**NAME**

**COLLEGE NUMBER**

**ARTIFICIAL INTELLIGENCE AND DATA MODELLING**

The terms AI data and Big data are used interchangeably

**Abstract**

The development of database systems should be considered and looked at as a wholesome approach. Since data is the heart of the organisation, organizations and individuals are working their way up the ladder to come up with more refined ways of offering information to management. There are three processes to consider when trying to deal with the data system. They include collection, storage, manipulation and visualisation.

**Data collection:**

Is considered the first process in decision making. Top management relies on decision support tools to come up with ready to use algorithms and tools that will necessitate the process of highlighting key decisions to be taken in different methods of addressing the company issues. Sources of these data can also be referred to as data points. Data points are key locations where data gets into the storage devices.

Usually, the company has servers within the applications upon its network that are meant for the consumption and full use of the information. Examples of such data points include. Customer relationship arrangements, company mobile apps, website portals, points of sales, contact forms and links, call logs, email conversations and social media responses. Now, what happens if the company is receiving such too much information that it may not be able to handle, this is where Big data now comes in. AI data will try to answer the following questions;

1. What type of data is this?
2. Where is it coming from?
3. Who is sending it in?
4. How should we store it?
5. How can we analyse it?

Some of the features of big data include the fact that it’s fast-moving, it’s in high amounts and it contains larger amounts of truth in it.

In the next step of the data, the process includes data storage. Also, some kinds of online data storage equally occur and leave traces of data along the path. An instance situation is where you go shopping and then you swipe your card across the POS, card reader or o some machine, the traces of information is left all over. Alternatively, when we also go online to search for a particular product to buy, Google search engine automatically and secretly monitors our key-words, maps these to our locations and then use this information to send use recommendations from other sites based on the type of item and the want on what to buy and how closely it is related to the items we are looking for online.

This is one of the reasons why sometimes after searching for a particular product, you get so many recommendations from other sites, mobile apps and social media pages that you visit. This kind of targeted selling may appear illegal and uncalled for but in some way, tots are still allowed. Maybe we need to review further the kind of laws and policies that we put in place to safeguard human data privacy, such privileges should not just be protected by some weird terms and conditions that usually consumers do not even read or understand.

Further looking at data storage, we can consider that applications majorly store data on three different types of formats.

* Structured
* Unstructured
* No storage (web 3.0)

**Structured database systems** are the kind of traditional database systems that are stored in the form of organised rows and columns, the attention here is paid to the organised rows that contain the unique primary and secondary keys that are used to access these databases. These data values are stored in an existing table. A table could contain several columns for storing different tables names also known as variables. A variable is a storage name for a particular type of data. They can store strings, numbers also known as integers, characters, Booleans and even float variables.

The data is managed by the language known as the structured query language also known as SQL the SQL runs by requesting formation stored on several tables and columns. It can also give database information, schemas, metadata, and table information and then join different tables together as we shall see later in this discussion. A database management system (DBMS), on the other hand, is a tool for managing structured data. Examples of this DBMS include MySQL, Oracle, DB2 and Postgres.

Almost all these databases work in the same way, the difference however could be a little syntax agreement here and there, the company name and the whether or not to accept capital SQL queries or just to ignore. Postgress accepts lower case written queries. Any other query that is written in the upper case will automatically be converted back to the lower case. The database is case sensitive.

Consider the below table sample for organised /structured database system

|  |  |  |  |
| --- | --- | --- | --- |
| Student\_id | name | wing | grade |
| 0003 | Alex maercies | Blue | 4 |
| 773 | Jontahan Mario | Red | 3 |
| 6443 | Simon Trucey | Green | 1 |
| 5243 | Lucia Degraada | Yellow | 2 |
| 1109 | Kimberly Golden | Purple | 3 |

Structured databases systems have been used for period’s ad and are the oldest in history considering their wide range of use. They are the legacy systems and databases. Their use and span from financial systems, healthcare education systems, research and development studies and in the financial stocks too, among other different developer environments.

Unstructured database systems on the other hand are the exact opposite of the structured database, Instead of data being stored in organised rows and columns, the data is tied in some unstructured format, meaning one cannot exactly tell if the data is in which particular row or index value position.

