BUSA3020 - Advanced Analytics Techniques

Assignment 2: Predictive Analytics

by John Allagiannis - 45410399

TASK 1 - Experimenting with Alternative Algorithms & Programs

1. Cleaning the Titanic Data

I cleaned the Titanic Data in Excel and my final dataset has the following variables:

name	Names of the 1309 passengers on board the "Titanic" which sank after colliding with an iceberg in 1912
survived	Yes / No
pClass	Passenger Class: 1=First / 2=Second / 3=Third (socio-economic status)
sex	Male / Female
ageGroup	Child = 0-18 years old / Adult = 19-64 years old / Old = 65+ years old (replaced the missing values in Age with the median = 28)
fSize	Family Size: No of Siblings or Spouses on Board + No of Parents or Children on Board
hasCabin	0=No / 1=Yes
poe	Port of Embarkation: S: Southampton / C: Cherbourg / Q: Queenstown (replaced the 2 missing values with C as this information is now known)

```
> head(titanic)
                          name survived pClass    sex ageGroup fSize hasCabin poe
1              Abbing, Mr. Anthony       No      3   Male    Adult     0        0   S
2    Abbott, Master. Eugene Joseph       No      3   Male    Child     2        0   S
3      Abbott, Mr. Rossmore Edward       No      3   Male    Child     2        0   S
4 Abbott, Mrs. Stanton (Rosa Hunt)      Yes      3 Female    Adult     2        0   S
5      Abelseth, Miss. Karen Marie      Yes      3 Female    Child     0        0   S
6   Abelseth, Mr. Olaus Jorgensen      Yes      3   Male    Adult     0        1   S
```

2. Selecting 2 machine learning programs

The 2 Machine Learning programs I used are:

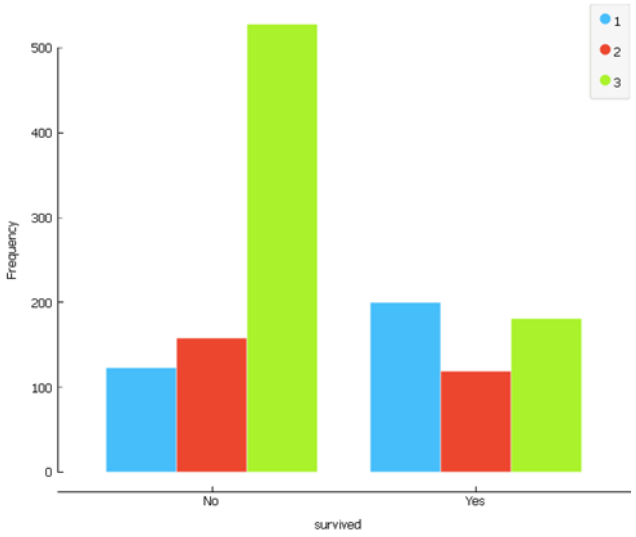
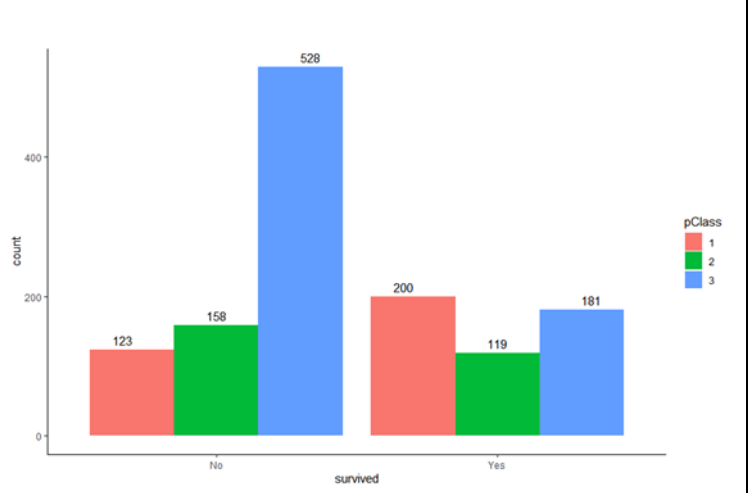
1) Orange Data Miner

2) R statistical programming language (RStudio)

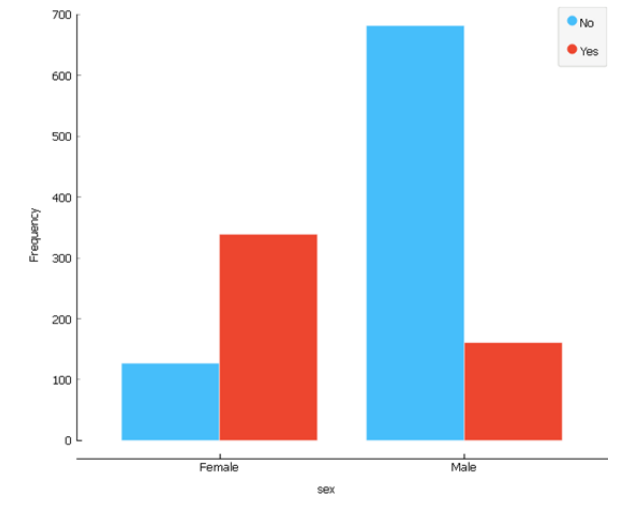
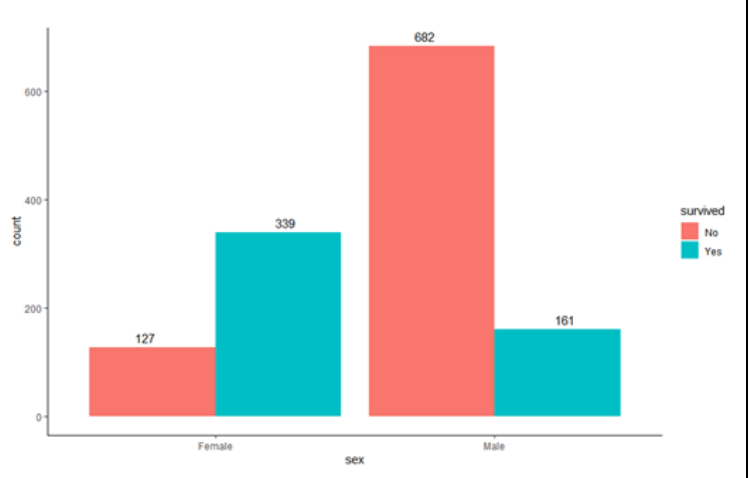
3. Experimenting & comparing my results between the 2 machine learning programs

