.

Big Data Concepts and Machine Learning on Rainfall Prediction

CHAN JUN QI, JULIUS

Table of Contents

Table of Figures

1.0 Big Data and SQL (Part A)

1.1 The 7 V's of Big Data

The Seven V's of Big Data Analytics are Volume, Velocity, Variety, Variability, Veracity, Value, and Visualization. This framework offers a model for working with large and complex data sets. (Trigyn Technologies, 2023).

Understanding these characteristics is critical for creating a successful big data strategy capable of managing, analysing, and extracting useful business insights from enormous datasets.

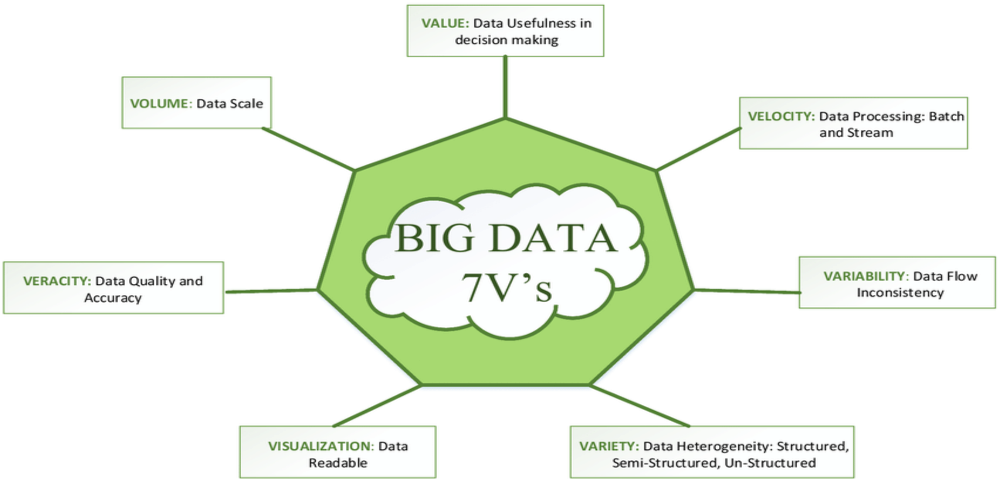

VALUE: Data Usefulness in decision making
VOLUME: Data Scale
VELOCITY: Data Processing: Batch and Stream
BIG DATA 7V's
VERACITY: Data Quality and Accuracy
VARIABILITY: Data Flow Inconsistency
VISUALIZATION: Data Readable
VARIETY: Data Heterogeneity: Structured, Semi-Structured, Un-Structured

Figure 1: 7 Vs Of Big Data (Soft Computing, 2022)

1.1.1 Volume

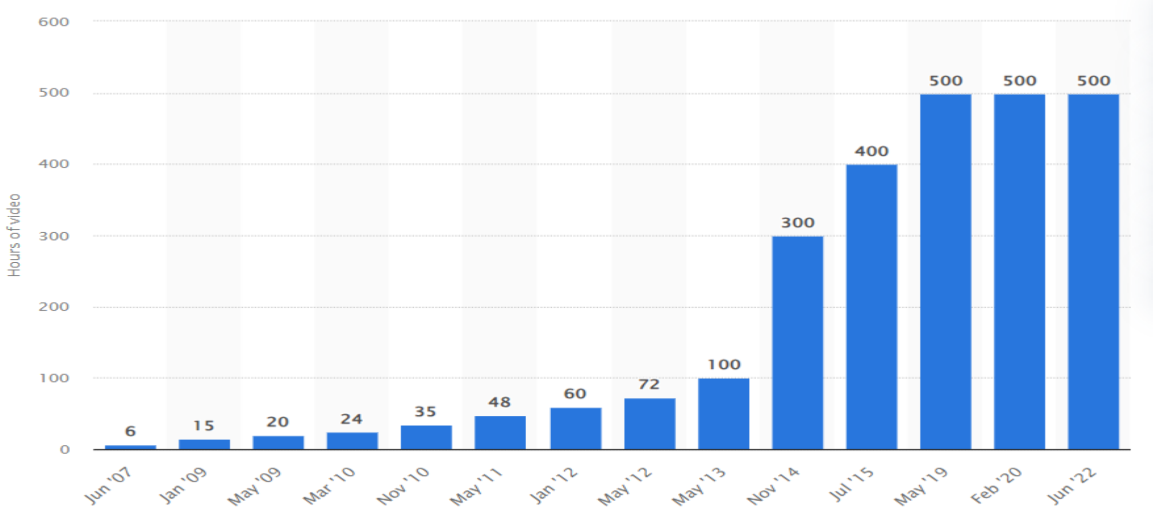


Figure 2: Hours of video uploaded to YouTube every minute 2007-2022 (Statista, 2024)

Definition: Volume in Big Data refers to the sheer size of the data. Whereas data used to be measured in gigabytes, it can now be measured in zettabytes (ZB) or even yottabytes (YB).

Example: Figure 2 shows that as of June 2022, YouTube processes over 500 hours of video uploads each minute. (Laura Ceci, 2022) That equates to 30,000 hours of newly uploaded content per hour and petabytes of new data daily.

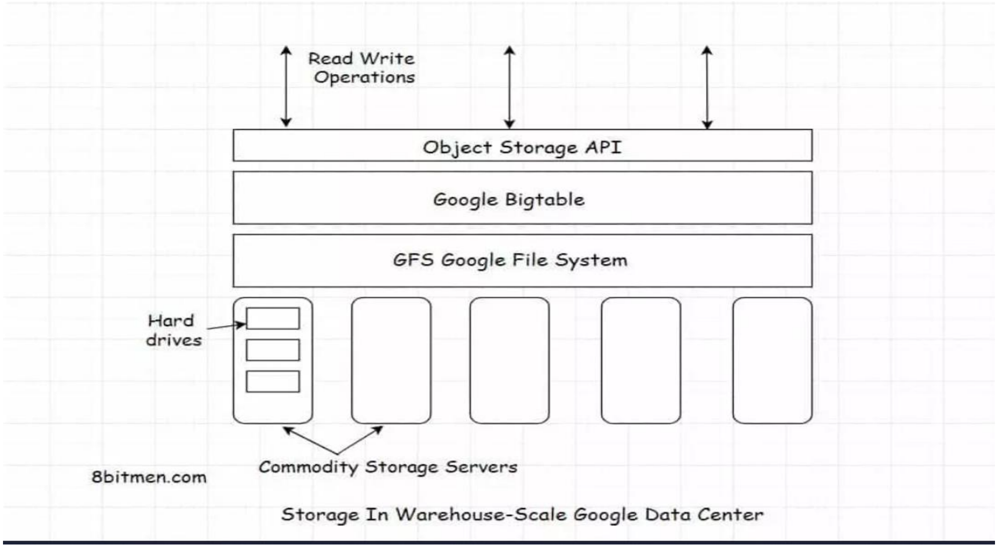


Figure 3: Warehouse-Scale Google data centers (Scale your app, n.d.)

As shown in Figure 3, YouTube stores video files on Google's File System (GFS) and BigTable, while metadata and user preferences are saved in MySQL databases. This method maintains content and user information effectively, allowing millions of users throughout the world to have personalized experiences.

1.1.2 Velocity

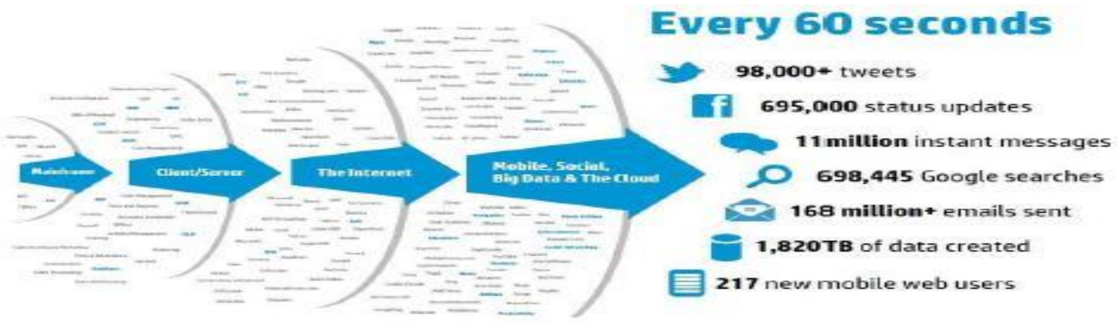


Figure 4: Velocity of Big Data Sources (ISU Corp, n.d.)

Definition: Velocity in Big Data refers to the speed at which data is generated, collected, and processed. As shown in Figure 4, examples include social media posts and emails sent in thousands per second should be accessible as soon as possible.