Unstructured data is the second latest type of database system, to get into the industry. Even some developers and software engineers are not well versed with the database. Sad but this indeed is where most applicators generating large amounts of data are going to shift their attention and focus, this is ideally true since UDBS is the only way the current applications tie with AI data. Since UBD does not consider any special kind of specific rows and columns, the UDBS will store images, audio, characters, integers and even documents that can only be accessed as an array. Consider the given example of an array calling the object below:

for (i in siteInfo .users)

{

    for (j in siteInfo.users[i])

    {

        x = siteInfo.users[i][j];

        console.log(x);

    }

}

The array above retrieved from an unstructured database fetches an object called site info and returns the associated array of items within the object. This is the way this type of database works.

Google has one particular product called Firebase. This is a tool that gives the developer and organisational access to so many features including the Firebase data tools. This type of database does not define data in any kind of row or columns. Another type of database in use is MongoDB, widely used for the same, reasons as the Firebase. Also, the latest introduction to data storage winch in essence is not even stored in web 3.0. You see when the internet was invented, it all began from a static site to dynamic systems, and now to smart systems and now the whole system is almost going back to web 1,0 but smartly and silently.

Instead of manually storing data in some scattered or unstructured database systems, the data now creates and destroys itself, users are identified by sessions and identifications. This is what cryptocurrency now trades in. No one knows the database or the particular data that is inside those applications. Instead, what we have are smart contracts. Smart contracts are barely owned by various partners who run the etherium grid and every computer on the networks identified by the key id of a particular user, when the user wants to trade or buy ether, the particular smart contract is run on the blockchain is called, stored in the system temporarily until l the transaction is marked as complete then contact is destroyed. Nowhere does the smart contact keep user details. This is for the same reasons why we say that blockchain technology is safe and secure for transactions, it’s also safe in the long run. Can be it hacked? No.

The next step in the data process is data analysis and storage. Developers and data engineers have come up with several tools, algorithms and languages that will sweeten the life of a data scientist/ scholar/ application engineer by making it easier to analyse data very fast. You see traditionally, after the introduction of the common business-oriented language COBOL, introduced the COBOL tool in the 1960s, other scientists in the same domain have introduced scientific analysis tools like FORTRAN and SCALA that focused on analysing data in the field of geography science.

Later on, developers would take advantage of the missing opportunity that was there for the engineers and thereafter leverage on the needs of getting calculations and arithmetic applications on their dataset. This is where the development of statistical languages such as R came in. R has been a powerful child in trying to help get quick summary statistics on the concerned measures of central tendencies, summaries on the values of comparison and also by helping give the required forecasting and the data descriptive features and values of such datasets. Moreover, the language has an inbuilt mode so which will enable the researcher to quickly get whatever kind of insights that they need from such dataset. Some of the models applied in this language include:

* Linear regressions
* Logistic regressions
* ARIMA models
* KNN model
* Random forest and decision trees

The models have been effective in trying to help scientists discover the relationships between datasets and still be able to predict the future outcomes of such inputs by forecasting the models through a fit function. Even though old and powerful, Python has been used to wrangle data during the whole data cleaning process, analysis and even building models to predict the dataset. Data analysis and features of this application include the ability to gather data from the internet through certain custom tools like beautiful soup. Once his data collection and preparation is ready then Pythons inbuilt libraries like Numpy, SciePy and pandas have been used as tools in trying to make the data visualisation process much easier.

Nonetheless, an old but still widely used approach to data analysis is a tool that we all know, Microsoft Excel is a tool that benefits almost every user and basic analyst regardless of the experience or age of the researcher. Using Microsoft office, as we shall see in the later chapters to come, an individual can quickly clean their dataset, get measures of central tendencies and then apply the relevant required functions on the data analyses tool pack to quickly get meaningful information from the dataset, as we shall discuss in the last part of this research papers in the analyses part.

