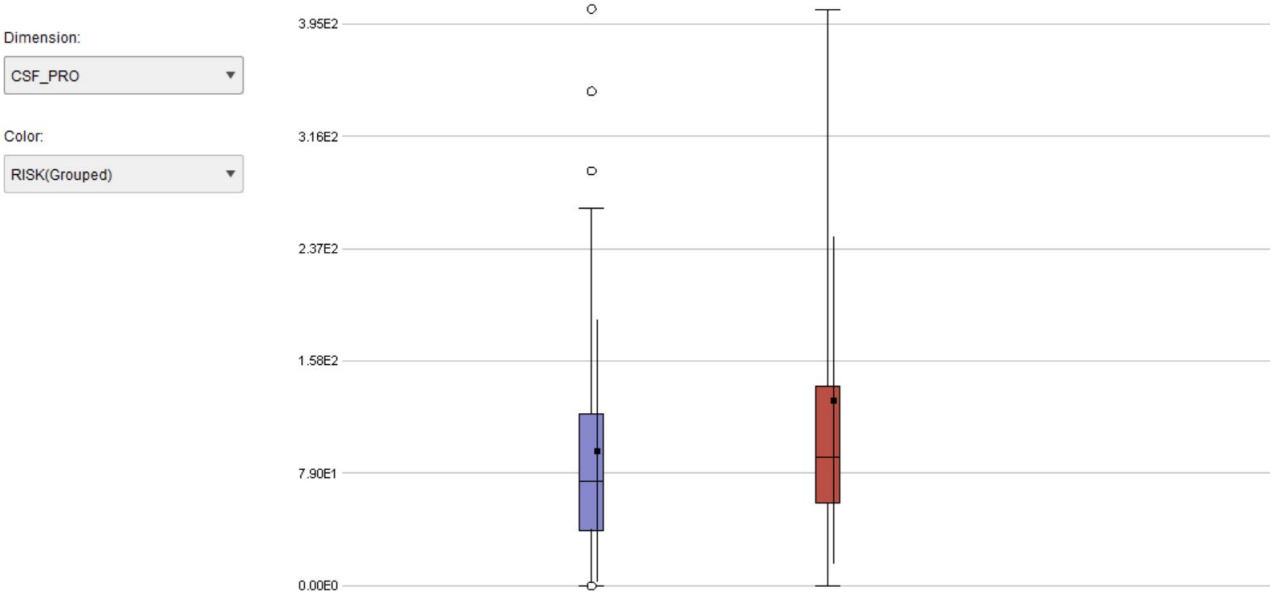


Figure 4: Protein in Cerebulospinal Fluid

Verifying Data Quality

Missing Values

Some missing values can be found in this data set. For example, row 122 contains no values at all. Additionally, CSF_CELL3 has a lot of missing data. There are a total of 119 out of the 141 instances of CSF_CELL3 that a missing values. This means that there are only 22 rows in the entire dataset that do not contain any missing data.

Noise, Bias and Outlier Values

Many outlier values can be seen. For instance, the FEVER attribute has a mean value of 6.343. But this value may be skewed as there is a high outlier value, 63, at row 8. Furthermore, the value 90009.0 in the white blood cell count (WBC). This may be an erroneous value as the next highest value in the column is only 19500. This outlier value may disrupt the accuracy of prediction in the model.

Sufficient Attributes

There are some redundant attributes found in the dataset. One example of this is the SEIZURE

attribute. There are only 8 individual and 4 distinct instances therefore it would be insufficient for representation of a class. Secondly, the CULT_FIND and CULTURE represent the same data. CULTURE would be more sufficient for prediction as it contain more varied values of bacteria/viruses found such as influenza, tb, herpes, strepto etc. rather than the true or false values of CULT_FIND. Therefore CULT_FIND can be removed from the dataset.

