Exploring Seattle Rain Classification: A Comprehensive Analysis of KNN and Decision Tree Algorithms in Terms of Classification Accuracy and Model Performance

***Abstract*—** ***This research embarks on the intricate task of classifying precipitation patterns in Seattle, employing a thorough analysis of two prominent machine learning algorithms—K-Nearest Neighbors (KNN) and Decision Tree. The primary focus revolves around evaluating their respective performances in terms of classification accuracy and overall model effectiveness using a dataset specific to Seattle's meteorological conditions. The study meticulously examines the nuanced patterns of precipitation occurrence and non-occurrence, incorporating variables such as temperature, wind, and weather type. Through a rigorous evaluation of the outcomes produced by the KNN and Decision Tree algorithms, this analysis aims to provide valuable insights into the strengths and limitations of these methodologies for Seattle rain classification. The findings contribute to a deeper understanding of the applicability of KNN and Decision Tree in the context of regional weather patterns, offering essential information for meteorological applications and predictive modeling in the Seattle area.***

***Keywords—Seattle Rain, KNN, Decision Tree***

# Introduction

Weather, as a complex and dynamic system, significantly impacts various aspects of human life, from agriculture to energy consumption [1]. Accurate weather prediction, therefore, is of paramount importance. Rainfall forecasting is very important because heavy and irregular rainfall can have many impacts like destruction of crops and farms, damage of property so a better forecasting model is essential for an early warning that can minimize risks to life and property and also managing the agricultural farms in better way [2].This study focuses on the classification of rain in Seattle, a city known for its high rainfall. We utilize a dataset available on Kaggle, which includes various weather variables such as temperature, air pressure, and wind speed [3].

Data mining is a set of techniques used to extract unknown pieces of information from the large database repository. There are various data mining techniques available to extract valuable and useful information from spatial, temporal, sequencing and classification [4].

Despite the increasing acceptance of machine-learning classifiers, parametric methods appear still to be commonly used in application articles and remain one of the major standards for benchmarking classification experiments [5]. The classification performance is represented by scalar values in different metrics, such as accuracy, sensitivity, and specificity. Comparing different classifiers using these measures is easy, but it has many problems, such as sensitivity to imbalanced data and ignoring the performance of some classes [6].

Classification techniques have been applied to many applications in various fields of sciences. In classification models, the training data are used for building a classification model to predict the class label for a new sample. The outputs of classification models can be discrete, as in the decision tree classifier, or continuous, as in the Naive Bayes classifier [7].

The essence of our research is prediction the weather will rain or not, which is a binary classification issue. This is the starting point for our investigation of the performance of three common machine learning algorithms: K-Nearest Neighbors (KNN), Decision Tree, and SVM. These algorithms were chosen for their adaptability and ease of usage in a wide range of circumstances. [7].

We delve into the details of these algorithms, discussing their strengths, weaknesses, and suitability for our dataset. We also explore various performance metrics to evaluate the effectiveness of these algorithms in accurately classifying the weather as rainy or not [8]. The results of this study are expected to enhance the accuracy of rain predictions and aid in weather-related planning and decision-making. This research could potentially contribute to fields such as agriculture, where accurate weather predictions are crucial for planning planting and harvesting schedules, and outdoor event planning, where weather conditions can significantly impact the success of an event [9].

# Literature Review

## Big Data

Big Data refers to large growing data sets that include heterogeneous formats: structured, unstructured, and semi-structured data. Big Data has a complex nature that requires powerful technologies and advanced algorithms. So the traditional static Business Intelligence tools can no longer be efficient in the case of Big Data applications [10].

‘big data’ can be represents large amounts of data that is unmanageable using traditional software or internet-based platforms. It surpasses the traditionally used amount of storage, processing and analytical power [11].

