**The Logbook** for the Data Science Revision Tutorial

This Logbook was created to log the process of creating the Data Science Revision Tutorial

2021

**Author:** Peter Lucan

Institute of Technology Carlow

Table of Contents

[1 Specification Breakdown 2](#_Toc70018041)

[2 The plan 2](#_Toc70018042)

[3 Selection of the Algorithms 3](#_Toc70018043)

[4 Data Science Revision Tutorial Format 3](#_Toc70018044)

[5 The Jupyter Notebook 3](#_Toc70018045)

[6 Creating the Algorithms 4](#_Toc70018046)

[7 Decision Trees 5](#_Toc70018047)

[8 Random Forrest 5](#_Toc70018048)

[9 Naive Bayes 6](#_Toc70018049)

[10 Support Vector Machines 6](#_Toc70018050)

[11Neural Networks 7](#_Toc70018051)

[11 Final Plan For The Presentation 7](#_Toc70018052)

[12 Personal Reflection 7](#_Toc70018053)

[14 Backups 8](#_Toc70018054)

[13 Conclusion 8](#_Toc70018055)

# 1 Specification Breakdown

A tutorial for Data Science that mainly focuses on Data Preprocessing and Machine learning Algorithms. Explain how it works using code examples. Think of it as if it was to teach a upcoming 4th year student.

Each Machine learning algorithm to include:

1. Introduction
2. Sources of useful information
3. Main body containing the algorithm
4. Conclusion

It must also be of a high-quality standard and be presented well.

# 2 The plan

1. Research each machine learning algorithm that I intend to use thoroughly before starting each algorithm Section.
2. Use GitHub to keep my work safely backed up and regularly commit my work.
3. Use this Logbook for planning and use separate smaller Jupyter notebooks to work out the algorithms I will use and once I am happy with them, I will copy and paste them into the Main Jupyter Notebook.
4. The Data Science Revision Tutorial will start with a Cover Page and a Table of Contents and will then proceed by introducing the student to Data Science and explain to him what Data Science is. As well as this some topics around data science will also be explained to the student by using definitions.
5. The next part of the tutorial will define what data preprocessing is and clearly explain how data pre-processing is carried out in detailed steps with appropriate diagrams and will also explain some common terminology and Theory that is needed to know and used in Data Science.
6. In the third part of the data science revision tutorial the chosen machine learning algorithms will be presented, each under their own heading. There will be 5 Machine Algorithms . These Machine Learning algorithms are: 1. Decision Trees, 2. Random Forrest, 3. Naïve Bayes, 4. Support Vector Machines and finally 5. Neural Networks. Each of these five sections will begin with a heading of the algorithm followed by an explanation of the algorithm. Next will be an introduction to the algorithm and the problem that I will be solving using this algorithm in this tutorial. Following on from this I shall discuss the data set that was selected for the algorithm and where it is from. After this I will give definitions and clearly explain some of the Terminology that is required to be known when dealing with this particular algorithm. This should equip the student with enough knowledge to understand what is going on when following this tutorial and doing their own research. After this will be the main body of the algorithm. In this section I will go through data preprocessing and the algorithm. I will only cover data preprocessing for the first or first few machine learning algorithms as data preprocessing is pretty much the same and repetitive for most machine learning algorithm. This will be skipped in some of the machine learning algorithm sections and the student will be told and presented with a clean data set. This is to save time and not to bore the student. The machine learning algorithm section will contain comments to go with the code in order to tell the student what is going on and how it is done. There will also be text markdown cells where appropriate where more explanation is necessary. After this section there will be a conclusion section for the algorithm which will highlight and discuss the algorithms strengths and weaknesses and give a recommendation to the student for what to do for future learning. Following on I will have the Special Thanks To section where I will list all the videos and articles, I have used to put this section together and give credit where credit is due. After this I will present the student with extra learning material in the form of clickable links. These will be handpicked to make sure they are the best learning materials out there for the student at the time of doing this tutorial.
7. The penultimate section of the Jupyter notebook tutorial will have a conclusion for the overall Data Science Revision Tutorial. It will summarize what the student has learned and give him tips for the future.
8. The last section of the Tutorial will have a list of the references used in the creation of this Data Science Revision Tutorial Jupyter Notebook.