Data Preparation

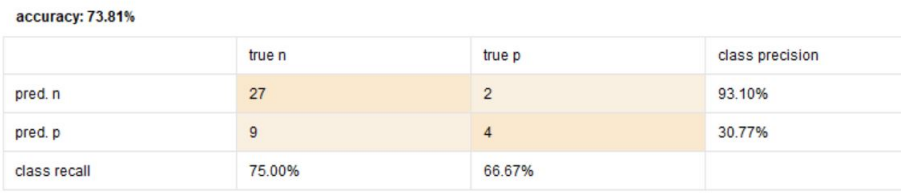
Firstly, the model was tested using a decision tree without any data preprocessing. Firstly, the attribute RISK(Grouped) was assigned as the class label. As many rows contain missing values, a filter was applied to ignore this row of data using the *no_missing_attribute* condition class. The split validation operator was then used for splitting the dataset in a training and testing data set with a split ratio of 0.7. Initially, linear sampling was used for testing the model. This gave an accuracy result of 57.14%. Shuffled sampling was used, which randomly selects 70% of the rows in the dataset as training data.This returned a result of 85.71% percent. Lastly, stratified sampling was used to select 70% of the rows from each class and this gave a result of 83.33%. Secondly, the k-Nearest Neighbor (k-NN) algorithm was used. Linear sampling and shuffled sampling with k-NN both provided an accuracy of 57.14% and stratified sampling gave an accuracy of 66.67%.
Although these accuracies are relatively high, they are only trained on 22 rows of values. This provides a false accuracy as the dataset is not large enough. Therefore, further preprocessing measures are needed to give a fairer accuracy result.

Select Data

There is not a high number rows need to be removed from the dataset as there should be at least 20 times more rows that columns in a dataset. However, some of the rows or columns may be transformed due to missing values, which shall be addressed in the following section.

Clean Data

Due to missing data in a number of cells in the dataset a number of transformational steps must be taken to ensure an increase in the overall accuracy of the final model. As row 122 is the only row in the dataset that has a large amount of missing data, it shall be removed from the dataset as it makes up less then 5% of the number of rows, therefore it will not significantly affect the final result. This was implemented using the *Filter Example Range*, which was used to filter out row 122 from the dataset. Using stratified sampling with a decision tree, this produced an accuracy of 73.81%. Please refer to Figure 5



accuracy: 73.81%

	true n	true p	class precision
pred. n	27	2	93.10%
pred. p	9	4	30.77%
class recall	75.00%	66.67%	

Figure 5: Stratified Sampling Using a Decision Tree

The attribute CSF_CELL3 lacks a lot of values. There are a total of 119 rows out of 141 that contain no value at all which means there is an 84.4% value absence. This column shall be removed as it is missing more than 40% of it's values. This is applied using the *Select Attributes* Filter and exclude CSF_CELL3 from the dataset. There are other attribute that should also be removed. For example, attributes such as LASUGUE, and GCS should be removed because to provide too little variance in values to the dataset. Following the application of these steps, the accuracy of the model has risen from 73.81% to 78.57%, as seen in Figure 6.

accuracy: 78.57%

	true n	true p	class precision
pred. n	31	4	88.57%
pred. p	5	2	28.57%
class recall	86.11%	33.33%	

Figure 6: Results After Filtering Useless Attributes

More ever, WBC possibly has a very high value outlier value of 90009. To accommodate for this *outlier detection* is used to to filter out this value. However, this did not increase the accuracy significantly so it was not kept for further testing. In order to reduce the dimensionality of the data, correlating attributes were removed. This was done by using the *Remove Correlated Attributes* operator. A threshold of 95% was used to filter out the highly correlated columns. In turn, this increased the accuracy by of the model from 78.57% to 85.71%. Although this accuracy is relatively high, the class recall and class precision for **true p** was at a 50/50 split. Therefore it would not accurately predict positive risk subjects. Construction of new data will be used to address this problem, which will be touched upon in the next section.