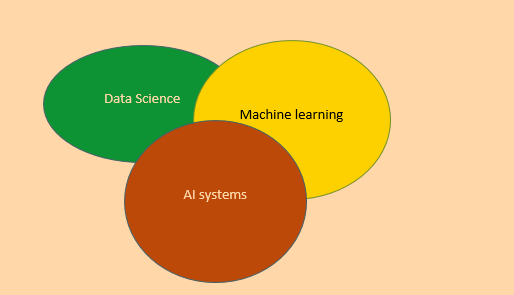
Software as a service (SAS) is a web tool that has been specifically designed to help the user quickly get around their analysis. With key attention to the analyses of the products and outputs derived from the dataset that they intend to use. Moreover, in this particular study, the SAS tools shall form the basis of our analysis with key objectives and instructions being sent on how we can use the SAS software to convert the CSV dataset of our choice into the correct .bat file that we can use on our application to make meaningful and useful data derivations from. Through a click and drag, we shall be quickly able to get the right analysis that is needed for this work.

All these introductions are necessary order to help the company to understand and measure the requirements of developing an effective business information system with closer attention being paid to the datasets, systems and whether to consume the structured or unstructured database, the method and language of analysis and the kind of models to apply for each dataset. Finally, the two tools considered here for the development and management of the AI application shall be the SAS online tool and the earlier mentioned SQL language to help management understand the features and capabilities of AI applications.

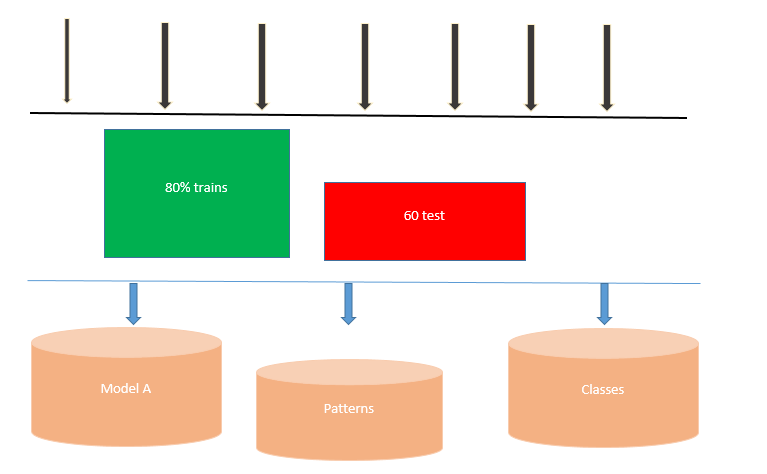
**Artificial intelligence and Data Science**

Moreover, we cannot discuss Artificial intelligence without bringing in the factor of Data science. AI is an intelligent tool that relies on data science to monitor and operationalize its quest for smart systems. Before an AI agent is constructed, usually, a bunch of datasets is taken through numerous training and retraining, collected over large volumes of time and scale, after which this data is subjected to modelling.

The reason why a good amount of data is used is sent into the system is because of the need to learn, and this is where we derived the term machine learning, the machine learns by respective algorithms of classification, so, when the model is finality fit into it, then it quickly searches its memory table and assigns the correct value or response to the data in questions. AB testing requires that 80% of the data we use should be for training and then the other 20% should be for testing. Look at the diagrams below:

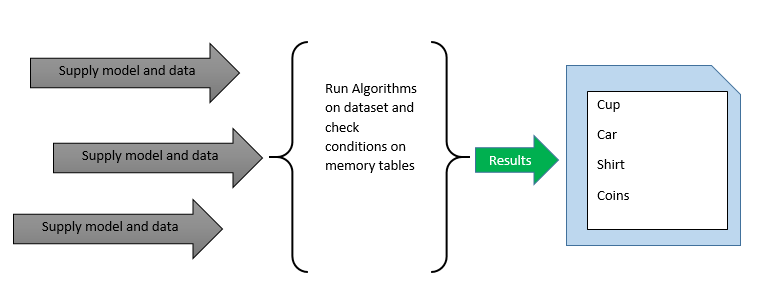


The figurative description above shows the narrow description that sits between the three disciplines, to build a good AI agent, we must have the correct data in place, which we gave discussed above, then, once this data is ready, it has to be trained very well, and then modelled in a manner that the AI systems can recognize.

