

The key frameworks and concepts covered in modules 1–10 are relevant for this assignment.

Assignment 3 relates to the specific course learning objectives 2, 3, 4 and 5: 2. analyse and

apply strategies processes and underlying technologies for effective management of data to

make evidence based decisions;

3. critically analyse organisational and societal problems using descriptive and

predictive analysis and internal and external data sources to generate insight, create value

and support evidence based decision making;

4. examine legal, ethical and privacy dilemmas that arise from the use of business

intelligence, analytics and evidence based decisions making to comply with legal and

regulatory requirements;

5. communicate effectively in a clear and concise manner in written report style for both

senior and middle management with correct and appropriate acknowledgment of the main

ideas presented and discussed.

**Note you must use RapidMiner Studio for Task 1** in this Assignment 3. Failure to do so

may result in Task 1 not being marked and zero marks awarded.

**Task 1 Predictive Analytics Case Study (40 Marks)**

The goal of the Predictive Analytics Case Study is to predict whether it is likely to rain

tomorrow or not based on previous weather conditions recorded by 49 weather station

locations in the weatherAUS.csv data set provided (see Table 1 Data Dictionary for

weatherAUS.csv data set). You should review the data dictionary for weatherAUS.csv data

set. The Australian Weather dataset contains over 190,000 daily observations from January

2008 through to July 2021 from 49 Australian weather stations. The daily observations are

available from http://www.bom.gov.au/climate/data Bureau of Meteorology. Definitions for

each variable are adapted from http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml.

In completing Task 1 you will apply business understanding, data understanding, data

preparation, modelling and evaluation phases of the CRISP DM data mining process. It is

important that you understand this data set to complete Task 1 and four sub tasks.

|  |  |  |
| --- | --- | --- |
| **Table 1 Data dictionary for weatherAUS.csv** | | |
| **Variable Name** | **Data Type** | **Description** |
| Date | Date | Date of weather observation |
| Location | Text | Common name of the location of the weather station. |
| MinTemp | Real | Minimum temperature in degrees Celsius. |
| MaxTemp | Real | Maximum temperature in degrees Celsius. |
| Rainfall | Real | Amount of rainfall recorded for the day in mm. |
| Evaporation | Real | So-called Class A pan evaporation (mm) in the 24 hours to 9am. |
| Sunshine | Real | Number of hours of bright sunshine in the day. |
| WindGustDir | Polynominal | Direction of the strongest wind gust in the 24 hours to midnight. |
| WindGustSpeed | Integer | Speed (km/h) of the strongest wind gust in the 24 hours to midnight. |
| WindDir9am | Polynominal | Direction of wind at 9am |
| WindDir3pm | Polynominal | Direction of wind at 3pm |
| WindSpeed9am | Integer | Wind speed (km/hr) averaged over 10 minutes prior to 9am. |
| WindSpeed3pm | Integer | Wind speed (km/hr) averaged over 10 minutes prior to 3pm. |
| Humidity9am | Integer | Relative humidity (percent) at 9am. |
| Humidity3pm | Integer | Relative humidity (percent) at 3pm. |
| Pressure9am | Real | Atmospheric pressure (hpa) reduced to mean sea level at 9am. |
| Pressure3pm | Real | Atmospheric pressure (hpa) reduced to mean sea level at 3pm. |
| Cloud9am | Integer | Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eighths. It records how many eights of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast. |
| Cloud3pm | Integer | Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cloud9am for a description of the values. |
| Temp9am | Real | Temperature (degrees C) at 9am. |
| Temp3pm | Real | Temperature (degrees C) at 3pm. |
| RainToday | Nominal | Integer: Yes if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise No. |
| RISK\_MM | Real | Amount of rain. A kind of measure of the "risk". |
| Rain Tomorrow | Nominal | Target variable. Did it rain tomorrow? Yes or No |

**1.1 Exploratory data analysis and date preparation** Conduct an exploratory data

analysis and data preparation of weatherAUS.csv data set using RapidMiner to understand

the characteristics of each variable and relationship of each variable to other variables.