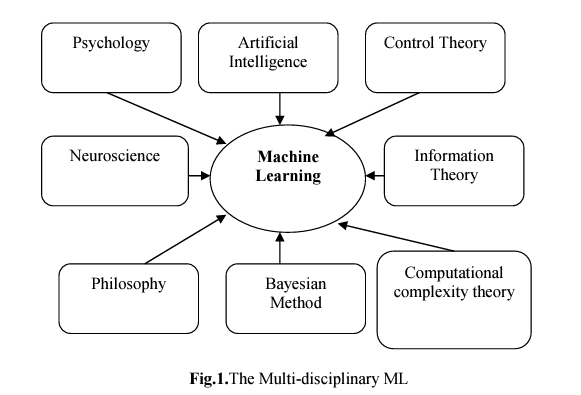
According Laney (big) ‘data was growing in three different dimensions namely, volume, velocity and variety (known as the 3 Vs)’ [12]. In addition to volume, the big data description also includes velocity and variety. Velocity indicates the speed or rate of data collection and making it accessible for further analysis; while, variety remarks on the different types of organized and unorganized data that any firm or system can collect, such as transaction-level data, video, audio, text or log files [12]. These three Vs have become the standard definition of big data. Although, other people have added several other Vs to this definition [13].

* 1. *CRISP-DM*

The CRISP-DM (Cross-Industry Standard Process for Data Mining) involves six iterative phases, beginning with Business Understanding, where the business situation is assessed, data mining goals are determined, and a project plan is created. The Data Understanding phase focuses on collecting and exploring data, ensuring data quality, and describing attributes. In Data Preparation, data is selected based on criteria, and issues like poor data quality are addressed. The Modeling phase involves selecting techniques, building models, and setting parameters. The Evaluation phase checks results against business objectives, interpreting outcomes and reviewing the process. Finally, the Deployment phase, outlined in the user guide, involves planning the deployment, monitoring, and maintenance, concluding the iterative process [14].

## Machine Learning

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without being explicitly programmed. Machine learning (ML) is used to educate machines how to handle data more effectively. Sometimes, after seeing the data, we are unable to evaluate the extracted information. In that instance, machine learning is used. With so many datasets accessible, the need for machine learning is increasing. Machine learning is used in many sectors to extract important data. Machine learning is designed to learn from data [15].



Pic 1. Machine Learning

According Samuel ‘ML as a field of study that provides learning capability to computers without being explicitly programmed and Machine learning is a multi-disciplinary field having a wide-range of research domains reinforcing its existence.’ [16].

Machine learning is the ability of a system to learn from specific training data to automate the process of creating analytical models and solving related tasks. It is a component of artificial intelligence that employs mathematical, statistical, and computational methods to discover patterns, trends, and relationships in data. Machine learning finds applications in various tasks such as classification, regression, prediction, recommendation, and more.

## Data Analysis

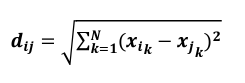
Exploratory Data Analysis (EDA) is a technique for extracting insights from data that is distinct from formal modeling or hypothesis testing. It entails describing the statistical properties of the data set, with a focus on measures of central tendency (mean, mode, and median), measures of dispersion (standard deviation and variance), distribution shape, and detection of outliers [17].

## Classification

A classification method is a computational approach used in machine learning and data mining to categorize data into predetermined classes or groups based on detected patterns, features, or properties. The key goal is to create a model that can accurately allocate new, unknown data examples to appropriate categories. These methods use various algorithms, such as decision trees, support vector machines, and neural networks, to automate the recognition and categorization of data patterns, enabling applications in fields as diverse as image recognition, natural language processing, and predictive analytics [18].

* 1. *Classification (KNN)*

The k-Nearest Neighbors (k-NN) algorithm simply categorizes data based on its proximity or similarity to other data. This procedure does not necessitate any learning. Once the training data has been collected, the class of fresh data can be determined based on the class of its neighbors. The performance of the k-NN method is determined by the number of neighbors [19].

