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

Business Understanding and Data Understanding

Business Understanding

This dataset provides data on subjects that 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 neuroligical 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 nominal 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.

Nominal 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: Nominal 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.

Data Exploration

Verifying Data Quality

Data Preparation

- three data preparation techniques to use - Justify the choices made: Discuss why your chosen techniques are appropriate/required for this data set and mining objective. - Document the improvements,

Select Data

Clean Data

Construct Data

Modeling / Data Mining

- Use at least two mining algorithms on the dataset. -

Modeling technique

Test Design

- Explain how you will evaluate the tests

Build and Assess the Model

Evaluation

- Discuss the overall accuracy of your final model - non-technical terms, what information you have learned from the dataset. - Also discuss any limitations of the dataset that may have effected model accuracy