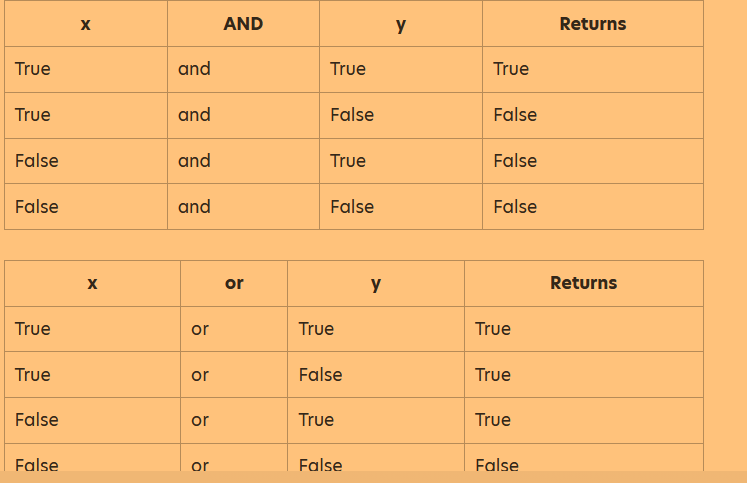


Incoming datasets

Then the resulting model can now be used to build the smart AI agent as a case example below:



The models and the datasets are supplied to the logical memory of the AI agent and then suing the earlier predefined classifier, a new pattern or cause for action is identified in the memory tabled and assigned to the object in question. Generally, this is how smart systems work. However. This may look complex but in reality and behind the scenes, the AI algorithms have several conditional stamens that help it make the right choice on the item from the memory tables to pick and based on the closest, it decides which one best describes the subject in question:



A good case is with an AI agent that checks user geographical location and automatically sets time to the user:

import time;

IF userSays = = currentTime():

THEN:

return:

Lookup\_local \_time()

print("Speak “It is:" . + [%time%]"

END:

**Literature review**

In an article written by Cohen (2020), there is a need for organisations and corporations to ensure satisfactory data storage mechanisms. A lot of organisations do not store their data and neither do they keep track of it. Since data is the future gold of the inquires, cloud solutions and scalable servers are supposed to be impended to keep track of whichever changes are coming in and then historically start to analyse how the data trend is performing over the years. This is important especially in fields where the subject matter keeps changing frequently. An example here is with the health care industry, Organisations need to assess how the pattern of the disease is changing with time and be able to use predictive analytics to simulate future variances.

Rabase et al (2020), producing future data scientists who will be able to implement effective AI systems will require adequate increments in institutions of higher learning by training more data scientists in this field. Currently, the world is still in shortage of adequate engineers who understand the value of the changing data field. New concepts are being introduced into the field and the best that organisations can do is to train more data scientists, our world is going to rely on the data we produce and consume today. Concepts such as machine learning, deep learning are still new in most countries and institutions need to train specialists in these fields.

As shared by Lughofer et al (2019), traditional decision support tools are changing and evolving to include a cross-organisational 360-degree view. Initially, this is the kind of work that was left to CRMS and intelligent systems, but this was only partial. More focus and attention is now being drawn to data modelling and forecasting, this was a big element that was missing in earlier applications of DSS. To achieve this AI elements have had to be incorporated into these legacies.

The earlier literature reviews by these scholars do not wholesomely describe the relationship between AI data, legacy systems and their stance in information security, this is still a wide-open topic that future researchers can look at.

**AI data modelling and experimentation**

Since artificial intelligence relies on machine learning to help build the correct agents, we shall try to build a successful model that can be used with any AI agent or robotics to help classify objects and materials.