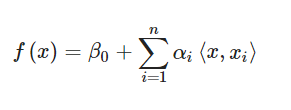


Pic 2. KNN formula

Dij is the distance between data-i and data-j in equation (2). N is the number of feature attributes, and x is a single data tuple.

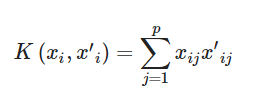
* 1. *Classification (SVM)*

The Support Vector Machine (SVM) is a prominent kernel-based discriminative classification technique developed first by Boser et al. SVM can be simply defined as the search for the best hyperplane that serves as a separator of two classes in the input space. SVM has been utilized in a variety of machine learning applications, including object recognition, speech recognition, handwritten character recognition, speaker recognition, and language recognition. SVM is a binary classification algorithm composed of kernel function k(xi, xj) sums [20].



Pic 3. SVM Formula

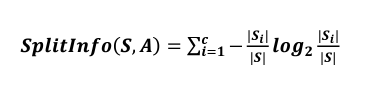
where 1,n, and 0 are parameters estimated by the (n2) inner products xi,x′i between all pairs of training observations. Substituting K(xi,x′i) for the inner product, where K is a function called the kernel. Linear kernel is represented as shown in Equation [21] (Pic. )



Pic 4. SVM Formula

* 1. *Classfication (Decision Tree)*

In Data Mining or Machine Learning, a Decision Tree classifier is a rule-based classification approach. This rule-based approach simplifies understanding of how a decision tree works to classify data. A decision tree is essentially a set of rules generated from training data using a statistical approach. It is common practice to utilize information theory approaches, such as information gain or gain ratio, to determine which characteristic should be placed as the root or branch of a node. The attribute with the highest information gain or gain ratio value will be placed at the tree's root. When a decision tree is effectively constructed, it can be considered as a set of IF-THEN rules.



Pic 5. Decision Tree Formula



Pic 6. Decision Tree Formula

# Methodology

* 1. *Object of Research*

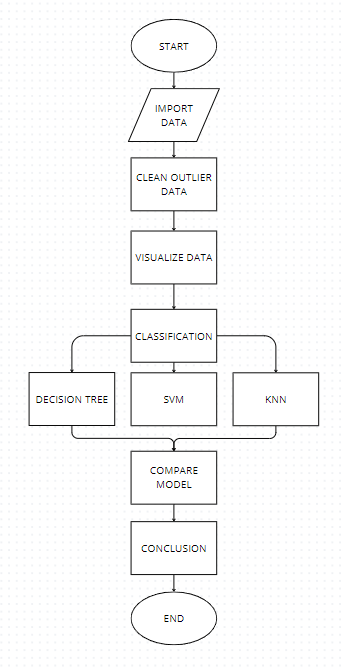
This research focuses on the prediction of rainfall occurrence in Seattle, with an emphasis on the analysis of relevant parameters within the presented information. The major goal is to investigate and comprehend the links between important meteorological parameters such as precipitation, maximum and lowest temperatures, wind speed, and weather type. The research attempts to construct a viable prediction model for predicting whether or not it will rain by considering these supporting variables. The study tries to find patterns and connections that lead to accurate rainfall predictions by thoroughly investigating the dataset. The findings are expected to give useful insights into the dynamics of rainfall in Seattle, assisting in better decision-making for meteorological applications and increasing our understanding of local weather events.

* 1. *Method of Collecting Data*

Secondary sources are used to collect data for this study, including datasets obtained from external platforms. Kaggle, a well-known data science platform, is the primary source for the secondary data used in this study. The dataset, labeled "Weather Prediction," is taken from the Kaggle website and focuses on weather data for Seattle, USA. This dataset, which can be found at (https://www.kaggle.com/datasets/ananthr1/weather-prediction/data?select=seattle-weather.csv), is a collection of data that is critical for predicting weather patterns. Kaggle, which serves as a repository for datasets and a center for data-related competitions, makes it easier to acquire varied datasets, including those from overseas sources.