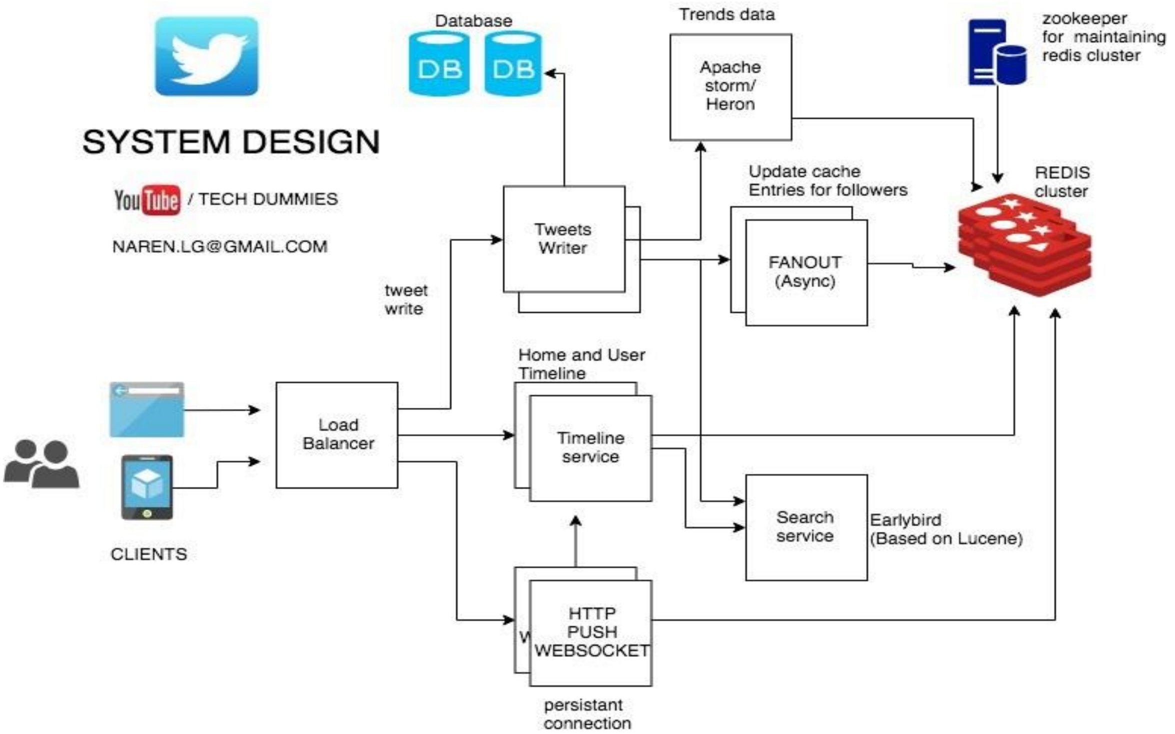


Figure 5: Twitter System Design (Narendra Lakshmana, 2018)

Example: As shown in Figure 5, Twitter uses Apache Storm for real-time data processing, trending hashtags based on tweet volume seconds after they appear. Apache Storm effectively manages massive amounts of high-speed data, allowing for real-time analytics.

1.1.3 Variety

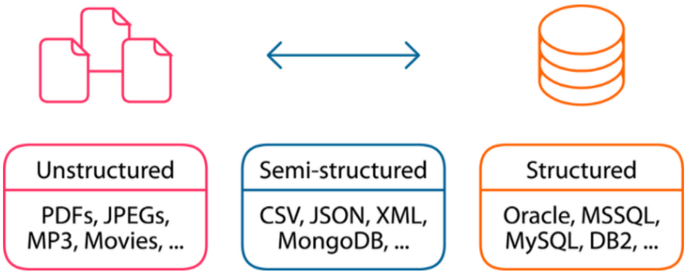

Unstructured
PDFs, JPEGs, MP3, Movies, ...
Semi-structured
CSV, JSON, XML, MongoDB, ...
Structured
Oracle, MSSQL, MySQL, DB2, ...

Figure 6: Different types of data (IET Digital Library, n.d.)

Definition: Variety in Big Data refers to the different types of data such as structured, semi-structured, and unstructured. Unstructured data, such as audio files, videos, and photos, is the most common in today's world due to social media.

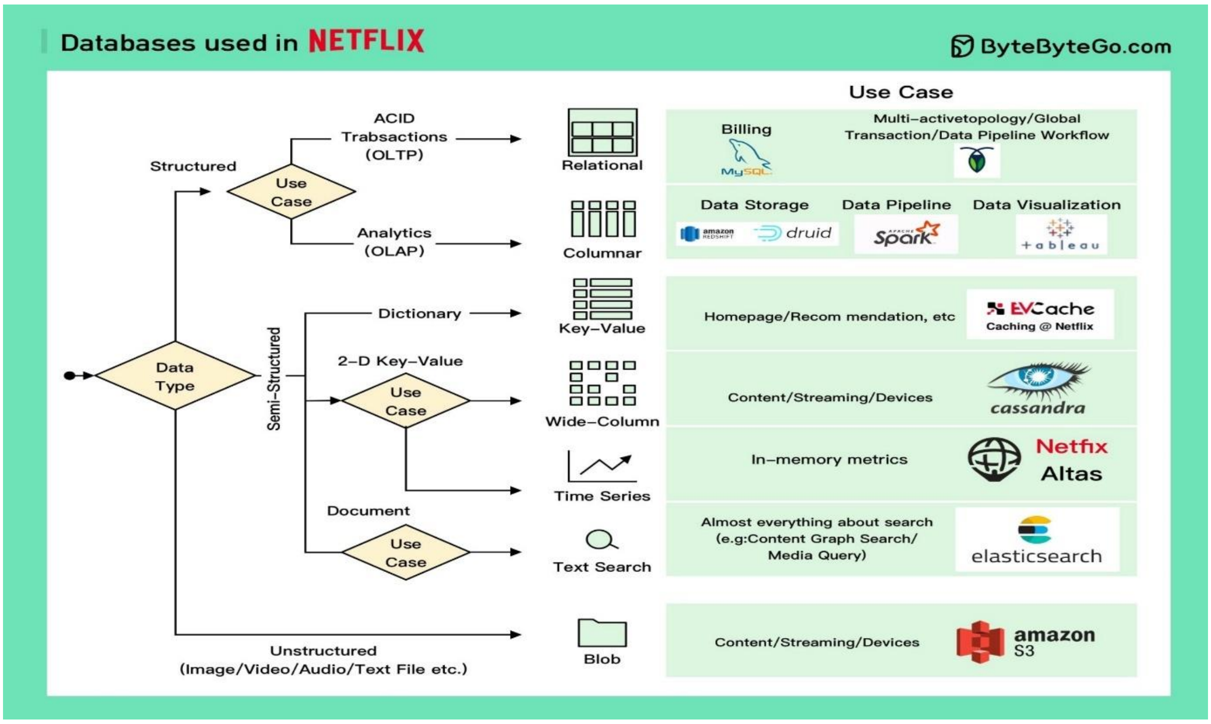


Figure 7: Netflix Tech Stack - Database (Alex Xu, 2023)

Example: As shown in Figure 7, Netflix uses relational databases and ACID transactions to manage structured data such as watching times, durations, social interactions, search queries, and stream metadata. Unstructured data, such as thumbnails and multimedia, is saved as BLOBs (Binary Large Objects), ensuring excellent data integrity and efficient administration of varied data types.

1.1.4 Veracity

Definition: Veracity in Big Data relates to the quality and reliability of the data, which is all about ensuring that the data gathered is accurate and keeping erroneous data out of the system. Bugs, human error, biasedness, and noise are factors that might have an impact on the reliability of data.

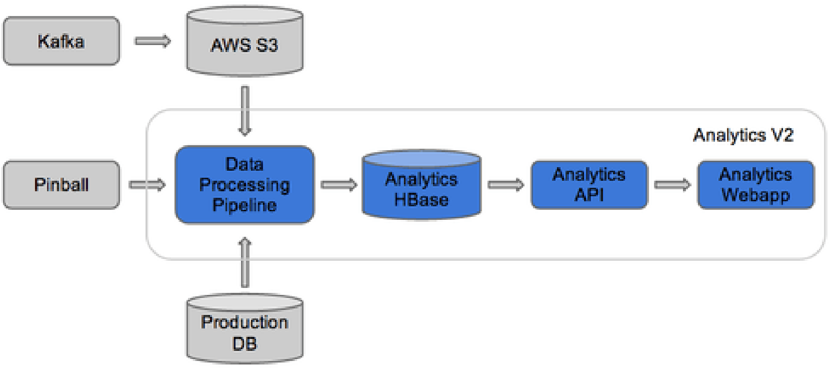


Figure 8: Pinterest Data Architecture (Ore Otegbade, 2020)

Example: As shown in Figure 8, Pinterest maintains data integrity by managing big data volumes and giving reliable insights to advertisers. It uses technologies such as Apache Kafka, Amazon S3, and HBase to maintain the stability and integrity of its data architecture.

1.1.5 Visualization

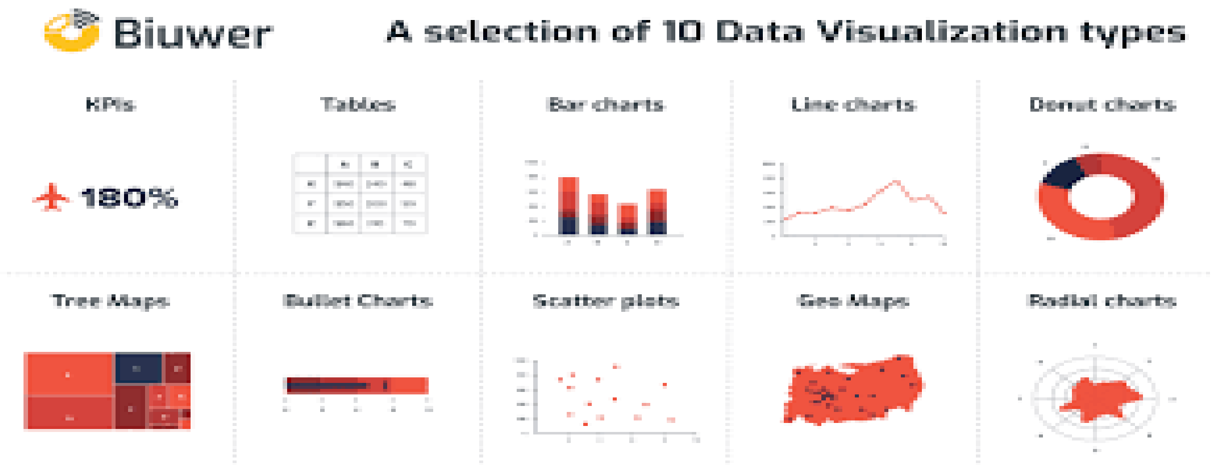


Figure 9: Data Visualization types (Biuwer, 2020)

Definition: Visualization refers to presenting data to executives for decision-making objectives. The data may be presented using graphical charts in a variety of forms.

