Coventry University

psb ACADEMY

Bachelor of Science with Honours in Computing Science

COURSEWORK – 1

MODULE: BIG DATA

CODE : PSB302IT

WEIGHTING : 40%

DUE DATE : 28 Jan 2022

Submitted By: hein thihan

Name : Thihan hein

Index No. : 066H9FCZ

Word Count : 1700(Approx.)

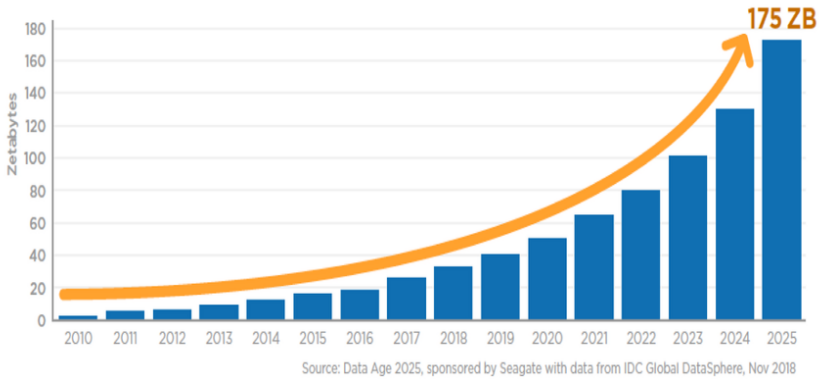
Introduction

Many data are generated in the form of texts , messages, photos, searches, phone calls, emails, videos in a single user using smartphones. And if we multiply by 5 billion which is half of the world population, that data is quite a lot that our daily computing system is difficult to handle and this kind of massive data is what we term as Big Data. 2.1 millions of snaps are shared on snap chat, 3.8 billion of information were searched and 1.8 millions of Facebook are created within a single minute and that is a lot of data. So how do we clarify that any data as big data? And we can describe by the concept of 7 V's (Volume,Velocity,Variety,Variability,Veracity ,Visualization, Value) which will be cover in the following topic in detail

7 V's of Big Data

Volume

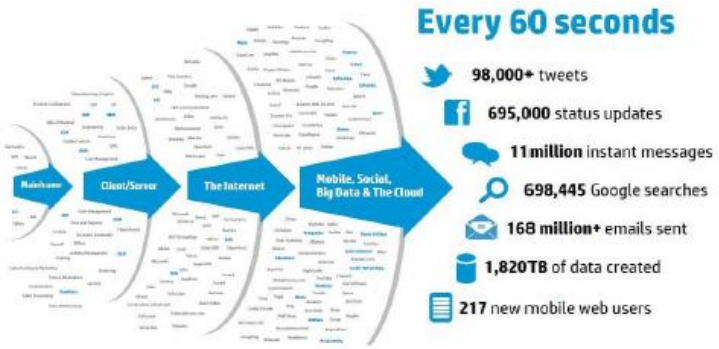
Volume is defined as the quantity of information we get and we can measure as Gigabyte but now it used to measure with Zettabytes (ZB) or even Yottabytes (YB). Data volumes that, in fact, can reach unbelievable altitude. As reported by research, there are over quintillion bytes of data are produced every single day, which will result in 40 zettabytes of data by 2020.



Data Growth Over the Year (medium,Surya,2020)

Velocity

It means the speed at which data is generated and transported. This is an important point to consider for companies who want their information move quickly so they may create the correct business resolution as possible. Twitter tweets, Facebook postings, Google searches are examples of data that is created at a high velocity.



Example of velocity. (researchgate, Khalid, na)

Variety

Structured (Excel sheet), semistructured (Log file), and unstructured (videos, images) data are all examples of variety Unstructured data is disorganized data that arrives in a variety of files and formats. It is not suitable for RDBMS (relational database management systems) since it is not suitable for traditional data models. Information that was not organized into a certain archive but includes connected data , is referred to as semi-structured data. Semi-Structure data is faster and more accurate than unstructured one. Information that is organized into a specific order is referred to as structured data. This tells that the data is more transportable, accepting for more efficient data computing and investigation. As an example, audio and video footages that are captured by city's CCTV are example of variety, it can contain structure data like city, street name in spreadsheets and audio and videos are unstructured data.

Variability

Data variability is the extent to which a set of data is dispersed. Variability allows users to define how much data sets differ and to compare their data to all other sets of data utilizing statistics which means data is changing constantly. It can be described with four main ways -

Range - The range is the difference between the set's smallest and largest item.

Interquartile Range - The interquartile range (IQR) is a number that reflects how evenly distributed scores are and informs consumers of the range in the center of a collection of scores.

Variance - Users may get a basic indication of how spread out data is by looking at the variance of a data collection.

Standard deviation - determines how closely a user's data is clustered around the mean.

Variability refers to the quantity of discrepancies in the data in big data. Irregularities in which large amounts of data are brought into the database are often referred to as volatility. As an example, when a user sees a different recommendation every 10 seconds on an

commerce-website, that is that variability. That variability can affect the quality and accuracy of our data