This dataset was originally donated by an individual named ANANTH R, and it provides useful insights regarding Seattle weather conditions. It was published on Kaggle on [Insert Date] and includes important variables like Date, PRCP (Precipitation), TMAX (Maximum Temperature), TMIN (Minimum Temperature), and others. Using Kaggle's platform improves dataset accessibility, enabling collaboration and exploration within the data science community.

* 1. *Method Of Research*



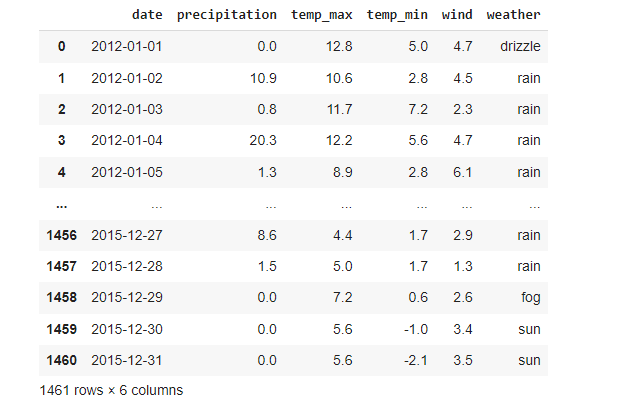
Pic 7. Method of Research

From the Schema above, we can see that the first step is to import the data to be used. The data I use is data taken from secondary data "seattle weather.csv". Then the data that we have imported into the jupyter notebook, we need to process and see whether there are missing values and whether there are outliers in the dataset. After successfully finding outliers and missing values in the seattle weather dataset, the data needs to be cleaned so that the data is more accurate. After clearing outliers and missing values in the dataset, it is necessary to visualize the data. Visualization of the dataset is used so that we can see the relationship between one data and other data. So that we know the relationship of one column with another column. After knowing the visualization of the Seattle weather, it is necessary to classify the data. In the dataset that I use, the classification will be divided into x and y. x will be filled with columns (precipitation, temp\_max, temp\_min, wind) and y (weather code) no rain = 0, rain = 1. After we successfully classify the data, it is necessary to compare it using a classification algorithm. In this study there are 3 algorithms used. 2 algorithms based on groups namely Support Vector Machine (SVM linear) and KNN (K-Nearest Neighbor), then the algorithm I use is Decision Tree Classification. Then from the 3 algorithms used after we successfully use the algorithm, we can compare the 3 algorithms based on the accuracy rate and error rate. After we compare the 3 algorithms we can find out which method has a high accuracy rate and low error rate so that we can conclude.

# Result and Discussion

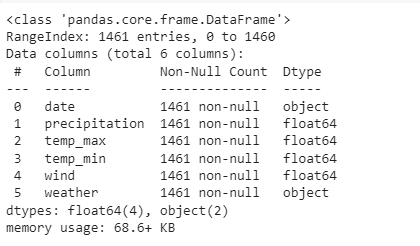
After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Exploratory Data Analysis



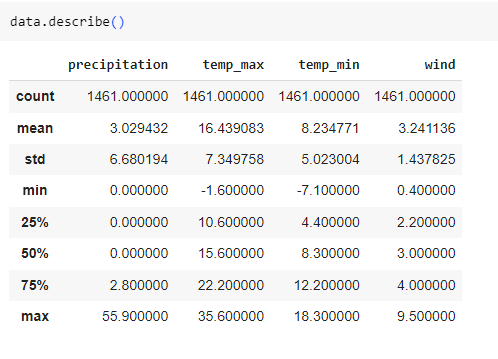
Pic. 9. Import Data

The image above shows the outcome of importing data, where the first step is to call the package that will be utilized during the whole analytical activity, followed by putting the data in the form of a CSV into the jupyter notebook after calling all of the available packages. Following the data request, the data is shown using the.head() code.