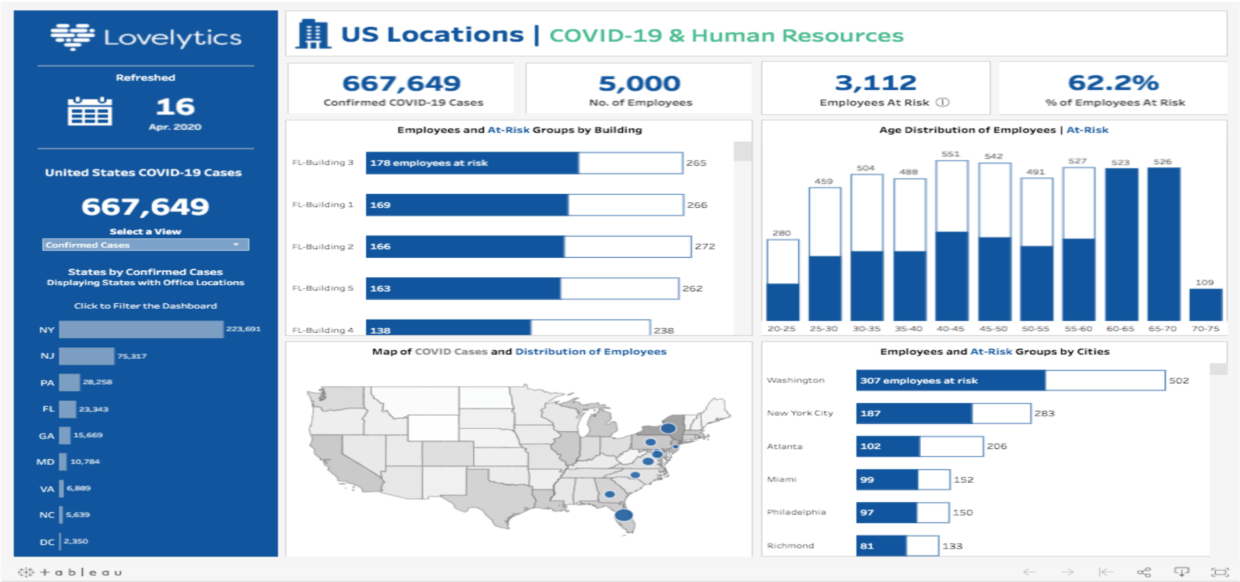


Figure 10: Tableau Dashboard (Tableau, 2020)

Example: As shown in Figure 9, Tableau provides a collection of data visualization tools that allow users to build interactive and shared dashboards, reports, and charts from big data sets. Its straightforward UI and robust analytics features making it a popular choice for businesses and organizations seeking to get insights from their data.

1.1.6 Variability

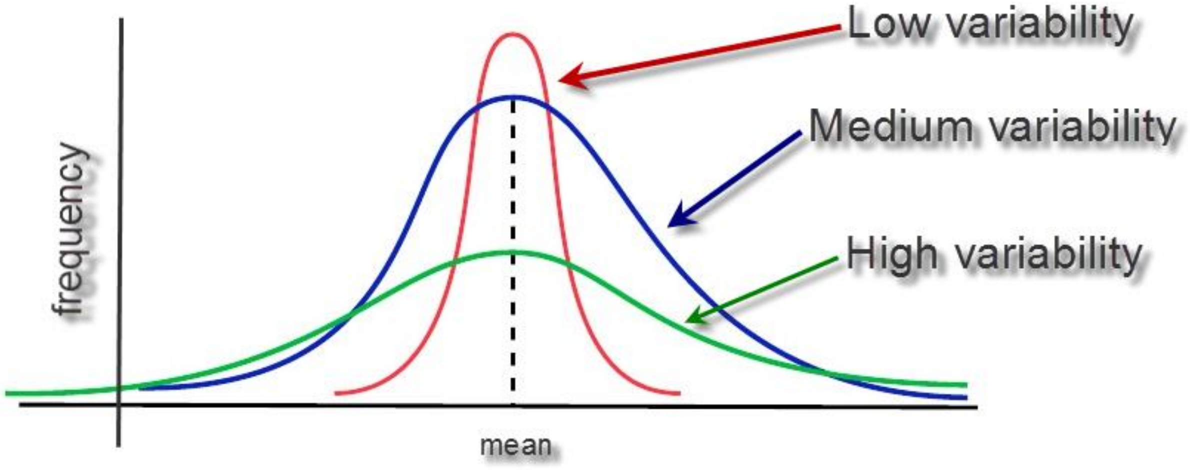


Figure 11: Variability Metrics (Mohsin Shaikh, 2023)

Definition: Variability in Big Data refers to inconsistencies or spread out in data flow, such as fluctuations and quality issues in the data source, demanding an understanding and interpretation of raw data's accurate meanings.

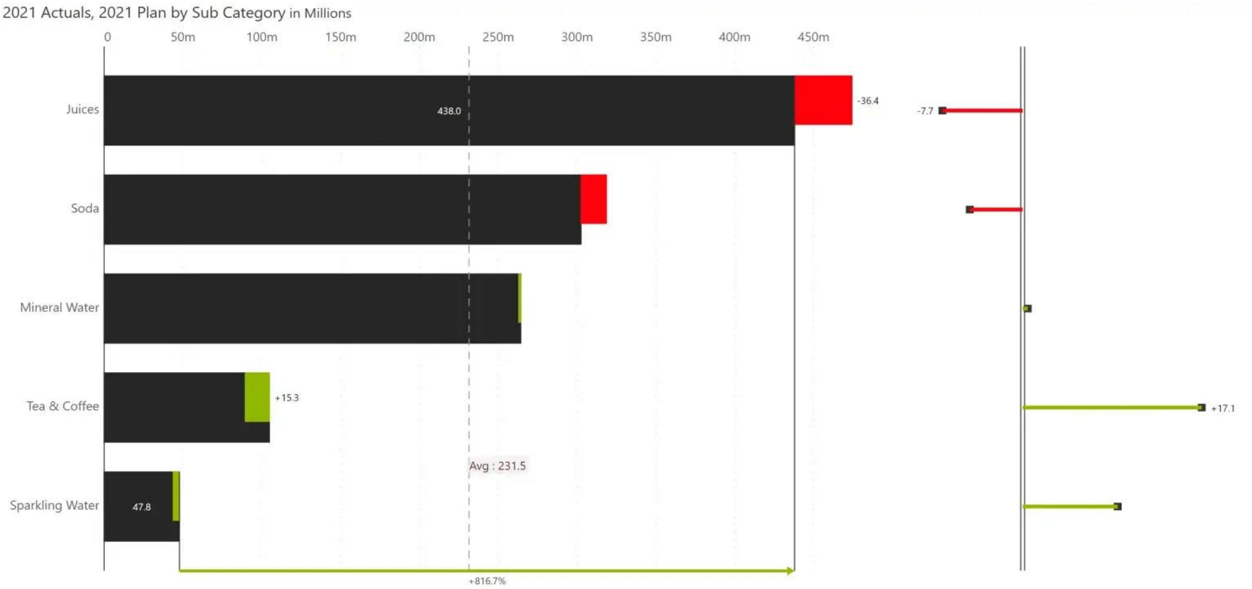


Figure 12: Inforiver Chart in PowerBI (Inforiver, 2022)

Example: Power BI can monitor and analyze variability in large data, allowing organizations to acquire insights into trends and patterns and then make informed choices based on these insights.

As shown in Figure 10, Inforiver Charts for Power BI provide users the capability to compare performance and benchmark with variance visualized in both absolute & % terms. (Inforiver, 2022) .

1.1.7 Value

Definition: Value in Big Data refers to the ability to turn the data into value for the organization. Every user must recognize that the company needs value after efforts and resources are expended on the aforementioned 7Vs of big data.

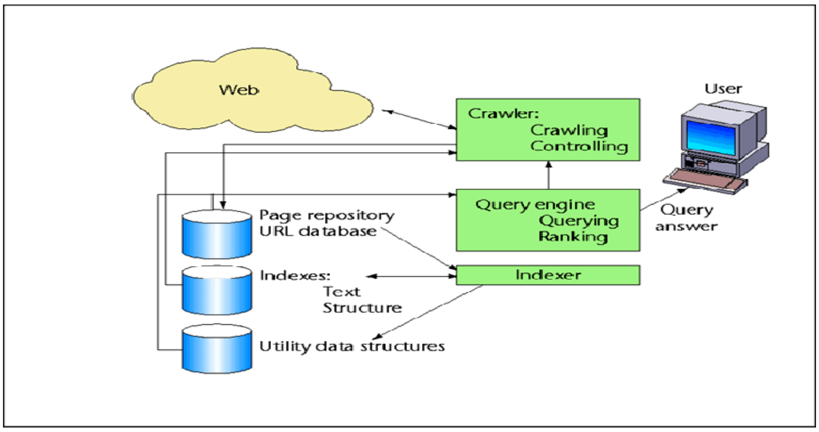


Figure 13: Search Query Database (Vikas Honmane, 2019)

Example: Google leverages big data from its Web index to first match requests with potentially valuable results, highlighting big data's importance in making business decisions. It uses machine-learning algorithms to analyze data credibility and rank the sites appropriately. (Santosh D, 2022)

1.2 The Four (4) Types of SQL's Table Joins

There are four types of SQL Table join: Inner Join, Left Outer Join, Right Outer Join, and Full Outer Join. These Join operations join data or rows from two or more tables that have a common field.