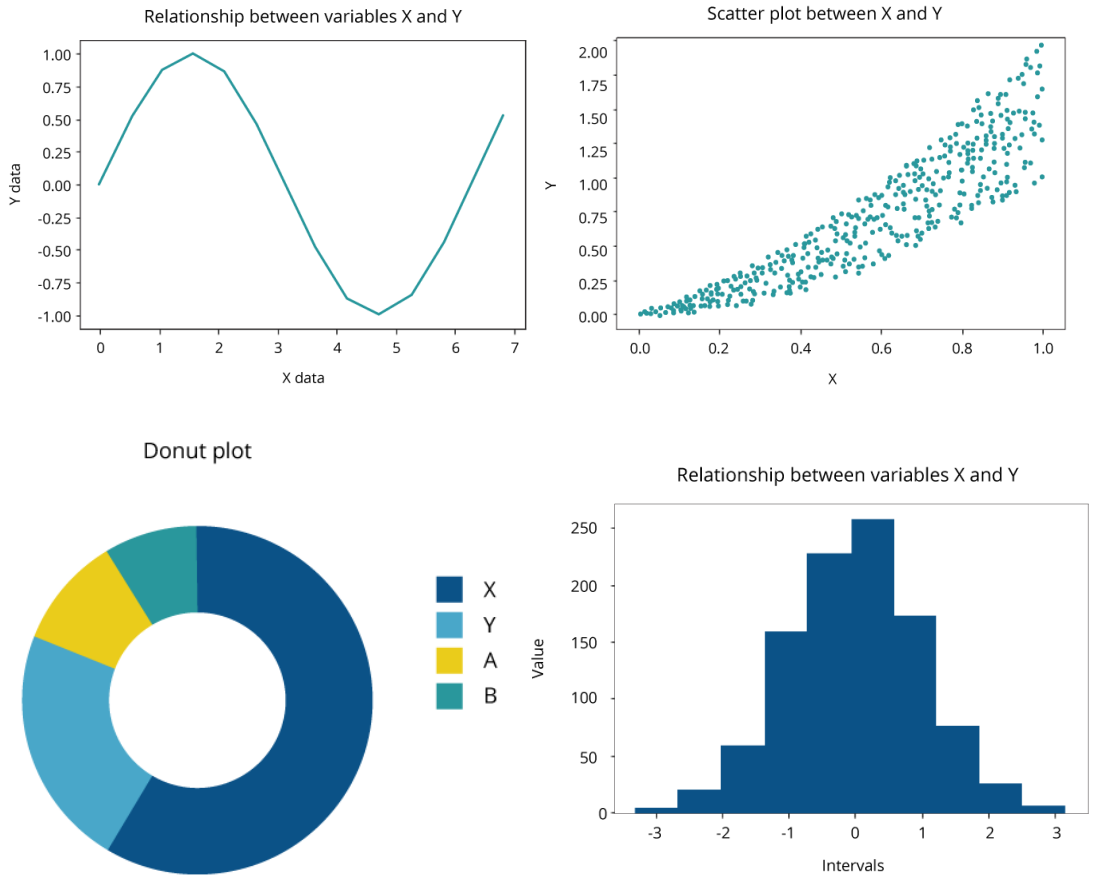
Veracity

Accuracy and quality of the generated data termed as veracity. Collected data might be insufficient, erroneous for providing actual, actionable information. Overall, veracity refers to the amount of confidence in the data collected.

Data may become cluttered and difficult to utilize at times. If the data is incomplete, a big volume of data might produce more complex. In healthcare industry, the accuracy and quality of raw data of cancers are very important to detect whether a patient is in cancer or not. If it does not have any accuracy, doctors can have wrong information about patient's health condition.

Visualization

In today's environment, visualization is essential. Using pie charts and line graphs to present large quantities of intricate data is significantly more powerful than using spreadsheets(i.e Excel) and using graphs is easier to visualize for many users.



Examples of Data Visualization (medium, Sociforce, 2019)

Value

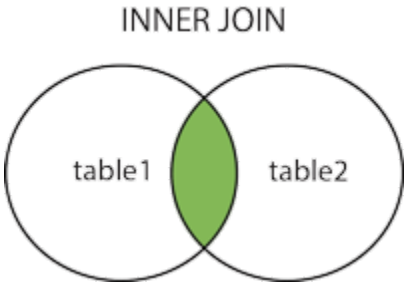
Big data value refers to the usefulness of gathered data for a business. Data must be transformed into something useful in order to obtain information. We can capture value by using the following one of them. We can identify the value of big data. As an example of an automation process such as the autopilot of Tesla is the usefulness of gathered data. Without having the correct value, Tesla's autopilot will crash every single person or object on the street.

Four Types of SQL table joins

SQL joins are used to join two or more tables based on the relationship of primary key and foreign keys which are related columns between them. In this report four different types of SQL joins will be discussed with examples in the following topic.

Inner Join

It returns a new result by selecting records which contain match values in both tables. However inner join will not show the result that does not have match value in both tables.



SQL Inner Join(w3school, na)

Demo Database Table

For demonstration, Student and Address table will be used for all table joins.

student_id	Name	email	age
001	michael	michael123@gmail.com	18
002	John Doe	johndoe123@gmail.com	18
003	John Martin	john123@gmail.com	20
004	Maria Anders	maria123@gmail.com	22
005	Kamado Tanjiro	kamado123@gmail.com	19

address_id	address	student_id
1	First Street plot No 4	001
2	Second Street plot No 5	002
3	Asakusa	005
4	1904 Clark Street, Long Island City	004

We have a student and address which is a relation of 1 to many relationships. We have five data in pas_students and 4 data in addresses. So it will show the result which matches the primary and foreign key.

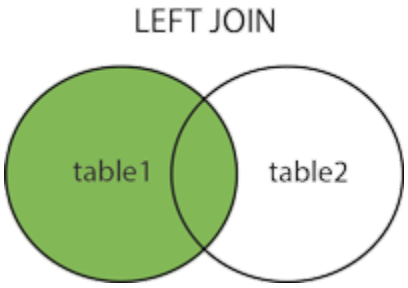
```
SELECT
psb_students.student_id,psb_students.name,psb_students.email,psb_students.age,addressess.address
FROM psb_students
INNER JOIN addressess
ON psb_students.student_id = addressess.student_id;
```

student_id	name	email	age	address
001	michael	michael123@gmail.com	18	First Street plot No 4
002	John Doe	johndoe123@gmail.com	18	Second Street plot No 5
005	Kamado Tanjiro	kamado123@gmail.com	19	Asakusa
004	Maria Anders	maria123@gmail.com	22	1904 Clark Street, Long Island City

Since student id with 003 does not have any relation, it will not be shown. This is how inner join works.