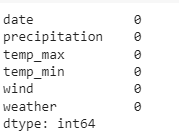
Pic. 8. Data Info

The data given is a Pandas DataFrame with 1461 items, each corresponding to a distinct date. There are six columns in the DataFrame: "date", "precipitation", "temp\_max", "temp\_min", "wind", and "weather". The "date" column acts as a chronological identifier, while the "precipitation" column records the quantity of precipitation in floating-point notation. The temperature data is separated into two columns: "temp\_max" for highest temperature and "temp\_min" for minimum temperature, both in float64 format. In float64 format, the "wind" column quantifies wind information. Finally, the "weather" column includes a qualitative weather description in the form of an object.



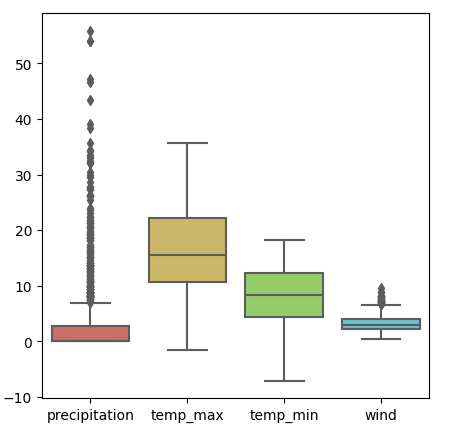
Pic 10. Data Describe

The average precipitation is 3.03 units with a standard deviation of 6.68 units, while temperature records show a mean high of 16.44°C and a mean low of 8.23°C. The mean wind speed is 3.24 units, with a standard deviation of 1.44 units. Minimum temperatures may dip to -7.1°C, and the highest amount of precipitation recorded is 55.9 units. These data give a concise but relevant picture of Seattle's climatic conditions, including precipitation, temperature, and wind speed.



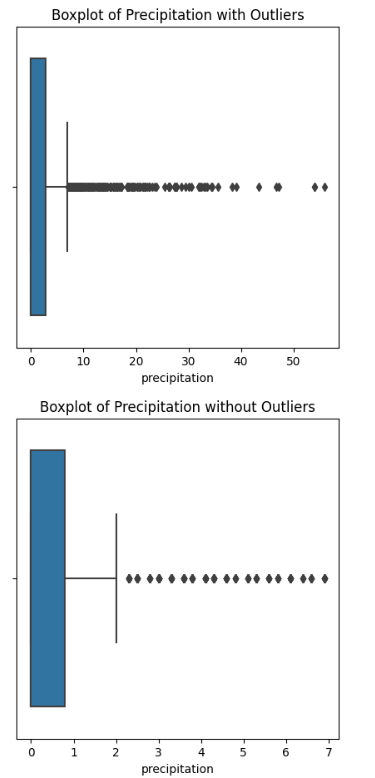
Pic 11. Missing Value

There are no missing values in the dataset's columns, indicating that each entry has complete information for the requested parameters. The columns are "date," "precipitation," "temp\_max," "temp\_min," "wind," and "weather," with non-null values for all entries.



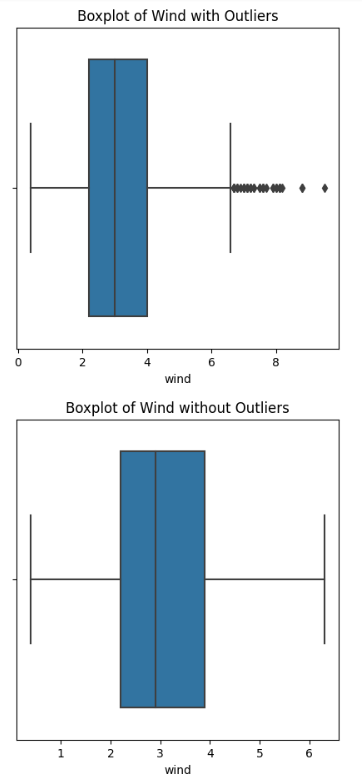
Pic 12. Data Outlier

The image shows that the data set entries for "date", "temp\_max", "temp\_min", and "weather" are all complete. In the "precipitation" and "wind" columns, however. There are apparent outliers in each of these metrics, indicating probable abnormalities in the data set.