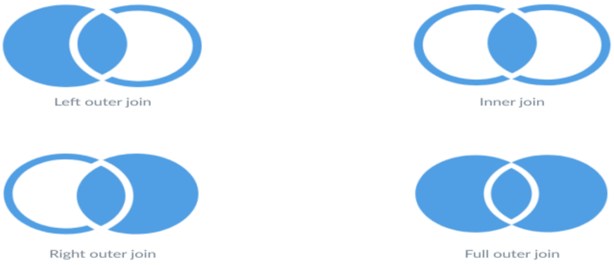
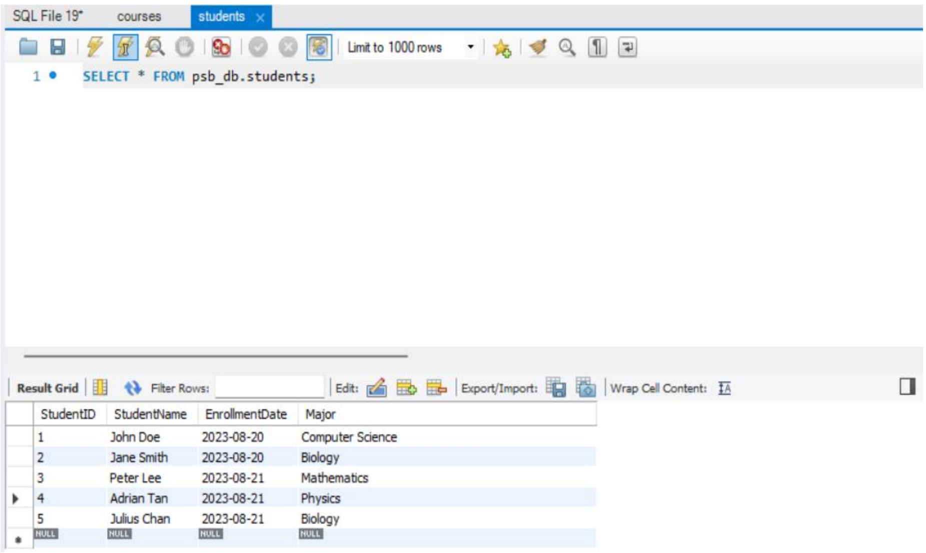


Figure 14: SQL Join Types (Metabase, n.d.)

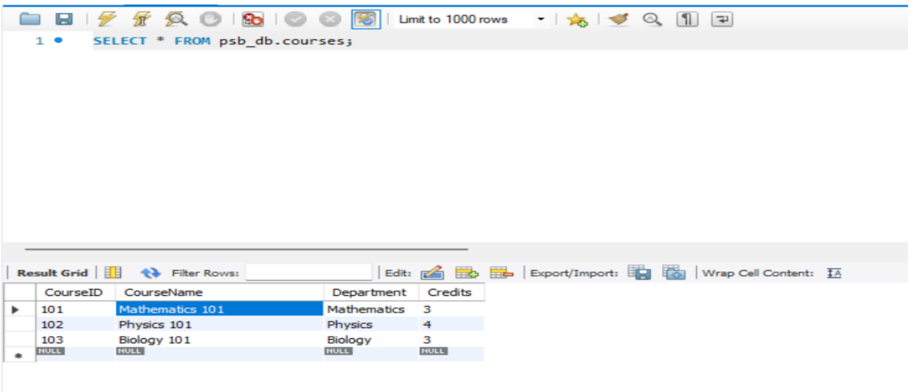
A database called "psb_db" has been created, containing the "students" and "courses" tables, to illustrate the join operations.



```
SELECT * FROM psb_db.students;
```

StudentID	StudentName	EnrollmentDate	Major
1	John Doe	2023-08-20	Computer Science
2	Jane Smith	2023-08-20	Biology
3	Peter Lee	2023-08-21	Mathematics
4	Adrian Tan	2023-08-21	Physics
5	Julius Chan	2023-08-21	Biology
NULL	NULL	NULL	NULL

Figure 15: Students Table (Julius Chan, 2024)

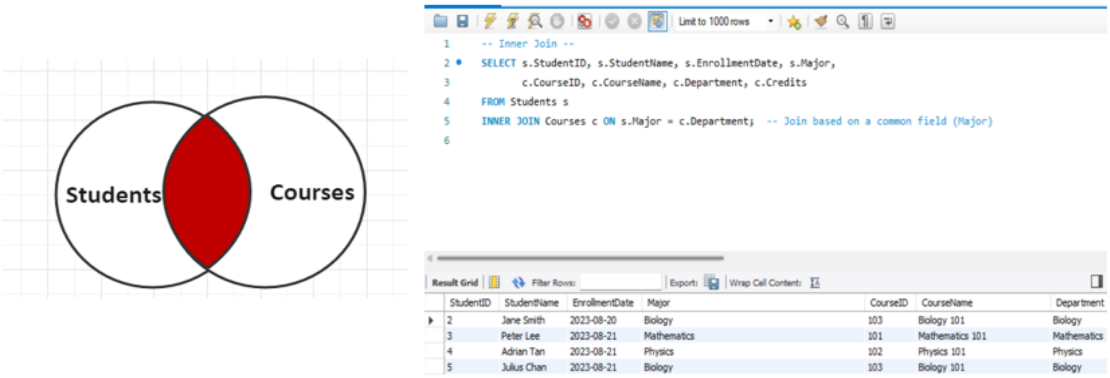


```
SELECT * FROM psb_db.courses;
```

CourseID	CourseName	Department	Credits
101	Mathematics 101	Mathematics	3
102	Physics 101	Physics	4
103	Biology 101	Biology	3
NULL	NULL	NULL	NULL

Figure 16: Courses Table (Julius Chan, 2024)

1.2.1 Inner Join



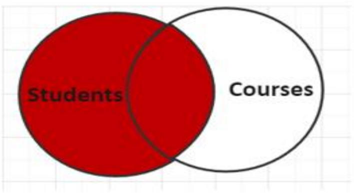
```
-- Inner Join --
SELECT s.StudentID, s.StudentName, s.EnrollmentDate, s.Major,
       c.CourseID, c.CourseName, c.Department, c.Credits
FROM Students s
INNER JOIN Courses c ON s.Major = c.Department;  -- Join based on a common field (Major)
```

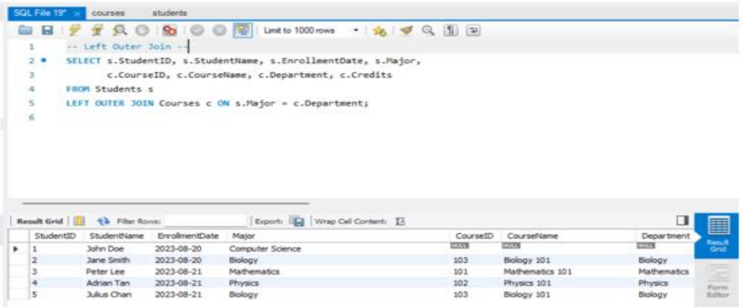
StudentID	StudentName	EnrollmentDate	Major	CourseID	CourseName	Department
2	Jane Smith	2023-08-20	Biology	103	Biology 101	Biology
3	Peter Lee	2023-08-21	Mathematics	101	Mathematics 101	Mathematics
4	Adrian Tan	2023-08-21	Physics	102	Physics 101	Physics
5	Julius Chan	2023-08-21	Biology	103	Biology 101	Biology

Figure 17: SQL Table, Code & Inner Join Diagram (Julius Chan, 2024)

As shown in Figure 17, Inner Join only returns information for students who are presently enrolled in at least one course. The join condition is based on a matching field between the Students and Courses tables, such as Major. The query would return data for students with a matching major in the Courses table.

1.2.2 Left Outer Join





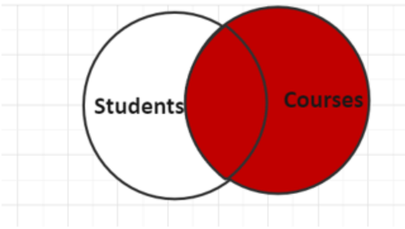
```
-- Left Outer Join --
SELECT s.StudentID, s.StudentName, s.EnrollmentDate, s.Major,
       c.CourseID, c.CourseName, c.Department, c.Credits
FROM Students s
LEFT OUTER JOIN Courses c ON s.Major = c.Department;
```

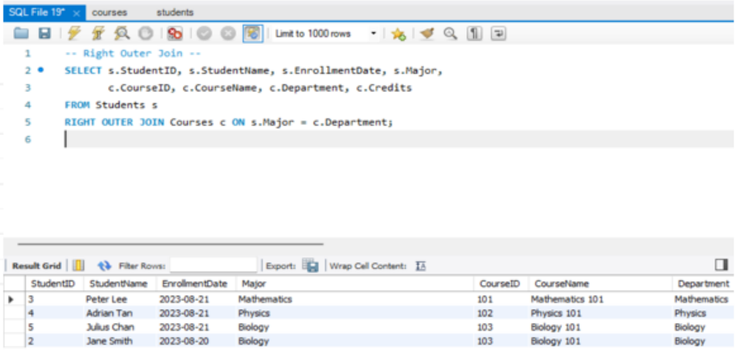
StudentID	StudentName	EnrollmentDate	Major	CourseID	CourseName	Department
1	John Doe	2023-08-20	Computer Science	NULL	NULL	NULL
2	Jane Smith	2023-08-20	Biology	103	Biology 101	Biology
3	Peter Lee	2023-08-21	Mathematics	101	Mathematics 101	Mathematics
4	Adrian Tan	2023-08-21	Physics	102	Physics 101	Physics
5	Julius Chan	2023-08-21	Biology	103	Biology 101	Biology

Figure 18: SQL Table, Code & Left Outer Join Diagram (Julius Chan, 2024)

As shown in Figure 18, the left outer join generates a student list by combining data, including those not enrolled in any courses. The student's table is the left table, and the course table is the right table. Unmatched students will have NULL values for course details. The query includes all students, highlighting both enrolled and not enrolled in courses.