Left Outer Join

It returns all the results of the rows on the left side which has a matching row for the table on the right side.



SQL Left Join(w3school, na)

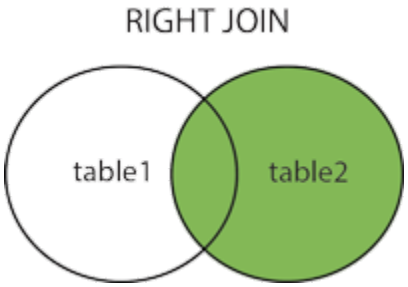
```
SELECT psb_students.student_id,psb_students.name,psb_students.email,psb_students.age,addressess.address
FROM psb_students
LEFT JOIN addressess
ON psb_students.student_id = addressess.student_id;
```

student_id	name	email	age	address
001	michael	michael123@gmail.com	18	First Street plot No 4
002	John Doe	johndoe123@gmail.com	18	Second Street plot No 5
005	Kamado Tanjiro	kamado123@gmail.com	19	Asakusa
004	Maria Anders	maria123@gmail.com	22	1904 Clark Street, Long Island City
003	John Martin	john123@gmail.com	20	*NULL*

Since the student table is on the left, it will iterate all the results on the left side(psb_students) and combine with the right side (address table). Since there is no data which is related to student_id 003 it will show Null .

Right Outer Join

It is opposite to the Left Outer Join. It will show the result of a row on the right side which has matching rows for the table on the left side.



SQL Right Join (w3school, na)

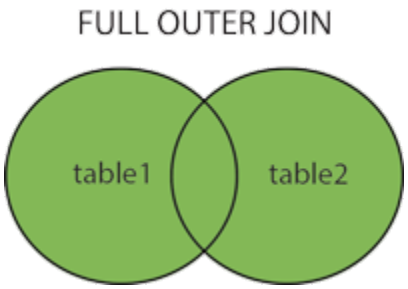
```
SELECT psb_students.student_id,psb_students.name,psb_students.email,psb_students.age,addressess.address
FROM psb_students
RIGHT JOIN addressess
ON psb_students.student_id = addressess.student_id;
```

student_id	name	email	age	address
001	michael	michael123@gmail.com	18	First Street plot No 4
002	John Doe	johndoe123@gmail.com	18	Second Street plot No 5
005	Kamado Tanjiro	kamado123@gmail.com	19	Asakusa
004	Maria Anders	maria123@gmail.com	22	1904 Clark Street, Long Island City

When we put the address table on the right join, the result will be different, since there is no data with student_id 003 it will not show that row even though they have a relation. It is because student_id 003 is not available in the address table.

Full Join

As it is named Full join, it combines data from both tables. It can generally return a large data set if we have too much data in our tables.



SQL Full Join (w3school, na)

```
SELECT psb_students.student_id,psb_students.name,psb_students.email,psb_students.age,addressess.address
FROM psb_students
FULL JOIN addressess
ON psb_students.student_id = addressess.student_id;
```

student_id	name	email	age	address_id	address	student_id
001	michael	michael123@gmail.com	18	1	First Street plot No 4	001
002	John Doe	johndoe123@gmail.com	18	2	Second Street plot No 5	002
005	Kamado Tanjiro	kamado123@gmail.com	19	3	Asakusa	005
004	Maria Anders	maria123@gmail.com	22	4	1904 Clark Street, Long Island City	004
003	John Martin	john123@gmail.com	20	*NULL*	*NULL*	*NULL*

It will iterate data from both tables, if there is no data in one of those tables, it will replace it with null as the picture shows above.

We have heard the value of Big Data. To archive that value, we need something that can detect the trends and patterns of the data. So, Machine learning can help this process by incorporating it's algorithms. There are many machine learning tools and in this report Orange will be the chosen one.

Orange Data Mining Tool and Its History

Orange is an user-friendly user interface open source data visualization, machine learning tool written in Python,Cython, C and C++. It started being released in 1996 and it has been 25 years since then. It provides an understandable GUI application for data analysis and data visualization. Even if they want to study and utilize data science, not everyone is eager to learn coding. GUI-based solutions can help in this situation.

In 1997 Donal Michine thought that machine learning needed an open toolbox. Michie's concept of a workshop was represented application that work in browser (web application)that people can submit data mining programs, methods, testing scripts, and data to be shared in a collaborative web workspace.Despite the presence of IBM's Java team, which might aid them in developing toolbox, the solution was not ready, and WebLab's initiative had faded by the time the conference finished. But, at least for Janez and Donal, the concept piqued our curiosity, and development of what is now Orange began shortly after.

By using Orange machine learning tool, it will be demonstrated by using weatherAUS.csv and those processes will be explained in detail in the following content.