# 3 Selection of the Algorithms

There will be 5 algorithms in total. The minimum requirement is 4 but I feel that 5 will be a better number for a tutorial and is probably the most that can be done in the time frame that was given along with all the other work in this hectic year.

# 4 Data Science Revision Tutorial Format

The format for this data science revision tutorial will be a single Jupyter notebook, this logbook and a presentation that will be made at a later date. A single Jupyter notebook was selected as it will be the most efficient way for a student to learn about Data Science and the machine learning algorithms at it will all be in a single source and the student can navigate the notebook cell by cell and it is easy to explain the code that is being presented to the student as it can be executed cell by cell.

# 5 The Jupyter Notebook

I will start the notebook by constructing a cover page for the Jupyter notebook. These types of sections will be made in markdown cells using HTML. The cover page will contain an image and the Carlow IT logo, the title, my name, my student number, the purpose of this notebook and also the date.

The next section of the notebook will be a table of contents that will contain clickable link to different headings within the Jupyter notebook.

Then I will have a short introduction to the notebook followed with a section about Data Science. This section will explain what data science is and why one should study it. This will be a good motivation tool for the student. I think the best way to do this is to have text sections explaining the reasons to choose data science with an image following each reason.

Next, I should talk about Data Preprocessing as it is a major focus of the Jupyter notebook as mentioned by the specification. Once again, I will begin by explaining what Data Preprocessing is and then I shall go through the Data Processing Tasks one by one. I will also include a diagram to provide the student with a better understanding of the steps. I will use diagrams and image where appropriate. In this section I will also define and explain useful terminology and theory around Data Science and Machine Learning Algorithms. This will equip the student with the knowledge to understand some of the topics and theory that will be discussed further down in the tutorial.

I think now I are ready to move onto the Machine Learning algorithm sections there will be 5 of these sections on total because I have chosen to do 5 machine learning algorithms. As previously motioned the 5 machine learning algorithms that I will be implementing are: 1. Decision Trees, 2. Random Forrest, 3. Naïve Bayes, 4. Support Vector Machines and finally 5. Neural Networks. I believe that 5 machine learning algorithms are the maximum amount that can be done in the time frame allocated to this project along with all the other things that are going on as this project will be very lengthy to complete like the main project. I also believe that these 5 machine algorithms will give a good range of coverage across the machine learning algorithms out there and these machine learning algorithms that I will cover are some of the best and most popular ones out there.

In the Machine learning Algorithm section, I will begin by discussing the algorithm that I have chosen. Each of these five sections will begin with a heading of the algorithm followed by an explanation of the algorithm. Next will be an introduction to the algorithm and the problem that I will be solving using this algorithm in this tutorial. Following on from this I shall discuss the data set that was selected for the algorithm and where it is from. After this I will give definitions and clearly explain some of the Terminology that is required to be known when dealing with this particular algorithm. This should equip the student with enough knowledge to understand what is going on when following this tutorial and doing their own research. After this will be the main body of the algorithm. In this section I will go through data preprocessing and the algorithm. I will only cover data preprocessing for the first or first few machine learning algorithms as data preprocessing is pretty much the same and repetitive for most machine learning algorithm. This will be skipped in some of the machine learning algorithm sections and the student will be told and presented with a clean data set. This is to save time and not to bore the student. The machine learning algorithm section will contain comments to go with the code in order to tell the student what is going on and how it is done. There will also be text markdown cells where appropriate where more explanation is necessary. After this section there will be a conclusion section for the algorithm which will highlight and discuss the algorithms strengths and weaknesses and give a recommendation to the student for what to do for future learning. Following on I will have the Special Thanks To section where I will list all the videos and articles, I have used to put this section together and give credit where credit is due. After this I will present the student with extra learning material in the form of clickable links. These will be handpicked to make sure they are the best learning materials out there for the student at the time of doing this tutorial.