Construct Data

Seeing as the Meningitis Dataset is relatively small, only having 141 rows, it is not sufficient for modeling. There should be atleast 20 times more rows in a dataset than attributes. Proceeding the initial preprocessing steps, there are a total of 32 attributes in the dataset. Therefore, there should be at least 640 rows. To address this problem, a number of sampling steps were taken. Initially, progressive sampling was used to increase the size of the data. This returned an accuracy of 90.48%, also increasing the class precision from 50% to 75%. The second approach taken was stratified sampling, which proves beneficial when the number of possible classes are known. No increase in performance was seen in this approach so the next sampling technique was taken. Kennard-Stone sampling then tested. Kennard-Stone sampling selects the two rows that are the furthest from one another and then adds other rows are added by adding new rows thats are furthers from objects currently being sampled. Due

to the nature of the Kennard-Stone algorithm, only numeric values can be handled. The *Nominal to Numberic* operator was then used to handle this in regard to nominal values. This approach gave an accuracy of 90.48% and a class recall of **true p** of 33.33%.

The last approach taken was applying a bootstrapping technique to the data, which involves creating new data from the preexisting data. Using the *sample_bootstrapping* operator, an absolute value was initially created with 200 samples to start off with. This produced an accuracy of 95%. 300 samples were then tested, which gave back an accuracy of 93.33%. These 100 sample interval steps were taken all the way to 600 samples. Which then gave an accuracy of 98.33%. Although this accuracy may appear high, it should be acknowledged that when using bootstrap, this accuracy is an over estimation.

Due to class imbalance and the under representation of class of people who are positive of risk of meningitis, Sampling is needed to equalize the classes out. This is done by filtering class **p** and bootstrapping new data to the dataset. Considering the dataset is relatively small, bootstrapping is applied to both classes in order to have a sufficient amount of examples. In total there is a 50/50 split of class labels with a total of 640 examples (20 times the amount of attributes). Please see the Figure 7 for chart of the class label. At this stage, there is a 97.4% accuracy rate.

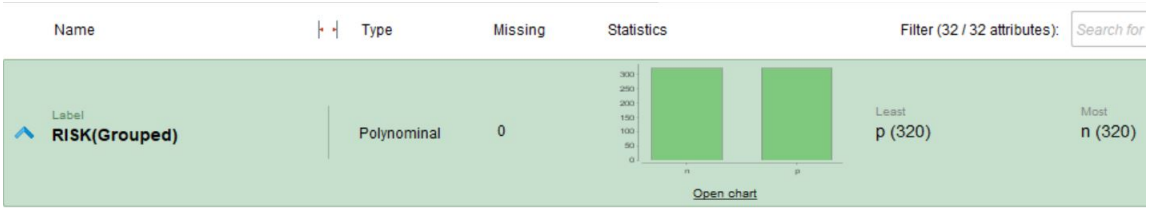


Figure 7: Balanced Classes for RISK(Grouped)

Modeling

Modeling technique

For this mining objective, algorithms that are good for classification rather than prediction is necessary as the the dataset contains nominal values. Therefore, some of the most suitable algorithms for this problem are Decision Trees as output can be easily understood, Artificial Neural Networks as they can be used for prediction or classification.

Test Design

Cross validation will be utilized for creating the test and training data. Where k=5 and using a shuffled sampling type.Using this approach allows for the entire dataset to be used for testing and training. For evaluation on the model, the accuracy will highlight for the different parameters being used in the model.

Build and Assess the Model

Decision Tree

The Decision tree model was first implemented using the preprocessing steps stated in the previous section. K-fold cross validation was then used to create the training and testing data. On the decision tree operator, pruning and prepruning is applied to reduce the complexity of the classifier. Firstly, 5 fold were used, working up until a maximal accuracy value was found. Please see the table below for accuracy results.