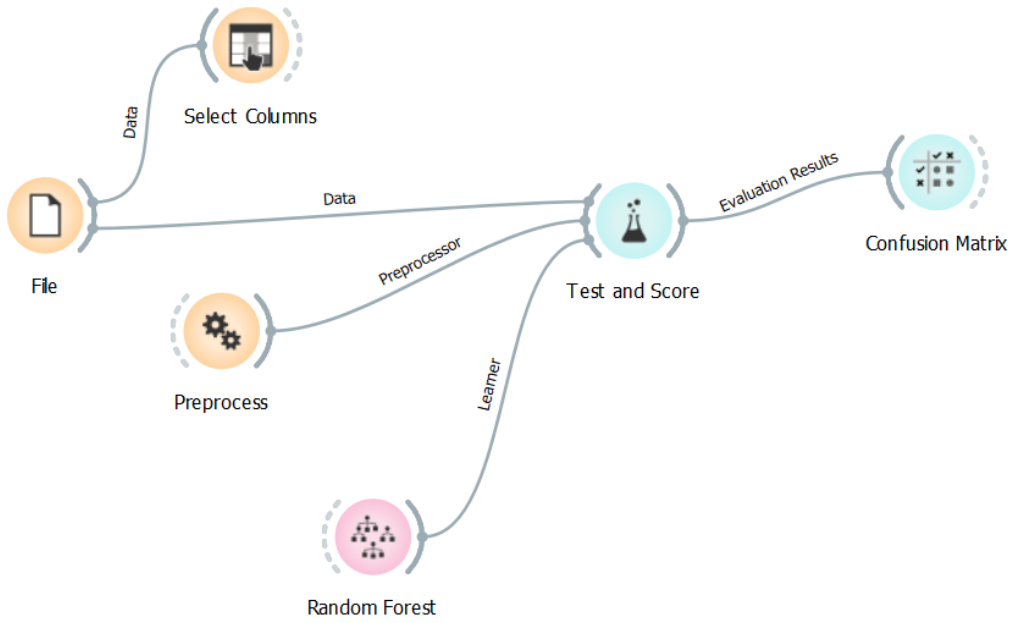
WeatherAus.csv

This csv contains the following headers and data. In this case the Rain Tomorrow column will be the target to predict the data. There are some NA (Not Available) data which will be replaced or removed during data clean up.

	A	B	C	D	E	F	G	H	I	J	K
1	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporatic	Sunshine	WindGust	WindGust	WindDir9	WindDir3
2	########	Albury	13.4	22.9	0.6	NA	NA	W	44	W	WNW
3	########	Albury	7.4	25.1	0	NA	NA	WNW	44	NNW	WSW
4	########	Albury	12.9	25.7	0	NA	NA	WSW	46	W	WSW
5	########	Albury	9.2	28	0	NA	NA	NE	24	SE	E
6	########	Albury	17.5	32.3	1	NA	NA	W	41	ENE	NW
7	########	Albury	14.6	29.7	0.2	NA	NA	WNW	56	W	W
8	########	Albury	14.3	25	0	NA	NA	W	50	SW	W
9	########	Albury	7.7	26.7	0	NA	NA	W	35	SSE	W

L	M	N	O	P	Q	R	S	T	U	V	W	X
WindSpee	WindSpee	Humidity9	Humidity3	Pressure9	Pressure3	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow	
20	24	71	22	1007.7	1007.1	8	NA	16.9	21.8	No	No	
4	22	44	25	1010.6	1007.8	NA	NA	17.2	24.3	No	No	
19	26	38	30	1007.6	1008.7	NA	2	21	23.2	No	No	
11	9	45	16	1017.6	1012.8	NA	NA	18.1	26.5	No	No	
7	20	82	33	1010.8	1006	7	8	17.8	29.7	No	No	
19	24	55	23	1009.2	1005.4	NA	NA	20.6	28.9	No	No	
20	24	49	19	1009.6	1008.2	1	NA	18.1	24.6	No	No	
6	17	48	19	1013.4	1010.1	NA	NA	16.3	25.5	No	No	

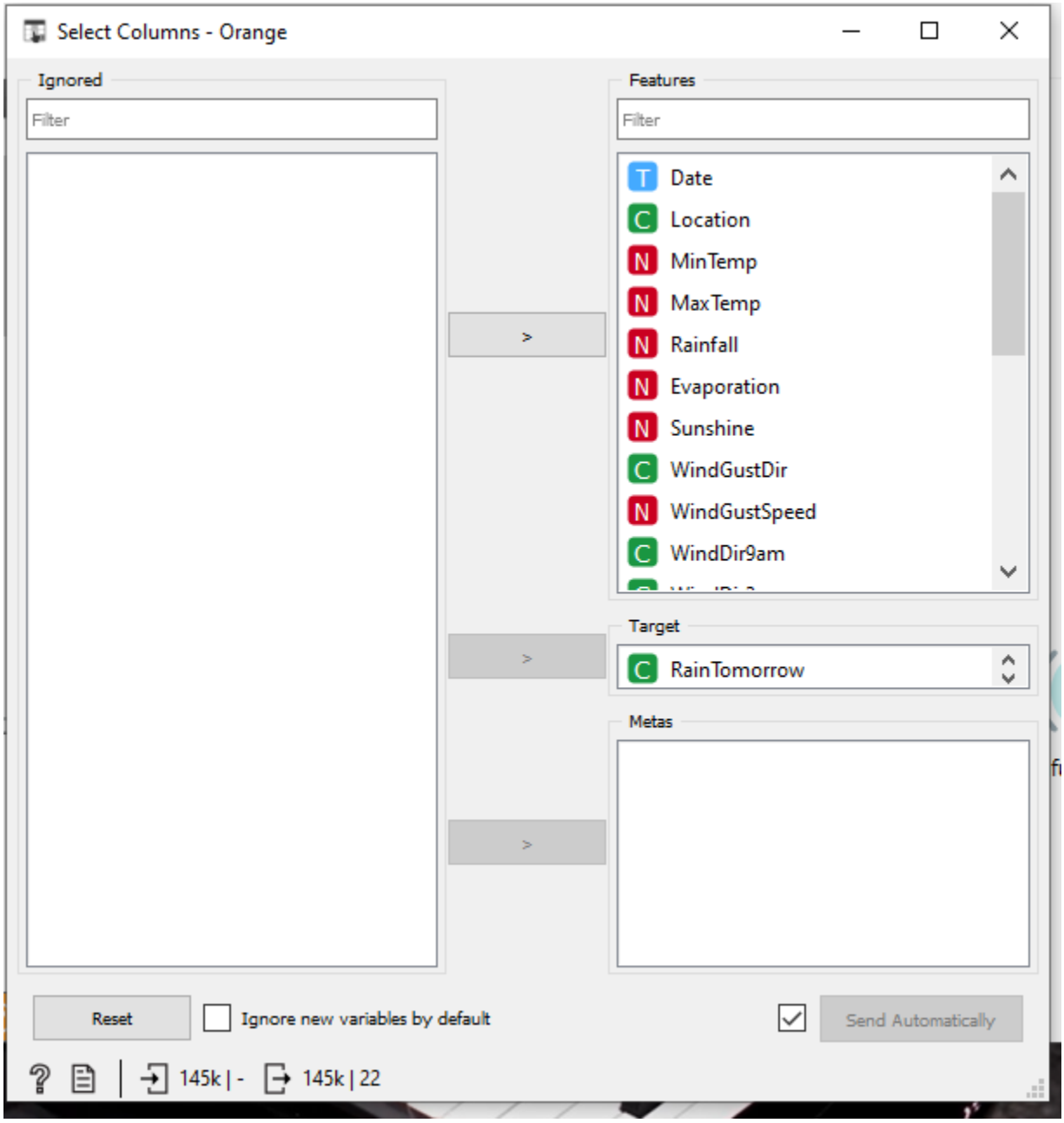
By using this data set and orange machine learning tool, results will be evaluated step-by-step.



