**Project for programming and scripting data analytics GMIT 2018**

**Fishing Fisher's Flower Figures - An exploration of the Iris Dataset**

**Abstract**

This report explores the iris dataset using data analytics techniques. The iris data consists of 50 samples from three species of iris namely setosa, virgincia and versicolor. It consists of measures of the flowers sepal and petals’ length and width in centimeters. The dataset was analysed using three approaches, descriptive statistics, graphical visualisations and inferential statistics. The method used was statistical analysis using python and python packages. Results showed clear differences between one species and the other two when the data is unlabelled and further separation of data when all species are labelled.

**Introduction**

The iris data set, also known as Fisher's iris data set, was popularised by the statistician Ronald Fisher. [Reference link] (https://en.wikipedia.org/wiki/Iris\_flower\_data\_set). It was collected by Edgar Anderson on a single day in 1936 [reference]. It is a small set of data consisting of 50 samples from three different species of the Iris flower, iris setosa ![iris setosa picture](img/setosa.jpg), iris virginica ![iris virginica picture](img/virginica.jpg) and iris versicolor ![iris versicolor](img/versicolor.jpg). Thus the dataset has five dimensions namely, 4 one dimensional measures in centimetres, sepal length (sepalL), sepal width (sepalW), petal length (petalL), petal width (petalW); And one categorical dimension of iris species (species) [reference the data.csv in the appendix]. The sepal is the small green leaf like structure below the outer petals where the stem and flower meet. Fisher analysed Anderson’s data set to illustrate discriminant analysis [reference]. Discriminant analysis is a statistical method developed by Fisher, for predicting a discrete categorical variable based on information from a continuous predictor variable or variables. Fisher explored if the petal/sepal measurements alone could predict which species of iris the sample came from [reference].

The Anderson measured (in centimetres). the length and width of the iris's petals and sepals (small green leave like structure connecting the flower to the stem). Fisher applied statistical methods to investigate if these measurements alone could identify the species of iris.

The iris data set is a useful and insightfull sample in which to apply data analysis. Its use in exploring machine learning algorithms and in data visualization is well documented. Historically, it is an early example of statistics applied to real data. The data set was collect and analysed by Roland Fisher, a famous biologist and statistian. The data set is small, 150 entries yet is sufficient to illustrate several areas where analysing data is insightful. Fisher's analysis showed two main clusters, with the iris setosa being clearly different from the other two. This seperation into two groups is an example of unsupervised clustering where the data is unlabled. If the samples are labled then three cluster group emerge, illustrating supervised clustering. [reference](https://en.wikipedia.org/wiki/Cluster\_analysis). The data consists of four measurments, the length and width of the petals in centimeters cm and the length and width of the sepal (the green leaf like structure below the flower attached to the stem) in cm ![picture of a sepal](https://amedia.britannica.com/700x450/39/91239-004-44353E32.jpg). 50 sampes of each of the three species were used giving a total of 150 datapoints. the iris data set is one of the most widely known ones in machine learning. Fisher's 1936 paper 'The use of multiple meausrments in taxonomic problems' has been widely cited. The data is used in machine learning as it shows clear encapuslation between the setosa and the other two species when the data is unlabled. if the data is labled by species then it is possbile to sepearate out the versicolor form the virginica species. The iris data is used in machine learning, particulary for testing sorting algorithms. This however maybe add a bit on python and analysis

**Method**

This report produced descriptive, inferential and graphical visualisations of the Iris data set.

**Data**

This project used a copy of the Iris dataset available from [reference]. A sample of which can be viewed in the appendix [reference].

**Apparatus and Materials**

It used an internet connected pc and software to perform the analysis. The computer language used was Python and functions from the Python packages listed below.

* Numpty
* Pandas
* Matplotlib
* Seaborn

**Procedure**

The data was downloaded as a cvs file as /data/iris.csv and imported to python 3.6.

* statistical methods uses Python 5? and the libraries pandas, numpty?, matplotlib seaborn and sklearn?. It produces descriptive statistics of the data including the sepal/petal length and width
* max
* min
* mean

in addition inferential statistics are produced with

* variance
* standard deviation
* skewness
* Kurtosis

This project also produces data visualizations, including scatter plot, box plot etc.....

**Design**

**Participants**

**Procedure**

research the dataset cite the references.

**Results**

put the findings here printouts graphs scatter plots etc

**Discussion**

discuss the findings maybe a discussion of cluster analysis data mining unstructured data and structured data mabye do colourful plots like maps?

from the

**References**

[wiki](https://en.wikipedia.org/wiki/Iris\_flower\_data\_set) [stack exchange](https://stats.stackexchange.com/questions/30788/whats-a-good-way-to-use-r-to-make-a-scatterplot-that-separates-the-data-by-trea/30789#30789) [link](https://stats.stackexchange.com/questions/74776/what-aspects-of-the-iris-data-set-make-it-so-successful-as-an-example-teaching) [link](https://archive.ics.uci.edu/ml/datasets/iris) [](https://www.kaggle.com/sridharcr/data-analysis-iris-dataset) [](https://www.kaggle.com/benhamner/python-data-visualizations) [](http://scikit-learn.org/stable/tutorial/basic/tutorial.html) []() # Appendices