The second to last section of the Jupyter notebook tutorial will have a conclusion for the overall Data Science Revision Tutorial. It will summarize what the student has learned and give him tips for the future.

Importantly I shall also have a References section at the bottom of the notebook, and I will fill this is as I use external material throughout the notebook.

# 6 Creating the Algorithms

Every algorithm will be made by experimenting and working things out in a different Jupyter notebook and then when the algorithm is complete this will be copy and pasted into the main notebook to make it look nice, neat, and tidy. These “Rough Work” Jupyter notebooks will be supplied along with the final Jupyter notebook. This decision was made because it is a lot more practical working things out on a Jupyter notebook where you can constantly execute cells in any order you wish rather than in this electronic logbook.

# 7 Decision Trees

First, I must find a data set which will allow us to work on a problem.

A Data set was found on Kaggle which will allow us to use Decision trees to predict whether and item as purchased through the use of online advertising.

Data Preprocessing has to be carried out on this data set as there are some columns I do not need.

I will also rename some columns to make it easier to work with.

After some research I have found out that Decision Trees only work with numerical data and therefore I must encode one of the columns that has the Gender into a numerical equivalent.

Male and Female was encoded into 1 and 0 retrospectively using LabelEncoder().

I will then check the information of the data set.

Now I will train the data by splitting it into a train and test set and I will train the model.

Now I will view the accuracy of the models

The models are very accurate so I may proceed

I then use this model to make predictions.

I put the predictions into the data frames that has all the other values to make it look presentable.

Now I compare the predictions to the actual results and calculate how many times the prediction was right or wrong.

Now I plot some Decision Trees using the model.

As a bonus I can also use the entire data set and also the train set to make predictions too. I repeat the steps above with these sets.

# 8 Random Forrest

I will pretty much do the same thing for Random Forrest as I have done for Decision Trees but with a twist.

I will introduce the confusion matrix to see how accurate my predictions are. I shall plot these on a confusion matrix using Seaborn’s heatmap which is an imported Pandas library.

Now I will make an algorithm that can plot any decision tree I wish by changing a variable from the random forest.

I will make predictions like I have done for Decision Trees also for this algorithm.

Finally, I will also look at the most important features in the dataset by using the important features imported function and I shall output a list that displays information about these features and also graph these using a bar chart using matplot.lib which is another imported pandas’ library.s

# 9 Naive Bayes

For Naïve Bayes I will be looking at a classification problem. I will look at the problem which for Naïve Bayes is most commonly used for. I will be separating messages from into normal messages and spam messages. To do this I will need a data set which will allows us to do this.

I found a dataset on Kaggle which will allow us to do this. This data set only contains two columns. Type of the message and the message itself. This data set has already been preprocessed and I do not have to show these steps for this section as I have done it in the previous two algorithms. I will insert link to the student in case they want to know how to clean the data.

I will again have to encode the labels as Naïve Bayes only works with numerical values. I will do the same as I have done previously.

Next, I will split the data into the train and test sets so I can train the model. I will also check if the data looks correct. The text data in the columns must also be converted into numerical values and this can be done using the Count Vectorizer method. This method converts the values into a matrix by taking the most unique words from each row and placing the number of times they appear into a matrix. After this I can fit the model and then test the accuracy of the test and train sets.

Finally, I will proceed to make predictions using the newly trained models and then I can. Then I will proceed to plot a confusion matrix as I have done before and then I will plot a pie chart as this will be a nice way to display the percentage of spam messages vs the percentage of normal messages.

# 10 Support Vector Machines

The problem that was chosen for the Support Vector Machine is a quite unique one. A data set from Kaggle was chosen which has data of the weight, height, and gender of people. I will use this data set to predict whether a person is male, or female based on their height and weight and in general this is a good way to predict whether a person is male or female even though their does exist the odd case where this is not correct.