Above picture is the process of prediction and for this random forest machine learning will be used to analyze the information. Each widget will explain in detail the following content. As a file widget, it will load the data set, in this case the downloaded data set, WeatherAUS.CSV will be loaded.

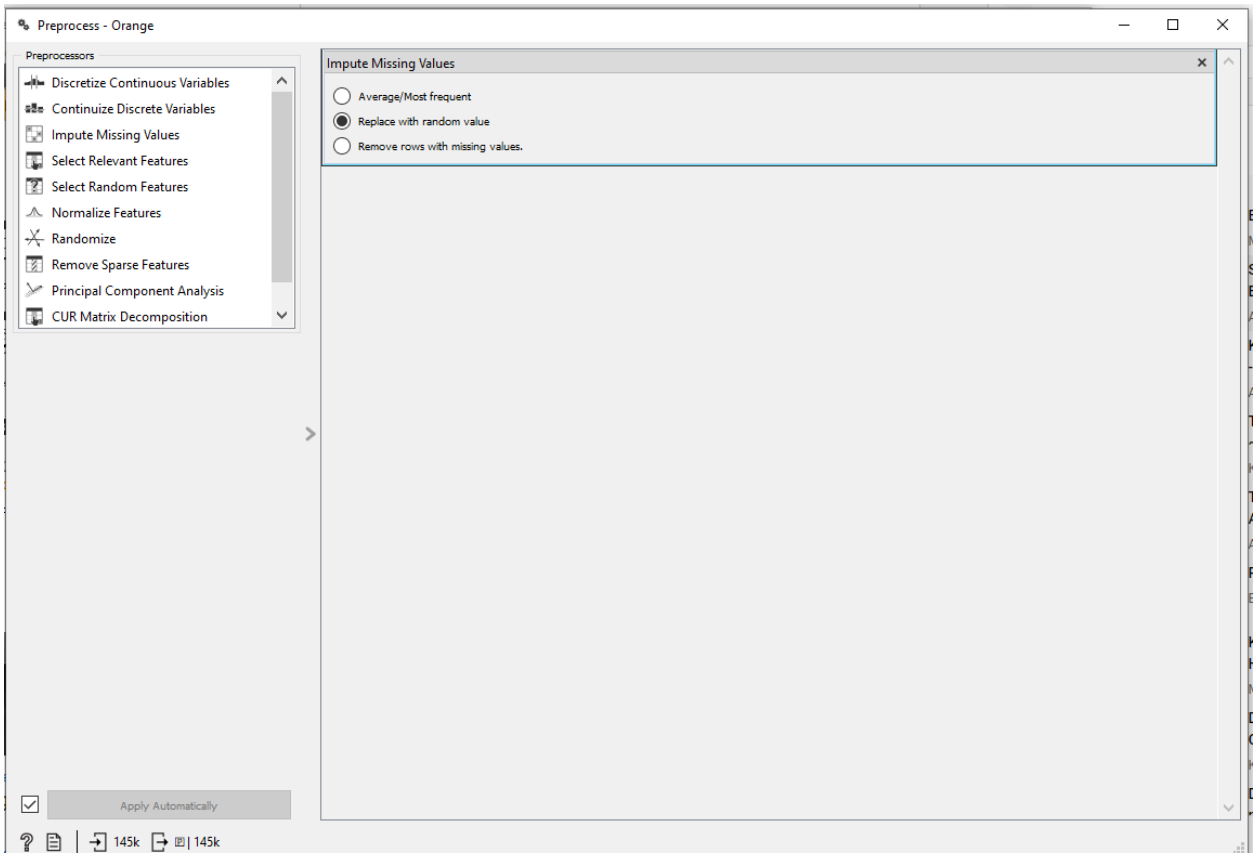
Select Column Widget

Select Column widget is to select a target column which will be Rain Tomorrow in this case.