Number of Folds	**Accuracy**
5	97.19%
6	97.34%
7	97.66%
8	97.66%
9	97.50%
10	97.81%
11	97.82%
12	97.81%
13	97.50%
14	97.65%

Table 3: Cross Validation - Decision Tree

As seen in the Table 3 the optimal number of folds is 11 for the highest accuracy at a total of 97.82%. The decision tree is properly fitted, as there isn't too little or too many branches in the model. See Figure 8 for a diagram of the decision tree

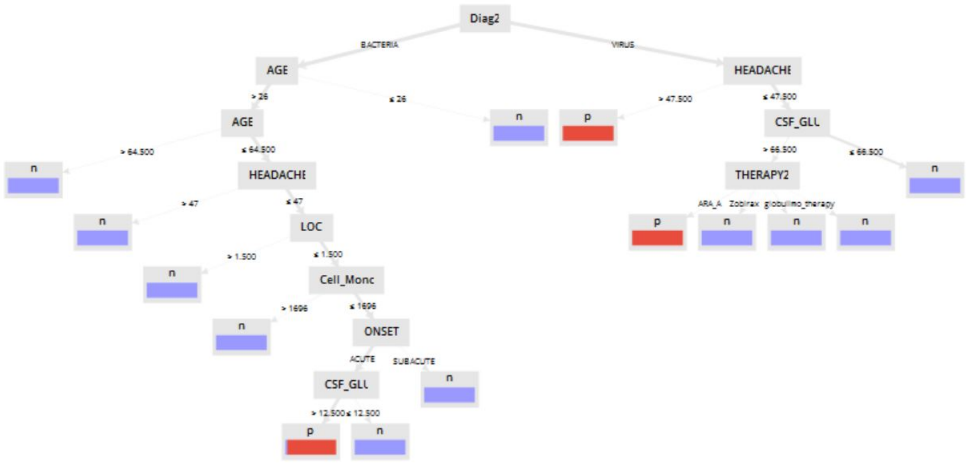


Figure 8: Decision Tree

Artificial Neural Network

The artificial neural was first implemented using the preprocessing steps stated in the previous section. For training and testing data, cross validation was utilized with 11 folds as it proved to be the optimal parameter for validation. The *Deep learning* operator was used as back-propagation is good for training model for higher accuracy. The model used ReLU as it's activation function. The number of epochs are then specified, incremental steps in the number of epochs were tested to find the best accuracy. Five runs were conducted for each interval of epochs tests, each accuracy score was recorded and an average percentage given.

Number of Epochs	**Average Accuracy**
5	97.81 + 97.66 + 98.44 + 97.81 + 97.65 /5 = **97.87%**
10	98.90 + 98.90 + 98.91 + 98.29 + 98.91 /5 = **98.78%**
15	99.69 + 99.37 + 99.37 + 99.37 + 99.06 /5 = **99.37%**
20	99.84 + 99.37 + 99.53 + 99.22 + 99.22 /5 = **99.43%**
25	98.44 + 98.43 + 99.69 + 99.07 + 99.53 /5 = **99.03%**

Average Overall Accuracy	**98.89%**

Table 4: Artificial Neural Network Test Results

As seen in the table 4 above, the optimal amount of epoch for this dataset found was 20, giving the highest accuracy of 99.43% and an overall average accuracy of 98.89%. In terms of performance, this approach could be improved in terms of speed as neural networks are computationally expensive, therefore faster hardware would boost training time.

Evaluation

To conclude, the mining objective have been achieved. A model that can accurately predict if a person is at risk of meningitis was successfully implemented using the Meningitis Dataset. Two techniques were used, giving accuracy results of 97.82% and 99.43%.

The models perform well given the limitations of the dataset, being it's small size, and quality issues. In terms of accuracy, Artificial Neural Networks out done the rest due to it's high results, but lacks in performance of speed due to it's slow computation time. Overall, the business and mining objectives have been met and the project has been completed successfully.