1.2.3 Right Outer Join





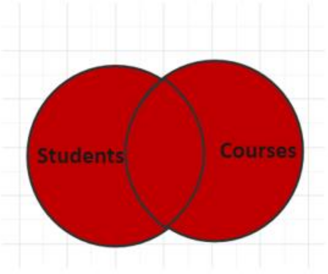
```
-- Right Outer Join --
SELECT s.StudentID, s.StudentName, s.EnrollmentDate, s.Major,
       c.CourseID, c.CourseName, c.Department, c.Credits
FROM Students s
RIGHT OUTER JOIN Courses c ON s.Major = c.Department;
```

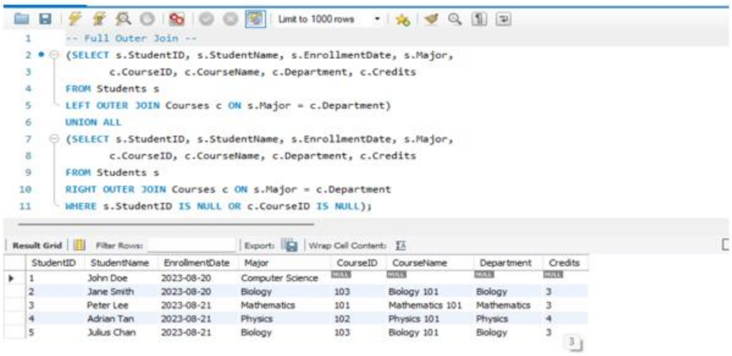
StudentID	StudentName	EnrollmentDate	Major	CourseID	CourseName	Department
3	Peter Lee	2023-08-21	Mathematics	101	Mathematics 101	Mathematics
4	Adrian Tan	2023-08-21	Physics	102	Physics 101	Physics
5	Julius Chan	2023-08-21	Biology	103	Biology 101	Biology
2	Jane Smith	2023-08-20	Biology	103	Biology 101	Biology

Figure 19: SQL Table, Code & Right Outer Join Diagram (Julius Chan, 2024)

As shown in Figure 19, a right outer join joins all entries from the right table with the corresponding value from the left table. This join type may be less useful in the student and course tables because it displays all courses available, even if no students are presently enrolled. This method may not be as useful as the two preceding join actions.

1.2.4 Full Outer Join





Limit to 1000 rows

```sql
-- Full Outer Join --
(SELECT s.StudentID, s.StudentName, s.EnrollmentDate, s.Major,
        c.CourseID, c.CourseName, c.Department, c.Credits
FROM Students s
LEFT OUTER JOIN Courses c ON s.Major = c.Department)
UNION ALL
(SELECT s.StudentID, s.StudentName, s.EnrollmentDate, s.Major,
        c.CourseID, c.CourseName, c.Department, c.Credits
FROM Students s
RIGHT OUTER JOIN Courses c ON s.Major = c.Department
WHERE s.StudentID IS NULL OR c.CourseID IS NULL);
```

StudentID	StudentName	EnrollmentDate	Major	CourseID	CourseName	Department	Credits
1	John Doe	2023-08-20	Computer Science	NULL	NULL	NULL	NULL
2	Jane Smith	2023-08-20	Biology	103	Biology 101	Biology	3
3	Peter Lee	2023-08-21	Mathematics	101	Mathematics 101	Mathematics	3
4	Adrian Tan	2023-08-21	Physics	102	Physics 101	Physics	4
5	Julius Chan	2023-08-21	Biology	103	Biology 101	Biology	3

Figure 20: SQL Table, Code & Right Outer Join Diagram (Julius Chan, 2024)

As shown in Figure 20, Full Outer Join combines the results of separate Left and Right Outer Join queries, ensuring that all data from both tables is included, with null values indicating unmatched records. This operation generates a complete report that consists of all students and courses, regardless of enrolment status. Unenrolled students will have null values for course details, whereas courses without enrolled students will have null values for student information.

2.0 Rainfall Prediction using Orange (Part B)

2.1 Introduction to Orange

Orange has been selected as the machine learning tool as it is a strong, user-friendly open-source machine learning and data visualization platform that has grown popular within the data science community.

In 1996, the Bioinformatics Laboratory of the Faculty of Computer and Information Science, University of Ljubljana, Slovenia, developed Orange in cooperation with an open-source community. (Meta Brown, 2016).

The application was designed to be an accessible data mining and analysis platform, appealing to beginners and experts.

2.1.2 Technical Features

Visual Programming: No coding is required. Orange's visual programming interface allows users to create data pipelines by dragging and dropping widgets.

Interactive Data Visualization: The widgets accept data from the input and send out filtered or processed data, models, or whatever the widget performs on the output.

Extensions: Allows specialized add-ons that can do natural language processing and text mining, network analysis, association rule mining, or handle fairness in machine learning.

2.1.3 Usage for Machine Learning

Exploratory Data Analysis: With Orange's visualization capabilities, users may quickly discover insights and trends in their data.

Model Training and Results: Orange facilitates the training and evaluation of predictive models by allowing users to experiment with multiple algorithms and parameter settings to see which models perform best.

Workflow Design Interface: Users may automate repetitive operations using the drag-and-drop interface, speeding the data analysis process and increasing efficiency.

2.2: Rainfall Prediction using Neural Network, Logistic Regression and Random Forest

2.2.1 Objective

The objective is to predict rainfall accurately for the following day. This prediction is vital for planning daily activities and mitigating the impact of unexpected weather events. Using Orange's machine learning tools and data visualization features, more precise predictive models can be developed to improve forecast accuracy, enabling better decision-making and preparation for weather-related challenges.

2.2.2 Expected Outcomes

Binary Classification: The forecast is usually a binary classification issue, with the goal variable indicating whether it will rain tomorrow (RainTomorrow).

Performance indicators: Key performance indicators for assessing the model may include model accuracy.

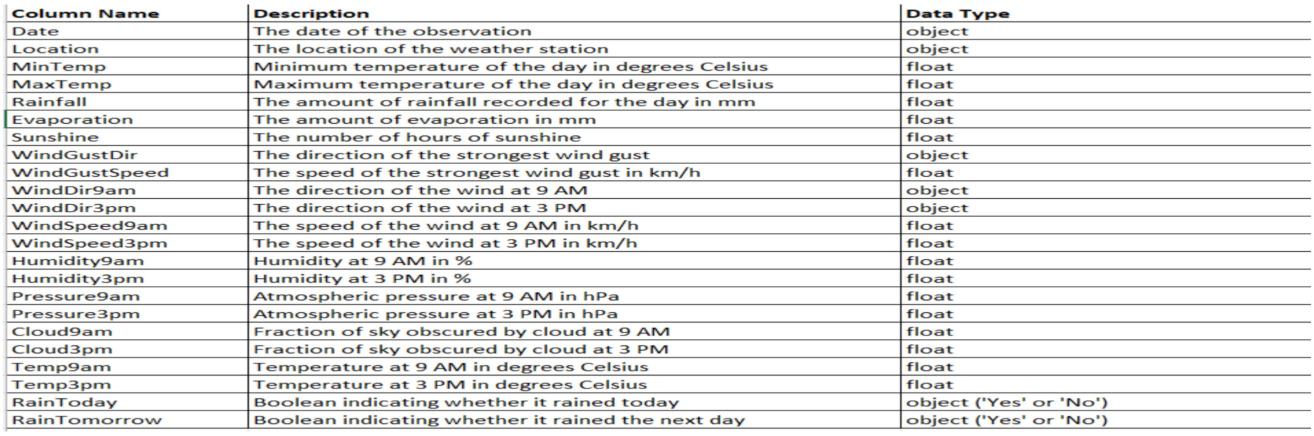
Relevant Features: Beyond prediction, the model may reveal which features are most significant in forecasting rainfall.

2.2.3 Dataset: weatherAUS

Source: The dataset is sourced from the Australian Bureau of Meteorology and comprises around ten years of daily weather observations from various places throughout Australia. It is accessible and acquired on Kaggle.

Data Structures: The dataset is a two-dimensional table with each row representing a daily weather observation from a specific place and each column representing an individual component of the weather data, such as numerical and category values.

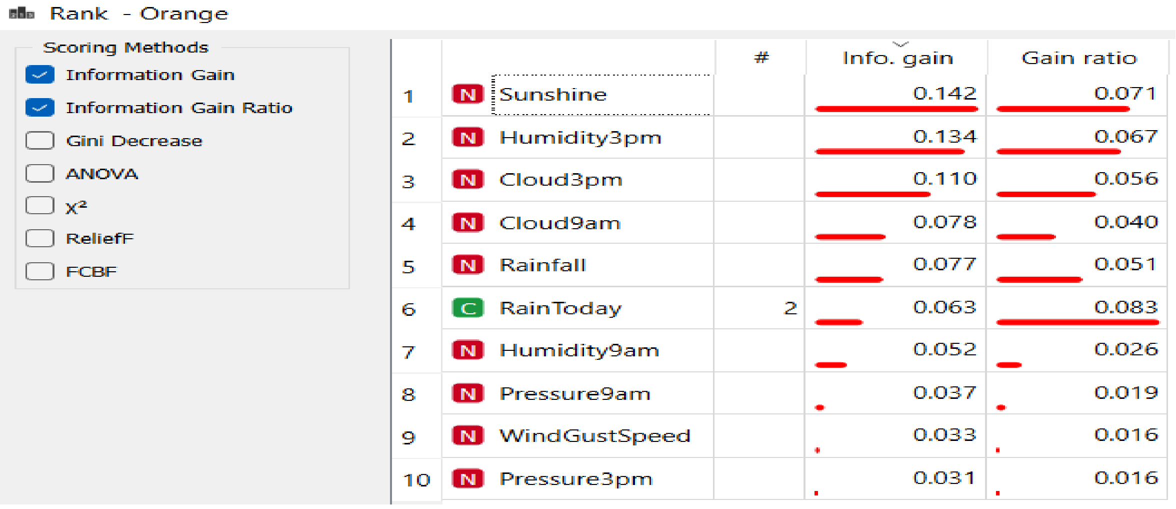
Data Labels & Types: A breakdown of data labels and types, as shown in Figure 21