Again, I will skip the preprocessing steps as I have covered this in the previous algorithms. The data set that was obtained has already been cleaned for use.

I will pretty much do what I have already done in the previous algorithms except make some more predictions for the student to practice.

# 11Neural Networks

As Neural Networks is the final algorithm that I will cover I will do something a bit different to enhance the students learning experience. I will introduce images and classify them using Neural Networks. More precisely I will be using convolutional neural networks (CNN) in this instance.

After some research I have discovered that there are two ways to do this. One way which is the easy way is to use a library that will allow us to use an inbuilt image set. However, to enhance the students learning experience. I will be importing my own data set of images.

I have obtained a data set that has a folder called shapes with folders of images of Circles, Squares and Triangles with 100 images in each category. This is not enough data to train a neural network, so I have taken these images and duplicated them to have 12,000 images in each category and I have saved this data set on my Google Drive. This modified data set is what I will be using.

I will begin by converting the images into a matrix and converting them into their numerical values. After this I will convert these matrices into arrays of a single column. Then I will split the data into the training and test sets and begin testing the model.

After some working with the code it was discovered that I must flatten the data and also divide it by 255 to “normalize” it as this is the number of pixels of each image.

Following on I will create the model using the kernas sequential function and then I will fit the model and all its layers. The amount of time that the neural network must be trained has to be specified and after trial and error I found that 20 times is the best number of times to train the model.

Finally, I can use this model to make predictions and then I will plot a confusion matrix as I have done for the previous couple of algorithms.

# 11 Final Plan For The Presentation

Before I start working on the final presentation, I will have to tidy up the project structure for a nicer presentation and easier navigation. The folder will contain the main Jupyter notebook, all the images that it uses, and the data sets it uses and the shape data folder. Then it will contain a folder for the rough work Jupyter notebooks and their data sets.

For the presentation I will screen record using MS Teams. I am going to briefly show this notebook and the rough work notebooks and go through the directories. Then I will quickly talk through the notebooks.

# 12 Personal Reflection

Personally, I have enjoyed working on this project. I have learned a great deal about Data Science, and it has sparked my interest this year. I may pursue a master’s in data science. I have decided on what problems to work on by researching each algorithm and then seeing what they are used for. Once I knew this, I obtained data sets that would allow me to work on such problems. I have learned how to solve those problems using research and various tutorials and articles that I have referenced in throughout the Jupyter notebook.

Some of the challenges that I have faced were figuring out some of the errors I had and researching the algorithms. I have to admit that this was very frustrating and time consuming. Especially being stuck at something for very long. One of the major issues for example came at the last algorithm where my laptop struggled with Kernas and Tensor flow after I installed it and I had to do a lot of manual fixes using the terminal.

Most of the other challenges were solved by extensively researching the web for ways that I could fix the problem or simply grabbing a piece of paper and beginning to write and figuring out a solution by myself.

# 14 Backups

I will provide two backups for this project my GitHub and My Google Drive. These links can be seen below.

GitHub: <https://github.com/PeterX12/Data-Science-Revision-Tutorial>

Google Drive: <https://drive.google.com/drive/folders/18QNPogTUctS1T7v1Et26sgsYAIjDdrs3?usp=sharing>

GitHub Cloning Link: <https://github.com/PeterX12/Data-Science-Revision-Tutorial.git>

# 13 Conclusion

In Conclusion I have found writing this logbook before doing some parts of the Jupyter notebook very useful as it allowed me to plan and focus in on things more precisely. It allowed me to plan my steps in advance and then go about solving them. I have planned my steps using this notebook in a general sense and then used the mini Jupyter notebooks to work away and work out any issues I faced and perfected the algorithms in those notebooks than will be attached along with the final deliverable. The logbook served me as a guide to the project and helped me keep track of where I was when I returned to the project.