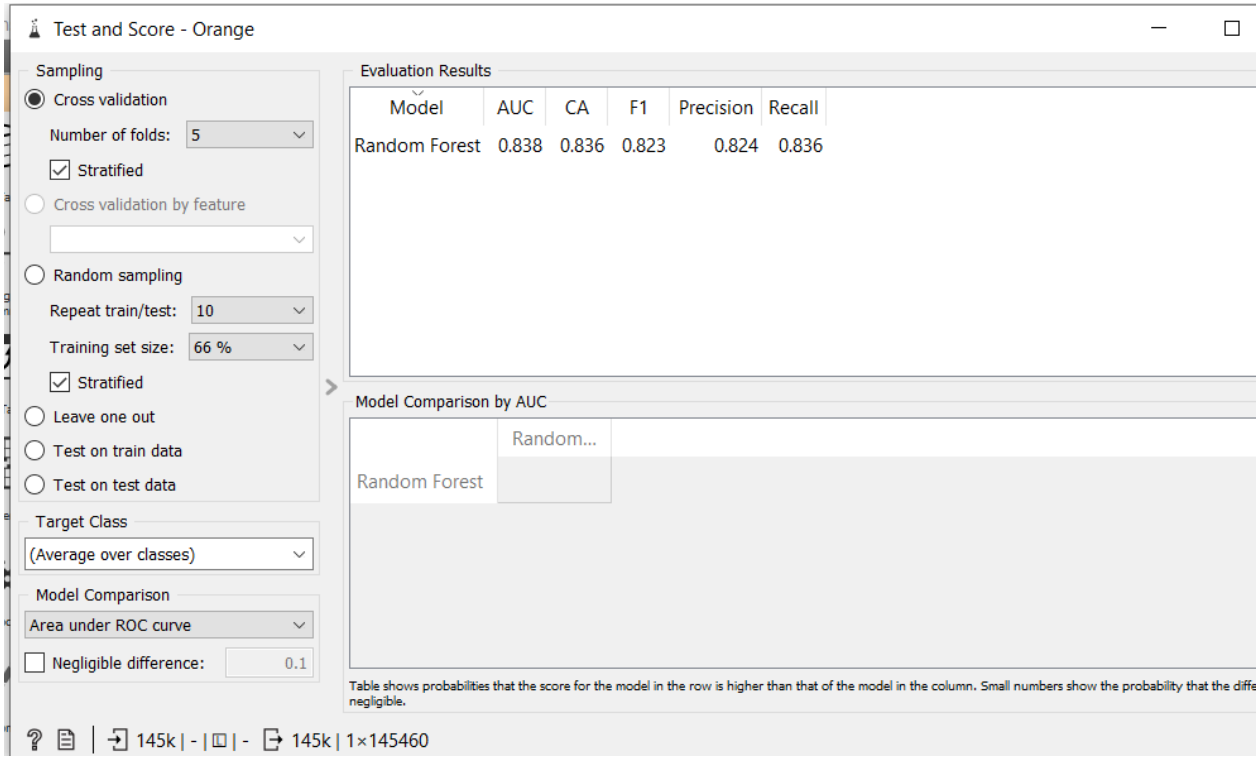
Process Widget

Process widget is that will make processing with NA (Not Available) data, for this case, it will be replaced with random value.



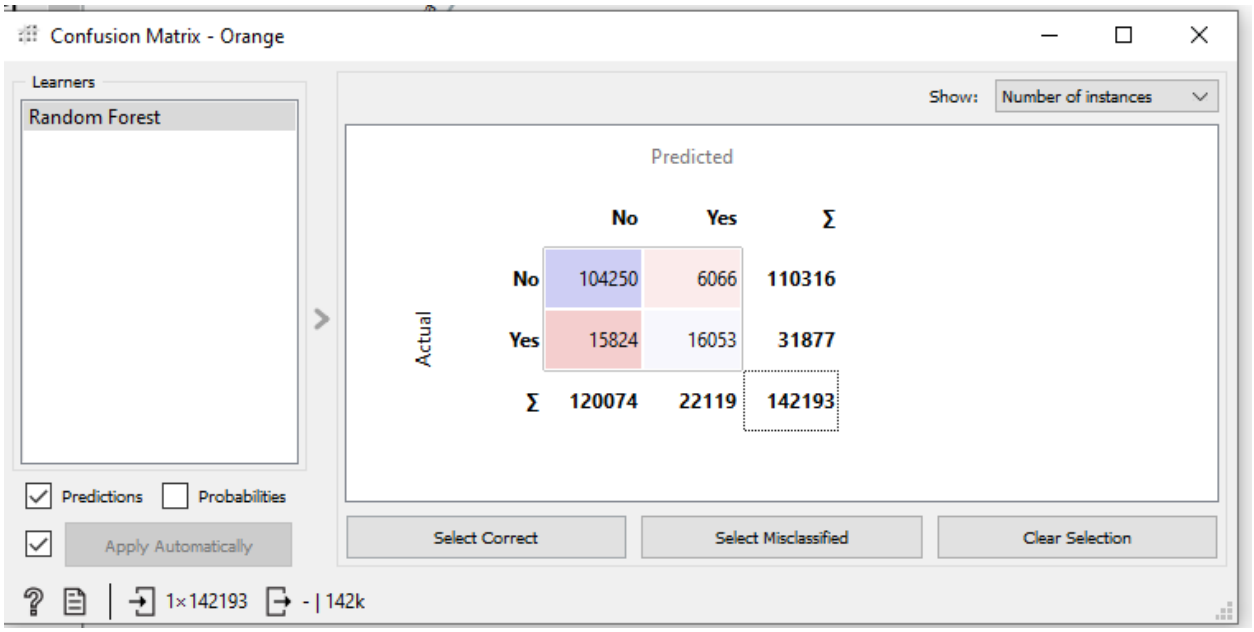
Test and Score Widget

Test and Score widget will be the result of a machine learning algorithm (Random Forest in this case) and based on a target size of 66% it will predict the data.