**Summarise the findings of your exploratory data analysis and data preparation** in terms

of describing key characteristics of each variable in the weatherAUS.csv data set such as

maximum, minimum values, average, standard deviation, most frequent values (mode),

missing values and invalid values etc and relationships with other variables, transformation

of existing variables, creation of new variables in a table named **Task 1.1 Results of**

**Exploratory Data Analysis and Data Preparation**.

**Hint:** Statistics Tab and Chart Tab in RapidMiner provide a lot of descriptive statistical

information and useful charts like Barcharts, Scatterplots required for Task 1.1 etc. You

might also like to look at running some correlations and/or chi square tests depending on

whether a variable is a categorical variable or a numeric variable. Indicate in Table 1.1 which

variables which contribute most to predicting whether it is likely to rain tomorrow or not.

You could also consider transforming some variables and creating new variables and

converting target/label variable into a binominal variable to facilitate analysis in Tasks 1.2,

1.3 and 1.4.

**Briefly discuss the key findings of your exploratory data analysis and data preparation**

and justification for variables most likely to predict whether it is likely to rain tomorrow or

not **(10 marks 500 words).**

**1.2 Decision Tree Model** Build a Decision Tree model for predicting whether it is likely

to rain tomorrow or not based on the weatherAUS.csv data set using RapidMiner and a set of

data mining operators in part determined by your exploratory data analysis in Task 1.1.

Provide these outputs from RapidMiner (1) Final Decision Tree Model process, (2) Final

Decision Tree diagram and (3) Decision tree rules. Briefly explain your final Decision Tree

Model Process, and discuss the results of the Final Decision Tree Model drawing on key

outputs (Decision Tree Diagram, Decision Tree Rules) for predicting whether it is likely to

rain tomorrow or not based on key contributing variables and relevant supporting literature

on interpretation of decision trees **(10 marks 150 words)**.

**1.3 Logistic Regression Model** Build a Logistic Regression model for predicting

whether it is likely to rain tomorrow or not using RapidMiner and an appropriate set of data

mining operators and weatherAUS.csv data set determined in part by your exploratory data

analysis in Task 1.1. Provide these outputs from RapidMiner (1) Final Logistic Regression

Model process (2) Key outputs from Logistic Regression Model. **Hint** for Task 1.3 Logistic

Regression Model you may need to change data types of some variables. Briefly explain

your final Logistic Regression Model Process and discuss the results of the Final Logistic

Regression Model drawing on the key outputs (Coefficients, Standardised Coefficients, Odds

Ratios, P Values etc) for predicting whether it is likely to rain tomorrow or not based on key

contributing variables and relevant supporting literature on interpretation of logistic

regression models **(10 marks 150 words)**.

**1.4 Model Validation and Performance:** You will need to validate your Final

Decision Tree Model and Final Logistic Regression Model using the Cross-Validation

Operator, Apply Model Operator and Performance Operator in your data mining processes.

Discuss and compare the performance of the Final Decision Tree Model with the Final

Logistic Regression Model for predicting whether it is likely to rain tomorrow or not based

on key results of the confusion matrix presented in Table 1.4 Model Performance Metrics

(Decision Tree vs Logistic Regression). Table 1.4 will compare the Final Decision Tree

Model with the Final Logistic Regression Model using following model performance metrics

– (1) accuracy (2) sensitivity (3) specificity and (4) F1 score **(10 marks 200 words)**.

**Note** the important outputs from the data mining analyses conducted in RapidMiner for Task

1 must be included in your Assignment 3 report to provide support for your conclusions

reached regarding each analysis conducted for 1.1, 1.2, 1.3 and 1.4. Note you can export

important outputs from RapidMiner as jpg image files and include these screenshots in the

relevant Task 1 parts of your Assignment 3 Report.