Column Name	Description	Data Type
Date	The date of the observation	object
Location	The location of the weather station	object
MinTemp	Minimum temperature of the day in degrees Celsius	float
MaxTemp	Maximum temperature of the day in degrees Celsius	float
Rainfall	The amount of rainfall recorded for the day in mm	float
Evaporation	The amount of evaporation in mm	float
Sunshine	The number of hours of sunshine	float
WindGustDir	The direction of the strongest wind gust	object
WindGustSpeed	The speed of the strongest wind gust in km/h	float
WindDir9am	The direction of the wind at 9 AM	object
WindDir3pm	The direction of the wind at 3 PM	object
WindSpeed9am	The speed of the wind at 9 AM in km/h	float
WindSpeed3pm	The speed of the wind at 3 PM in km/h	float
Humidity9am	Humidity at 9 AM in %	float
Humidity3pm	Humidity at 3 PM in %	float
Pressure9am	Atmospheric pressure at 9 AM in hPa	float
Pressure3pm	Atmospheric pressure at 3 PM in hPa	float
Cloud9am	Fraction of sky obscured by cloud at 9 AM	float
Cloud3pm	Fraction of sky obscured by cloud at 3 PM	float
Temp9am	Temperature at 9 AM in degrees Celsius	float
Temp3pm	Temperature at 3 PM in degrees Celsius	float
RainToday	Boolean indicating whether it rained today	object ('Yes' or 'No')
RainTomorrow	Boolean indicating whether it rained the next day	object ('Yes' or 'No')

Figure 21: 'WeatherAUS' Data Types & Labels (Julius Chan, 2024)

2.2.4 Data-Preprocessing



Rank - Orange

Scoring Methods

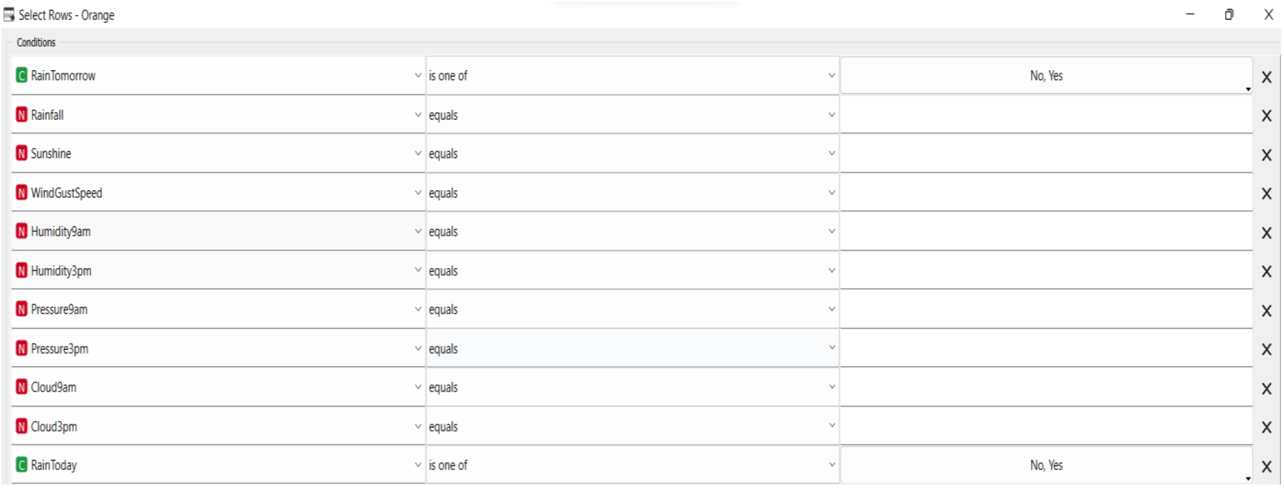
☑ Information Gain
☑ Information Gain Ratio
☐ Gini Decrease
☐ ANOVA
☐ x^2
☐ ReliefF
☐ FCBF

			#	Info. gain	Gain ratio
1	N	Sunshine		0.142	0.071
2	N	Humidity3pm		0.134	0.067
3	N	Cloud3pm		0.110	0.056
4	N	Cloud9am		0.078	0.040
5	N	Rainfall		0.077	0.051
6	C	RainToday	2	0.063	0.083
7	N	Humidity9am		0.052	0.026
8	N	Pressure9am		0.037	0.019
9	N	WindGustSpeed		0.033	0.016
10	N	Pressure3pm		0.031	0.016

Figure 22: Ranking of Features (Julius Chan, 2024)

The Rank Widget shown in Figure 22 helps to identify on what was the best 10 features based on Information Gain and Gain Ratio Score. The features are Humidity3pm, Sunshine, Cloud3pm, Cloud9am, Rainfall, RainToday, Humidity9am, Pressure9am, WindGustSpeed, and Pressure3pm.

Both Information Gain and Gain Ratio are significant indicators in the Rank widget. Information Gain estimates the reduction in uncertainty about the target variable (RainTomorrow) provided by knowing a feature, whereas Gain Ratio normalizes this metric by including the feature's related information.



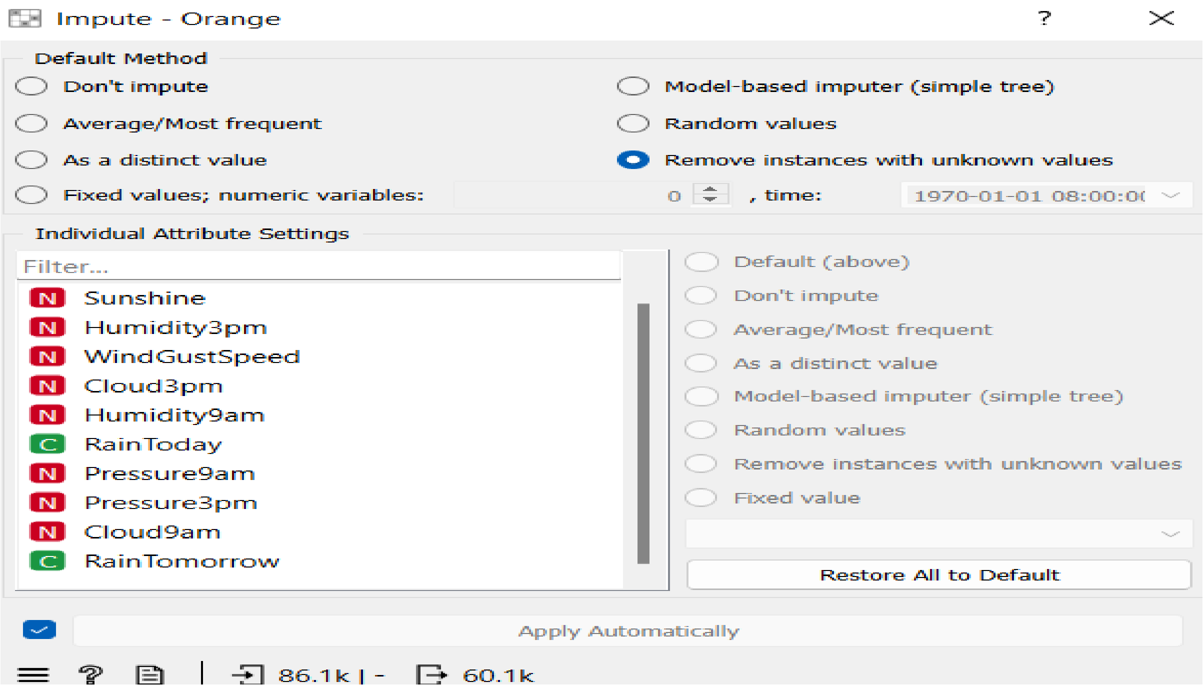
Select Rows - Orange

Conditions

C RainTomorrow	is one of	No, Yes
N Rainfall	equals	
N Sunshine	equals	
N WindGustSpeed	equals	
N Humidity9am	equals	
N Humidity3pm	equals	
N Pressure9am	equals	
N Pressure3pm	equals	
N Cloud9am	equals	
N Cloud3pm	equals	
C RainToday	is one of	No, Yes

Figure 23: Feature Selection for Data Pre-Processing (Julius Chan, 2024)

Figure 23 shows a pre-processing stage. The purpose is to choose the highly correlated features, as shown in Figure 22, thus these features improve the model's accuracy and efficiency by obtaining accurate information that predicts rainfall.



Impute - Orange

Default Method

○ Don't impute
○ Average/Most frequent
○ As a distinct value
○ Fixed values; numeric variables: 0, time: 1970-01-01 08:00:0(
○ Model-based imputer (simple tree)
○ Random values
● Remove instances with unknown values

Individual Attribute Settings

Filter...