Confusion Matrix Widget

Confusion Matrix which is a measurement of the classification problem where output can be two or more classes. Those Confusion Matrices are the measurement of the measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC-ROC curves from the test and score widget.

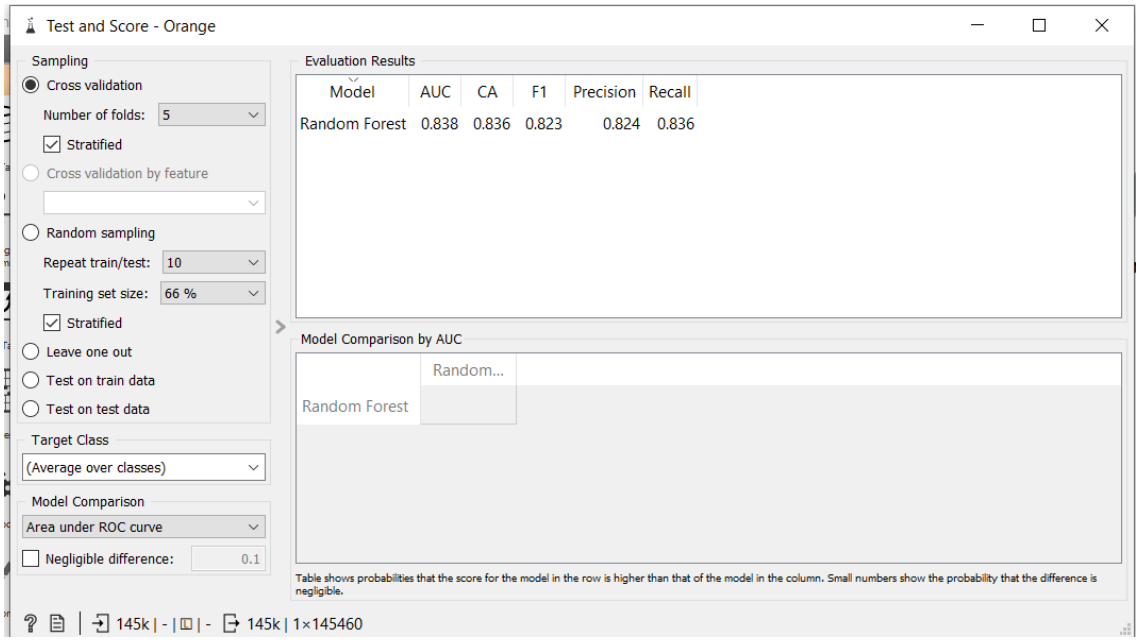


Random Forest Machine Learning Algorithm

There are many reasons for choosing this machine learning algorithm. Random forest algorithm is that it reduces overfitting in decisions that improve accuracy of prediction. It is simple and adaptable, so that it can be used for classification and regression tasks. it doesn not need hypet-parameter tweaking to get a great result. The accuracy of random forest is generally very high and its efficiency is compatible with large data sets. And it does not overfit unlike other machine learning models which means it reduces overfitting in decisions that help improve the accuracy. In addition, it automates missing values present in the data that make more fit with the WeatherAUS data set.

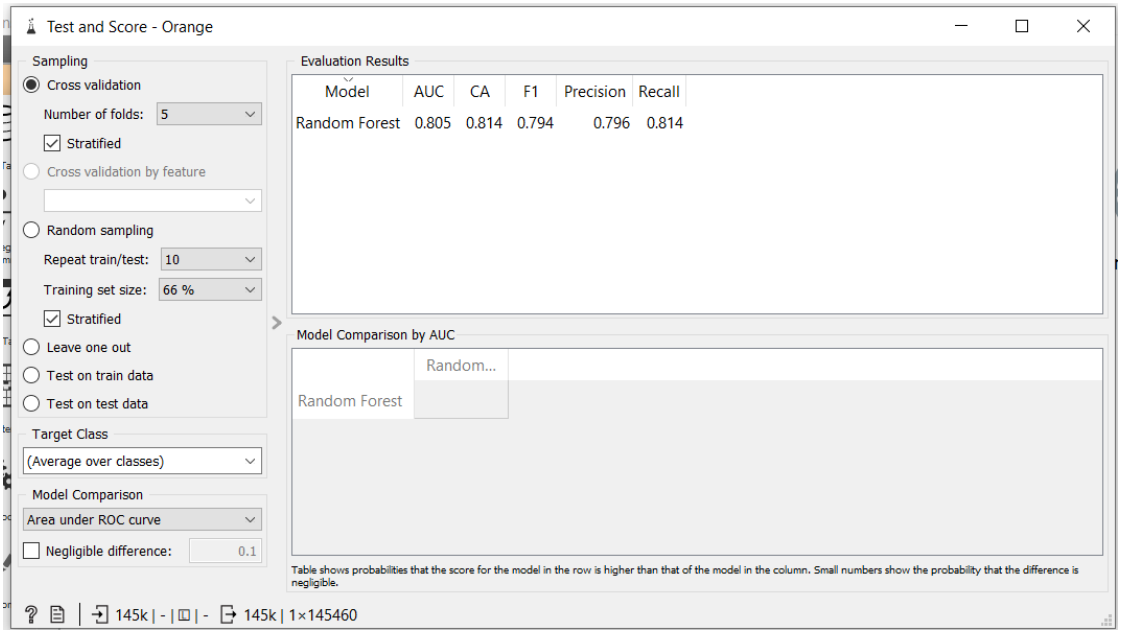
The Most Optimal Features

We have a target column: Rain Tomorrow. But we need other features (other columns) to predict the target value. However, not all columns are not necessary for the prediction. If we make little changes, the result in the test and score will be affected. Without any changes, the result of the test and score value will be as below.