**Note** you will find the North Text book and RapidMiner Tutorials useful references for the

data mining process activities conducted in Task 1 in relation to the exploratory data analysis

and data preparation, decision tree analysis, logistic regression analysis and evaluation of the

performance of the Final Decision Tree model and the Final Logistic Regression model.

These concepts are covered in Module RapidMiner Practicals and Chapters 3, 4, 9, 10 and 13

of North Textbook and RapidMiner Tutorials contained within RapidMiner.

Research and critically review the study materials and other relevant literature to provide a

suitable written response to each of the following tasks 2, 3 and 4 supported with an

appropriate level of in-text referencing:

**Task 2 Sentiment Analysis (15 marks 500 words)**

**2.1** Define the concept Sentiment Analysis and explain how Sentiment Analysis relates to

text mining (7 Marks 250 words)

**2.2** Identify and describe a widely used application area of sentiment analysis and explain

why sentiment analysis is used in this application: what business problem does sentiment

analysis address and how does it add value for an organisation and its customers: illustrating

your answer with a real-world example of the application of sentiment analysis by an

organisation (8 marks 250 words)

**Task 3 Big Data Technologies 15 marks 500 words)**

**3.1** Identify and describe each of the three prominent big data technologies using diagrams

where appropriate (8 marks 250 words).

**3.2** Explain the key role (s) that these three prominent big technologies play in managing big

data in an organisation including how these three big data technologies are interrelated and

integrated to achieve effective big data management (7 marks 250 words).

**Task 4 Artificial Intelligence: automation and augmentation in workplace**

**and ethical considerations (20 marks 1000 words)**

**4.1** First, discuss how configurations of humans and artificial intelligence will evolve in the

workplace as organisations drive automation and augmentation through the adoption of

artificial intelligence (10 marks 500 words).

**4.2** Second identify and discuss the ethical implications for organisations in relation to (1)

privacy (2) transparency (3) bias and discrimination and (4) governance and accountability

of using artificial intelligence to drive automation and augmentation in the workplace (10

marks 500 words).

**Report Quality: structure presentation writing and referencing (10 marks)**

**Structure and presentation:** Cover page, table of contents, page numbers, headings,

subheadings, tables and diagrams, use of formatting, spacing, paragraphs

**Writing quality:** Use of English, report written in a clear and concise manner for an

intended management audience (Correct use of language and grammar. Also, is there

evidence of spelling-checking and proofreading?)

**Quality of research evident by correct and appropriate use of referencing:** Appropriate

level of referencing in text, reference list provided, used Harvard Referencing Style correctly

**Assignment 3 Report must be structured as follows:**

Assignment 3 Cover page

Table of Contents

Task 1 Heading – Sub headings for Tasks 1.1, 1.2, 1.3 and 1.4

Task 2 Heading – Sub headings for Task 2.1 and 2.2

List of References

List of Appendices

**You must submit Assignment 3** in Word document format with extension *.docx* and use

**following file naming convention:** Studentno-Studentname-Ass3.docx

**You must use Harvard referencing style – Harvard referencing resources**

|  |  |  |
| --- | --- | --- |
|  | | |
| 1.1 Exploratory data analysis, data preparation of weatherAUS.csv data set | 0 | /10 |
| 1.2 Decision Tree Analysis of weatherAUS.csv data set | 0 | /10 |
| 1.3 Logistic regression analysis of weatherAUS.csv data set | 0 | /10 |
| 1.4 Model performance comparison Decision Tree versus Logistic Regression for weatherAUS.csv data set | 0 | /10 |
| 2 Sentiment Analysis | 0 | /15 |
| 3 Three Prominent Big Data Technologies | 0 | /15 |
| 4 Artificial Intelligence: Automation and Augmentation, Ethical Considerations | 0 | /20 |
| Report Quality: Structure, Presentation, Writing Referencing. | 0 | /10 |
|  | 0 | / 100 |
|  |  |  |