- N Sunshine
- N Humidity3pm
- N WindGustSpeed
- N Cloud3pm
- N Humidity9am
- C RainToday
- N Pressure9am
- N Pressure3pm
- N Cloud9am
- C RainTomorrow

○ Default (above)
○ Don't impute
○ Average/Most frequent
○ As a distinct value
○ Model-based imputer (simple tree)
○ Random values
○ Remove instances with unknown values
○ Fixed value

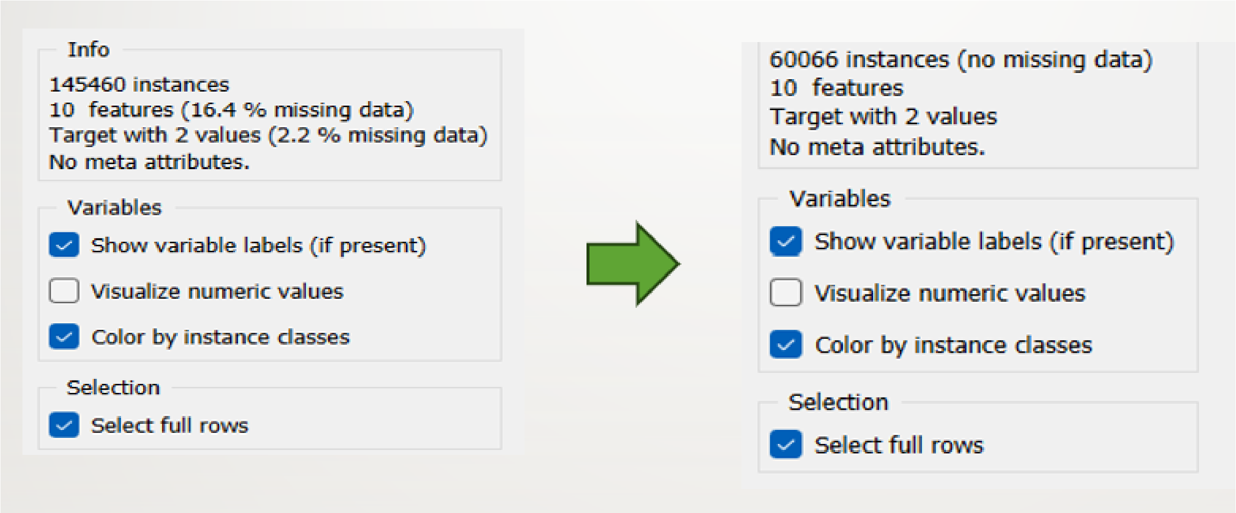
Restore All to Default

☑ Apply Automatically

86.1k | - 60.1k

Figure 24: Impute Widget (Julius Chan,2024)

The Impute widget in Figure 24 is a pre-processing step that manages missing data in a dataset. The "Remove instances with unknown values" option only eliminates data rows (instances) that have missing or unknown values.



Info

145460 instances
10 features (16.4 % missing data)
Target with 2 values (2.2 % missing data)
No meta attributes.

Variables

☑ Show variable labels (if present)

☐ Visualize numeric values

☑ Color by instance classes

Selection

☑ Select full rows

60066 instances (no missing data)
10 features
Target with 2 values
No meta attributes.

Variables

☑ Show variable labels (if present)

☐ Visualize numeric values

☑ Color by instance classes

Selection

☑ Select full rows

Figure 25: Data Post-Processed Data

After performing the Impute Function, no missing or unknown values are identified, as shown in Figure 25. This ensures the dataset used for model training and testing is cleaned, eliminating any concerns with missing data that could hinder model performance.

2.2.5 Machine Learning Algorithm Overview

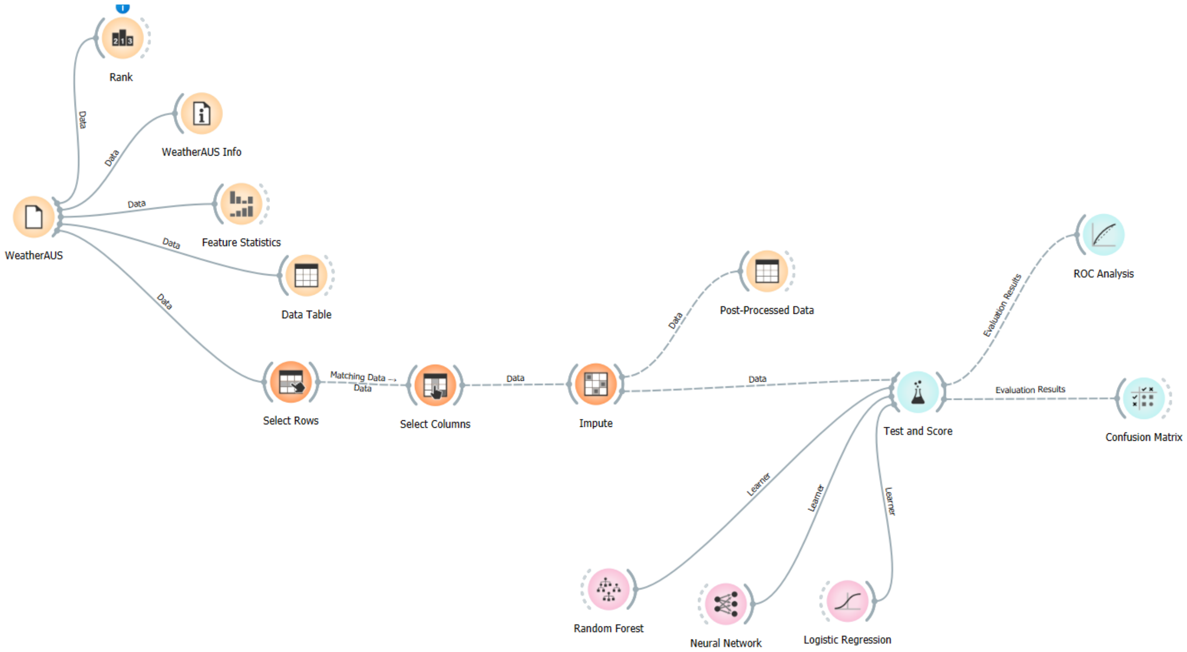


Figure 26:Construction of Machine Learning Program (Julius Chan, 2024)

Figure 26 shows an overview of the program construction using Orange and illustrates the complete workflow for rainfall prediction that includes feature selection, data pre-processing, model training, and assessment. The selected machine learning algorithm consists of Logistic Regression, Random Forest, and Neural Network.

2.2.6 Selected Machine Learning Algorithms

Random Forest: The Random Forest Algorithm is a commonly used machine learning algorithm that combines the output of multiple Decision Trees to achieve a single result. (GFG, 2022).

It works well with both numerical and categorical variables since it detects key features for rainfall prediction, making the weatherAUS dataset a good match due to its resistance to overfitting.

Rainfall is predicted differently by each tree using weatherAUS dataset features. The algorithm creates various decision trees during training by combining subsets of attributes and data.

The final prediction for input is based on the average or weighted average of all the individual trees' predictions. (AnalytixLabs, 2023)

Neural Network: The Neural Network widget uses sklearn's Multi-layer Perceptron algorithm. A Multi-Layer Perceptron (MLP) is a sort of artificial neural network that has multiple layers of neurons and is frequently used for different machine-learning tasks, including classification and regression. (GFG, 2023).

Neural networks excel at weather prediction because of the ability to capture complicated, non-linear correlations in data, flexibility to different datasets, and potential for high predictive performance, particularly on large and complex datasets.

During training, the network learns complicated, non-linear correlations in the weatherAUS dataset by passing input features through its layers and using non-linear activation functions which help to capture the relationship between features and rainfall outcome.

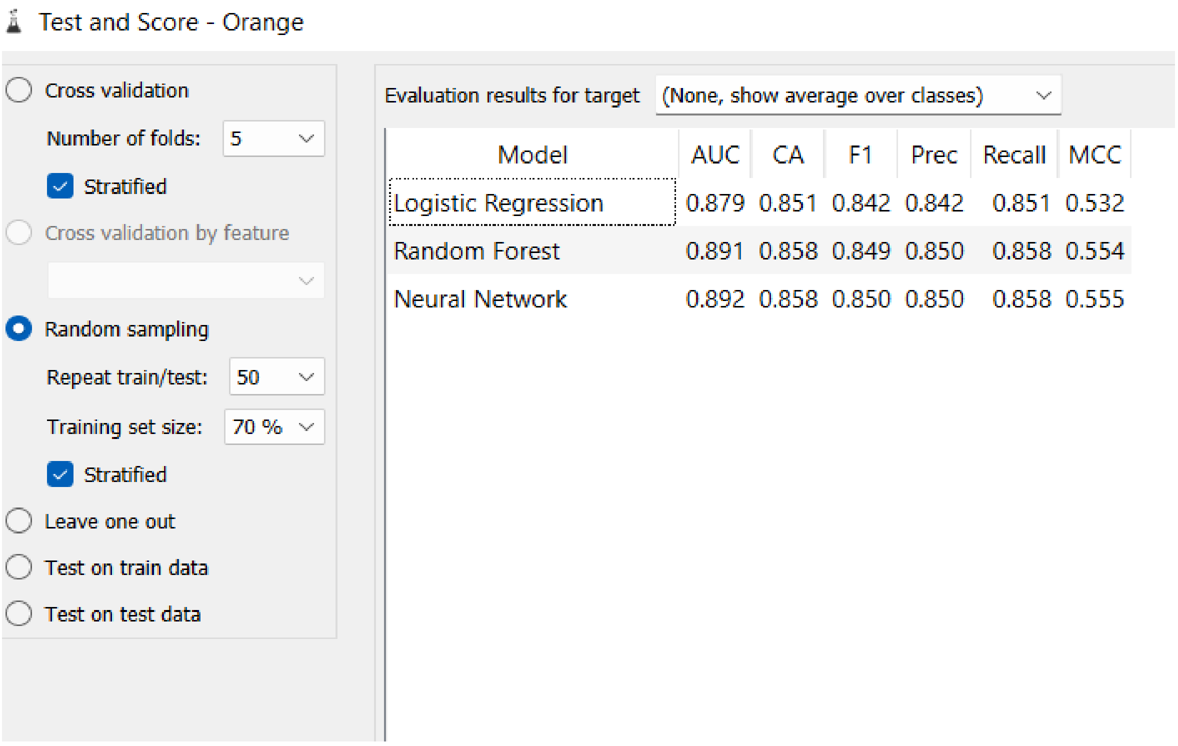
Logistic Regression: Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. (Vijay Kanade, 2022)

The algorithm simplifies and interprets the relationship between features and rainfall probability which is well-suited to binary classification like predicting rain tomorrow. It also provides a baseline for comparing complex models such as Random Forest and Neural Networks.