Firstly, I created column graphs in both programs to demonstrate the relationship that survived (my response variable) has with certain predictor variables (exploratory data visualisation).

1) survived vs. pClass: We can see that if a passenger was third-class, they were less likely to survive, and if a passenger was first-class, they were more likely to survive the Titanic disaster.

Orange Data Miner	RStudio
	

2) survived vs. sex: We can see that females were more likely to survive, and males were less likely to survive the Titanic disaster.

Orange Data Miner	RStudio
	

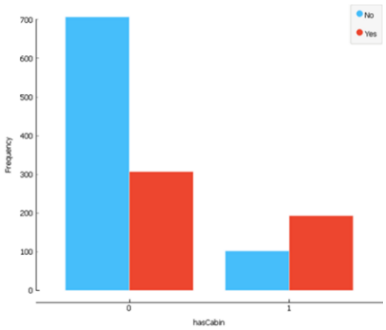
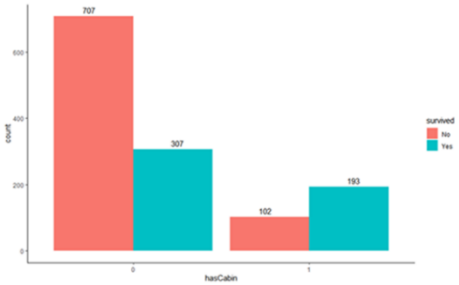
I then ran 4 predictive analytics algorithms to predict survival using training and test data (75%/25% split). The reason for selecting the following algorithms is that they are classification algorithms which allow me to determine how accurately my models are able to predict survival on unseen data and measure overall survival and deaths.

	Predictive Accuracy Comparison (Test Data)	
	Orange Data Miner	RStudio
Decision Tree	0.795	0.783
Logistic Regression	0.780	0.774
Naïve Bayes	0.758	0.761
Random Forest	0.783	0.786

We can see that the Decision Tree algorithm performed the best in Orange Data Miner whereas the Random Forest algorithm performed the best in RStudio. Naïve Bayes performed the worst in both programs.

TASK 2 - Advising a Client on Preferred Program

1. Establishing evaluation criteria for evaluating the 2 machine learning programs

	Evaluation Criteria	
	Orange Data Miner	RStudio
Learning How To Use The Program	- Appropriate resources to begin learning how to use the program including a "Getting Started with Orange" YouTube tutorial playlist, example workflows and a widget catalog	- Appropriate resources to begin learning how to use the program including "R Help" (HTML browser interface for help), cheat sheets and relevant documentation.
Ease of Use	- Simple and intuitive user interface where you place widgets on a drawing board and connect them together to perform tasks	- Simple and accessible user interface divided into panes for writing R code ("R Script"), viewing your data environment and data visualisation
Interpretation of Output	- Nice and easy-to-read output 	- Nice and easy-to-read output 
Integration with Outside Data Sources	- Suitable for Titanic Data (CSV file) as Orange Data Miner has a "File" widget and a "CSV File Import" widget which both recognise and understand the data	- Suitable for Titanic Data (CSV file) as RStudio can read CSV files and recognises and understands the data
Relevance	- Is still relevant today as you can visualise data and perform machine learning operations but feels and is limited	- Is still relevant today as you can visualise data, perform machine learning operations and undertake statistical computing and analysis

2. Advising a client on which is the better/preferred program to adopt for future predictive analysis tasks

The better/preferred program I would advise a client to adopt for future predictive analysis tasks is R statistical programming language (RStudio).

I enjoyed using Orange Data Miner. It is more welcoming to beginners than RStudio with excellent resources on learning how to use the program. Once you have successfully created some connections between different widgets on the drawing board, you have already begun your data analysis. It is very user-friendly, quick, easy but limited. It is limited to what has be done in this assignment, exploratory data visualisation and performing simple machine learning operations.

RStudio, however, a program which will be used in the future of business analytics, is most useful to a client. Compared to Orange Data Miner, beginning to use RStudio felt confronting and a challenge as you cannot just open the program and begin your data analysis. Despite its steep learning curve and having to teach yourself how to code in R, its power and capabilities extend beyond exploratory data visualisation and performing simple machine learning operations. RStudio allows the user to undertake complex and detailed tasks and due to its popularity over Orange Data Miner there are more resources and examples online that will help you achieve your goals when undertaking future predictive analysis tasks.