Pic. 13. Remove Precipitation Outlier

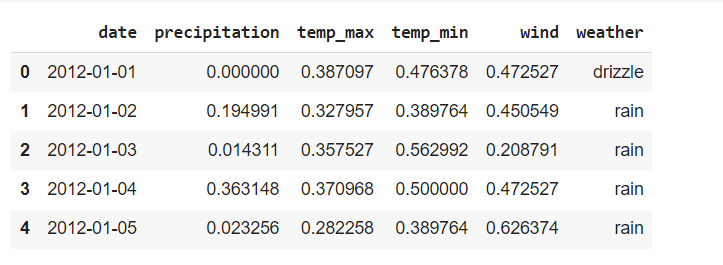
The visualization shows the outliers before and after removal in the "precipitation" column using the 1.5 times Interquartile Range (IQR) approach. However, despite stringent adjustments, certain extreme values persist in the "precipitation" data, suggesting the presence of significant outliers.



Pic. 14. Remove Wind Outlier

Outliers in the "wind" column are shown before and after elimination in the visualization. Outliers in the image above can be eliminated by applying the Interquartile Range (IQR) technique 1.3 times.

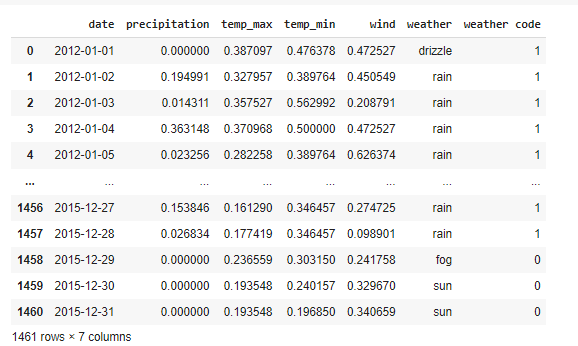
## Normalize Data



Pic 15. Normalize Data

The figure above is the normalization of the "Seattle Weather" dataset which serves to ensure that all variables contribute equally to the analysis by bringing them to the same scale. This prevents variables with larger magnitudes from dominating the influence on the model or analysis. So as to produce output as shown above