During training, Logistic Regression defines the parameters of a logistic function, which converts input characteristics to a probability score ranging from 0 to 1 for rainfall prediction.

2.2.6 Performance and Results

Figure 26 shows the selected performance matrix. Test & Score displays an overall model evaluation. ROC Analysis assesses the model's ability to distinguish between rainy and non-rainy days at various threshold levels. Confusion Matrix gives detailed error insights for refinement. This approach provides comprehensive evaluation and effective model improvement.



Test and Score - Orange

○ Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

● Random sampling

Repeat train/test: 50

Training set size: 70 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.879	0.851	0.842	0.842	0.851	0.532
Random Forest	0.891	0.858	0.849	0.850	0.858	0.554
Neural Network	0.892	0.858	0.850	0.850	0.858	0.555

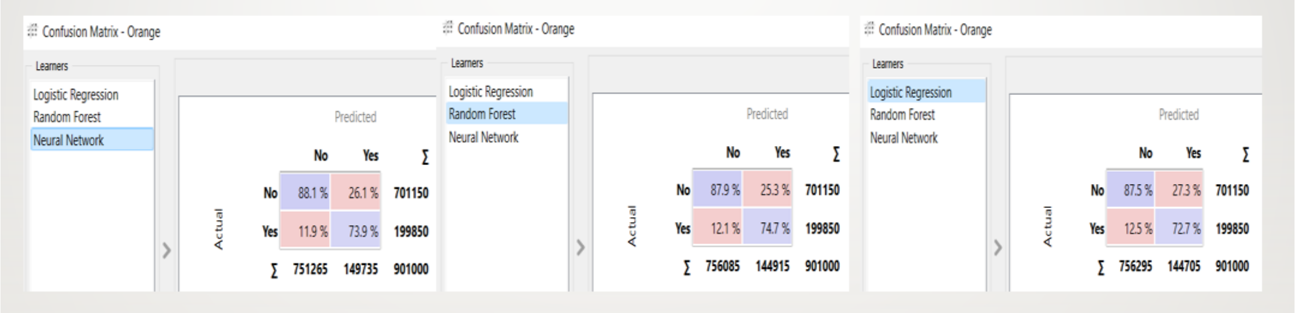
Figure 27: Test and Score of Machine Learning Algorithms (Julius Chan,2024)

Accuracy: Neural Network is the most accurate (0.892), topping Random Forest (0.891) and Logistic Regression (0.879).

Precision and Recall: Random Forest and Neural Network both have excellent accuracy and recall, which allows them to detect real positives while limiting false positives successfully. Logistic regression produces somewhat more incorrect predictions.

F1: The Neural Network has the greatest F1 score (0.850), suggesting the best balance of precision and recall, which is critical for reducing prediction mistakes.

MCC: Neural Network has the greatest MCC (0.555), suggesting the best-balanced performance across all classes and providing a trustworthy measure of model quality.



	No	Yes	∑
No	88.1 %	26.1 %	701150
Yes	11.9 %	73.9 %	199850
∑	751265	149735	901000

	No	Yes	∑
No	87.9 %	25.3 %	701150
Yes	12.1 %	74.7 %	199850
∑	756085	144915	901000

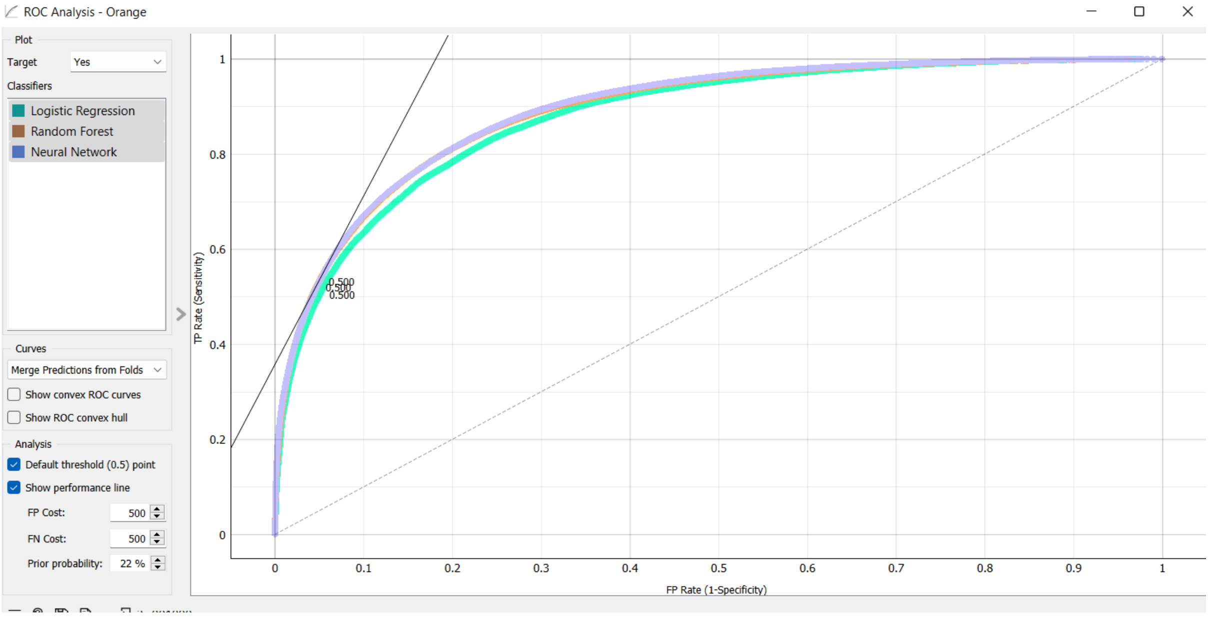
	No	Yes	∑
No	87.5 %	27.3 %	701150
Yes	12.5 %	72.7 %	199850
∑	756295	144705	901000

Figure 28: Confusion Matrix of Machine Learning Algorithms (Julius Chan,2024)

Neural Network: The best balance, with the fewest false negatives (11.9%) and the highest true positive rate (88.1%), suggests a good rainy day forecast.

Random Forest: Similar to Neural Network, but with a larger false positive rate (25.3%), suggesting a somewhat better likelihood of forecasting rain when it does not fall.

Logistic regression: This has a slightly greater false negative rate (12.5%), which might lead to more missed rainy day forecasts.



ROC Analysis - Orange

Plot

Target: Yes

Classifiers

- 🟩 Logistic Regression
- 🟫 Random Forest
- 🟦 Neural Network

Curves

Merge Predictions from Folds

☐ Show convex ROC curves

☐ Show ROC convex hull

Analysis

☑ Default threshold (0.5) point

☑ Show performance line

FP Cost: 500

FN Cost: 500

Prior probability: 22 %

Figure 29: ROC Analysis of Models (Julius Chan, 2024)

Additionally, the ROC curve for the 'Yes' target variable on RainTomorrow was evaluated and showed high ROC values for the selected models. This demonstrates greater performance in distinguishing positives from negatives, establishing these models as the most effective for rainfall prediction.

2.2.7 Conclusion

The project requires a combination of design and technical skills. A rainfall forecast model with good predictability was developed through design and execution processes.

The steps involve selecting features, using appropriate machine-learning algorithms, and measuring model performance using precise metrics.

The project offered valuable insights toward forecasting rain successfully. Using the capabilities of machine learning technologies such as Orange, weather forecasts were achieved with high accuracy and reliability, making them relevant to additional real-world data sets.

3.0 List of References

AnalytixLabs (2023) “Random Forest Regression — How It Helps in Predictive Analytics?” , https://medium.com/@byanalytixlabs/random-forest-regression-how-it-helps-in-predictive-analytics 01

Santosh D (n.d.) “Big Data Case Study - Google.” , www.linkedin.com/pulse/big-data-case-study-google-santosh-d/.

GeeksforGeeks (2023) “Classification Using Sklearn Multi-Layer Perceptron.” , www.geeksforgeeks.org/classification-using-sklearn-multi-layer-perceptron/

d.o.o, Arctur (n.d.) “ORANGE: Faculty of Computer and Information Science.” , www.inzenirji-bomo.si/en/kvizum/2021090213234141/orange-faculty-of-computer-and-information-science/.

GeeksforGeeks (2022) “Differences between Random Forest and AdaBoost.” www.geeksforgeeks.org/differences-between-random-forest-and-adaboost/.

Inforiver (2022) “Variance Analysis in Power BI Using Inforiver Charts.” www.inforiver.com/blog/general/variance-analysis-powerbi-inforiver-charts/.

Moore, Matt. (2016) “The 7 V’s of Big Data.” www.impact.com/marketing-intelligence/7-vs-big-data/.

Statista (2024) . “YouTube: Hours of Video Uploaded Every Minute 2019” , www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/.

Trigyn Insights (n.d.) | the Seven V’s of Big Data Analytics.” *www.trigyn.com*, www.trigyn.com/insights/seven-vs-big-data-analytics#:~:text=The%20Seven%20V.

Yildizbasi, Abdullah, and Yagmur Arioz. (2021) “Green Supplier Selection in New Era for Sustainability: A Novel Method for Integrating Big Data Analytics and a Hybrid

Fuzzy Multi-Criteria Decision Making." *Soft Computing*, https://doi.org/10.1007/s00500-021-06477-8.

Scaleyourapp (2019) "YouTube Database – How Does It Store so Many Videos without Running out of Storage Space? www.scaleyourapp.com/youtube-database-how-does-it-store-so-many-videos-without-running-out-of-storage-space/#:~:text=The%20videos%20are%20stored%20in.

Zach (2021) "How to Interpret a ROC Curve (with Examples)." www.statology.org/interpret-roc-curve/.