Result without ignoring Humidity and Pressure

If we skip or ignore Humidity9am, Humidity3pm, Pressure9am and Pressure3pm features, the result will be slightly different as follows.



Result after ignoring Humidity and Pressure

As we saw, Humidity and Pressure is the most optimal feature for predicting Rain Tomorrow. As a percentage it can clarify 3% difference between ignoring others such as Sunshine, Evaporation, MaxTemp, Wind Speed and Wind direction which only make only a decimal(0.2) difference by ignoring.

Evaluation Technique (Precision)

Precision is the proportion of true positives that are classified as positive. Precision is calculated as the number of true positives divided by the total number of true positives and false positives(Medium, shruti, 2018). And its formula is as shown below. Those positives can be found in the confusion matrix picture above.

Precision = TruePositives / (TruePositives + FalsePositives)

Reason for choosing Precision over others is that precision does not have false negatives in its formula and it can have a huge impact in prediction. Precision also gives us a measure of the relevant data points. If there are no false positives (FPs), the model is 100 percent accurate. The more FPs which are added to the mix, the worse that precision gets.

Evaluation on my Method

The result of precision with none ignoring features is 0.824 which is equivalent to 82% which is an outstanding result. We can assume that that is the accurate result of prediction. The proportion of true positive forecasts to all positive predictions made by precision, or the accuracy of classifier predictions, as you can see. Since we do not want any false negatives ,this emphasizes that, while accuracy is beneficial, it does not reveal the entire story. It makes no mention of how many true positive class cases were incorrectly classified as negative, resulting in so-called false negatives. For our problem statement, that would be a measure of

rain tomorrow that will correctly identify having raining out of all measures data. Since we do not want a wrong classification on tomorrow's rain and also we do not want to predict wrongly that it will not rain tomorrow(false negative) as a result. In this case, evaluation of precision is more important than others.s

Conclusion

During this course work and lecture time, I gained a lot of knowledge about BIG DATA and why it is important. Moreover, along with big data, I learned that machine learning is also important and I experienced it by using the Orange machine learning tool for this course work. I am certain that even though this is my first step, this lecture will help me to extend my knowledge in Big Data.

References

Medium. 2019. *Best Data and Big Data Visualization Techniques*. [online] Available at: <https://medium.com/sciforce/best-data-and-big-data-visualization-techniques-e07b897751dd> [Accessed 13 January 2022].

2019. [online] Available at: <https://www.bigdataframework.org/four-vs-of-big-data/#:~:text=Veracity%20refers%20to%20the%20quality%20of%20the%20data%20that%20is%20being%20analyzed.&text=An%20example%20of%20a%20high%20veracity%20data%20set%20would%20be,algorithms)%20to%20reveal%20meaningful%20information.> [Accessed 13 January 2022].

Donges, N., 2021. *A Complete Guide to the Random Forest Algorithm*. [online] Built In. Available at: <https://builtin.com/data-science/random-forest-algorithm> [Accessed 14 January 2022].

Framework, B., 2022. *Five ways to capture Value from Big Data - Enterprise Big Data Framework©*. [online] Enterprise Big Data Framework©. Available at: <https://www.bigdataframework.org/value-of-big-data/> [Accessed 12 January 2022].

YourTechDiet. 2022. *7 V's of Big Data Explained (Along with Infographic)*. [online] Available at: <https://yourtechdiet.com/blogs/7vs-big-data/> [Accessed 14 January 2022].

Data Science Stack Exchange. 2019. *When is precision more important over recall?*. [online] Available at: <https://datascience.stackexchange.com/questions/30881/when-is-precision-more-important-over-recall#:~:text=When%20we%20have%20imbalanced%20class,its%20formula%2C%20which%20can%20impact.> [Accessed 20 January 2022].

Bioinformatics Laboratory, U., n.d. *Test and Score*. [online] Orangedatamining.com. Available at: <https://orangedatamining.com/widget-catalog/evaluate/testandscore/> [Accessed 17 January 2022].

JORDAN, J., 2017. *Evaluating a machine learning model.*. [online] Jeremy Jordan. Available at: <https://www.jeremyjordan.me/evaluating-a-machine-learning-model/#:~:text=It%20can%20be%20calcula

ted%20easily,the%20number%20of%20total%20predictions.&text=Precision%20is%20defined%20as%20the,belong%20in%20a%20certain%20class.> [Accessed 13 January 2022].

Includehelp.com. 2022. *Big Data Analytics – 7 V's of Big Data*. [online] Available at: <https://www.includehelp.com/big-data-analytics/7-vs-of-big-data.aspx> [Accessed 19 January 2022].

Cb2.uahs.arizona.edu. n.d. *Orange - Machine Learning Training for Health Sciences | The Center for Biomedical Informatics and Biostatistics*. [online] Available at: <https://cb2.uahs.arizona.edu/orange-machine-learning-training-health-sciences#:~:text=Orange%20is%20an%20open%2Dsource,Perform%20simple%20data%20analyses> [Accessed 17 January 2022].

W3schools.com. 2022. *SQL Joins*. [online] Available at: <https://www.w3schools.com/sql/sql_join.asp> [Accessed 20 January 2022].