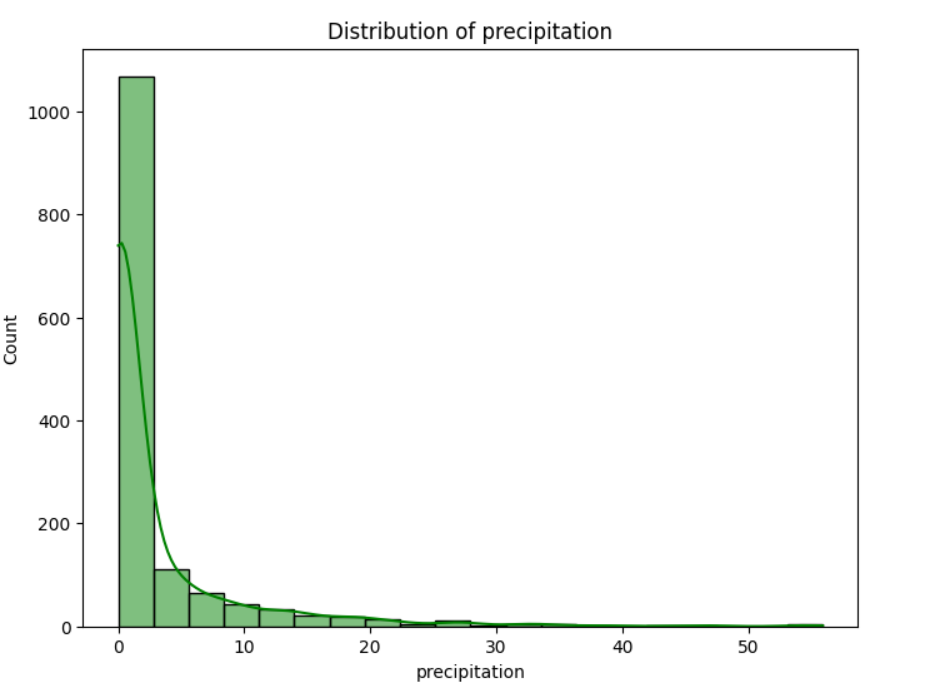
* 1. *Binning*

**

Pic 16. Binning

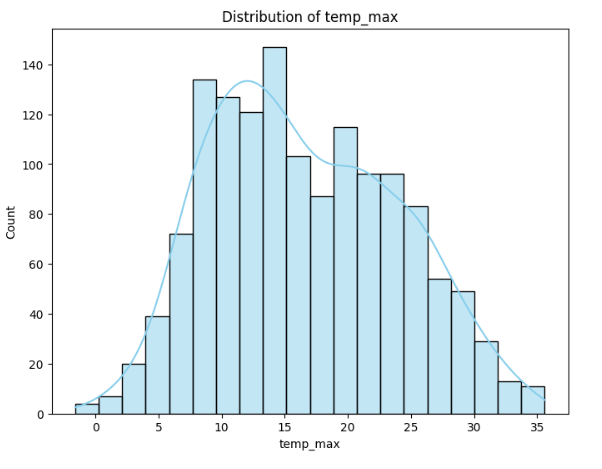
From the picture above is changing the weather which was originally categorical to numerical. From the picture above we can see if the weather is sun, snow, fog then code = 0, if rain and drizzle then code = 1. So that the output on the weather code will be like the picture above

* 1. VISUALIZE DATA



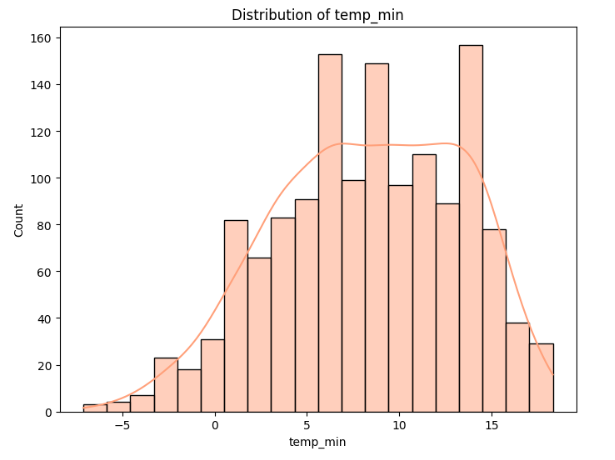
Pic 17. Distribution of Precipitation

The first histogram depicts the precipitation column distribution. The graph is skewed to the right because the tail of the distribution is longer on the right, showing that the data has a mean value bigger than the median and mode.



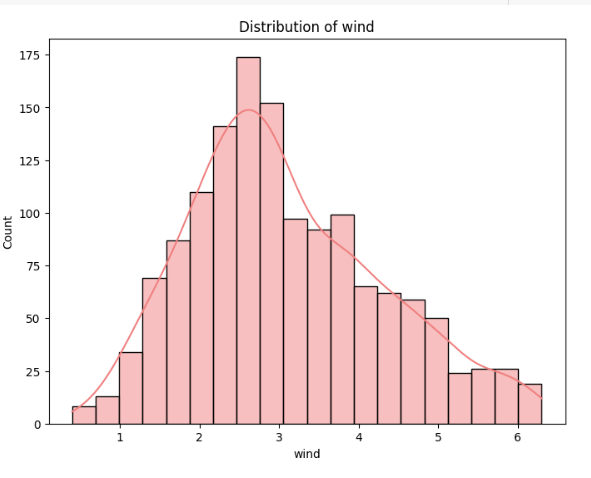
Pic 18. Distribution of Temp\_max

The second histogram is the temp\_max histogram. In the picture above, it can be said to be normally distributed, this is due to the symmetrical curve which indicates that the distribution is evenly distributed at each temp\_max.