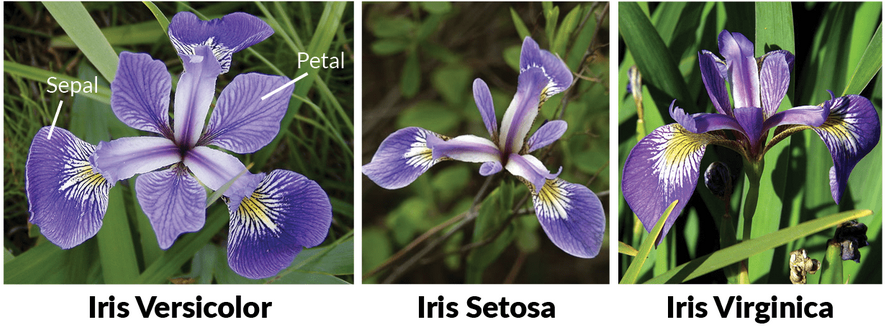
**The methodology:**

We shall use the dataset IRIS, availed as an open ’/public dataset and can be accessed via the open link available at: <http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data>, the dataset can also be accessed on R as an inbuilt package. Further, we shall use the Rstudio IDE, together with several packages that come inbuilt with R.

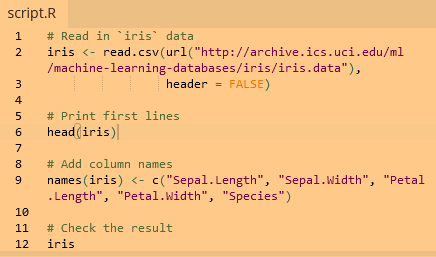
**Aims and objectives**

1. To use machine learning to classify different flowers

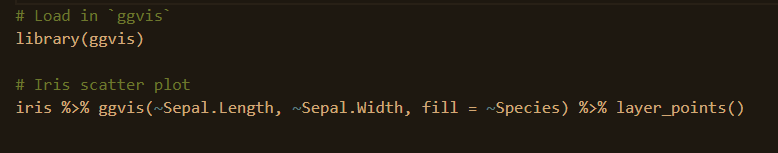
During the experiment, we shall try to establish whether a given property fits the classification of the flower as either Iris flower or not. Traditionally, we are all aware that a flower has a sepal and a petal, and in most cases the sepals are green, but this is not true for the Iris flower family, as below:

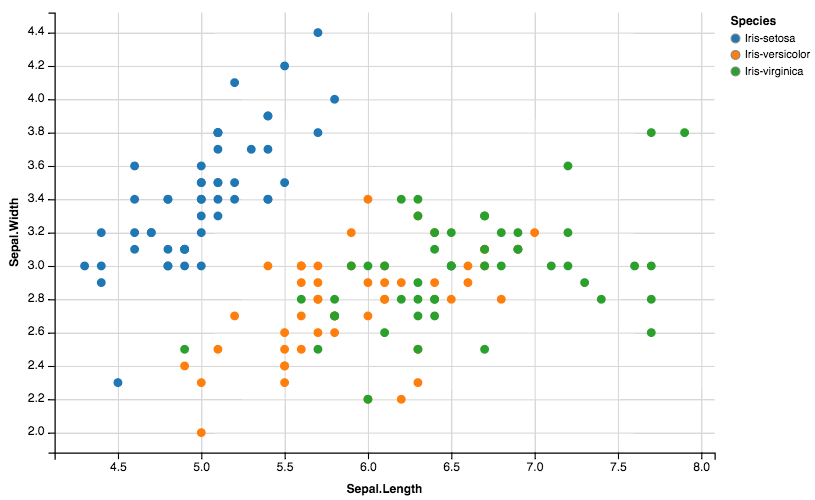


The first process is to run quick visualisations on the dataset by loading it into the Rstudio IDE:



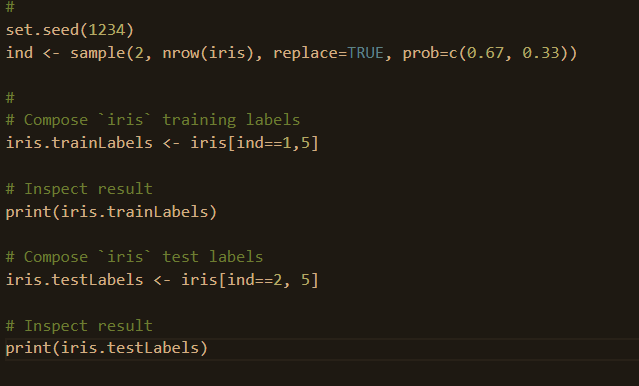
Get a scatter plot distribution of the dataset:



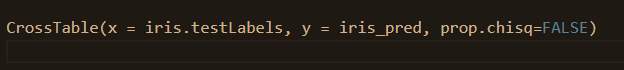


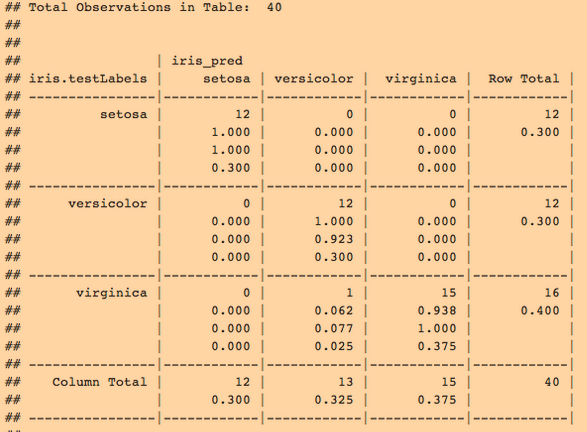
There is a high correlation in the dataset between in the setosa family, unlike the other strains where the relationship is much more spread. Now that we have an understanding of the dataset, let’s apply the machine learning model to it, which in this case we shall employ the KNN model:

The next step is to set our seed population then split our set into train and test while assigning in the relevant expected probabilities during prediction



Evaluating the model on Chisq and comparing the prediction models against the test tables reveals that almost all objects in the flower family were correctly classified and so what this means is that our model is correct and may not need any longer adjustments:





**Summary and conclusions**

Since robotics is coming a key player in the industrialisation and automation of the process, serious studies need to go into how we handle data. The understanding of the data components that include data acquisition, processing and modelling are significant to building efficient artificial intelligent machines. However, this data needs to be well stored, storage mechanisms can either be on structured or unstructured database systems.

Further, the current and future smart agents are going to rely heavily on big data, which is the primary food for smart agents. These agents need to process big data continues to make accurate predictions and causative behaviours that match their real-time environment. There are quite several models that can be applied to smart agents to help make the right decisions, however, in this project, the KNN model was proved to have almost correctly classified all the datasets supplied to it. Finally, smart technology is the future of science and innovation, there is a need to train more data scientists and engineers who will readily pick up such roles in the future.

**REFERENCES**

Cohen, R., 2020. Introduction to Data Management, Life History, and Demography.

Corea, F., 2019. Introduction to Data. In An Introduction to Data (pp. 1-5). Springer, Cham.

Rabasa, A. and Heavin, C., 2020. An Introduction to Data Science and Its Applications. In Data Science and Productivity Analytics (pp. 57-81). Springer, Cham.

Cuocolo, R., Perillo, T., De Rosa, E., Ugga, L. and Petretta, M., 2019. Current applications of big data and machine learning in cardiology. Journal of geriatric cardiology: JGC, 16(8), p.601.

Laughter, E. and Sayed-Mouchaweh, M. eds., 2019. Predictive maintenance in dynamic systems: advanced methods, decision support tools and real-world applications. Springer.

Tran, T.O., 2020. Database Systems.

Ilin, I., Klimin, A. and Shaban, A., 2019. Features of Big Data approach and new opportunities of BI-systems in marketing activities. In E3S Web of Conferences (Vol. 110, p. 02054). EDP Sciences.

Grover, V., Chiang, R.H., Liang, T.P. and Zhang, D., 2018. Creating strategic business value from big data analytics: A research framework. Journal of Management Information Systems, 35(2), pp.388-423.

Zhu, H., 2020. Big data and artificial intelligence modeling for drug discovery. Annual review of pharmacology and toxicology, 60, pp.573-589.

Golding, N., 2019. greta: simple and scalable statistical modelling in R. Journal of Open Source Software, 4(40), p.1601.