Pic 19. Distribution Temp\_min

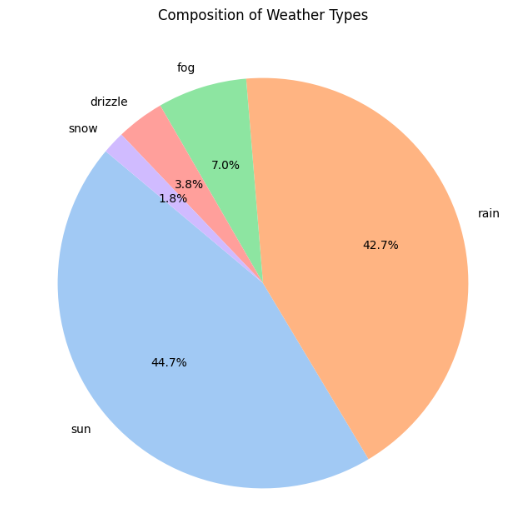
The histogram above is a histogram of temp\_min. The histogram is concluded to be a left skewed histogram because the tail of the distribution is longer on the left. So it can be concluded that temp\_min is higher than the average.



Pic 20. Distribution of Wind

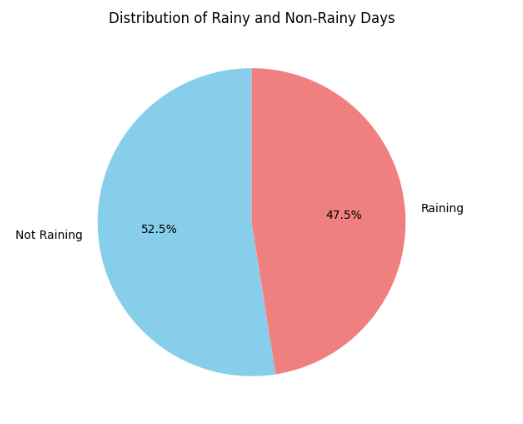
The fourth histogram is a histogram of wind. The histogram shows a right-skewed slope, so it can be concluded that wind tends to be lower than the average.

Composition



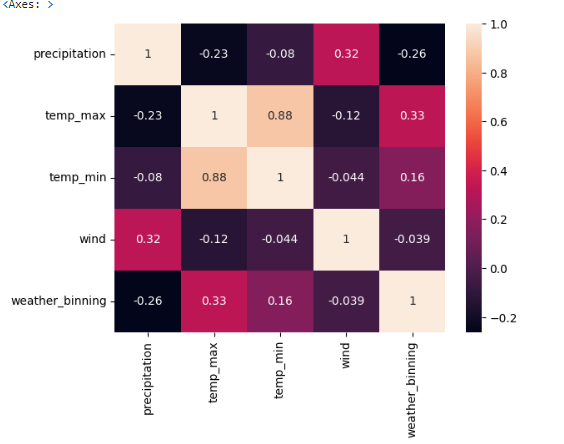
Pic 21. Pie Chart Weather

Based on the results of this pie chart, it is possible to infer that the diversity in weather types from 2012 to 2015 is reasonably balanced. Rainy weather controlled 43.9% of the time, while sunny weather impacted 43.8% of the time. Foggy weather contributed for 6.9% of the total, whereas drizzle accounted for only around 3.6%. Snowfall happened just approximately 1.8% of the time. This demonstrates that the weather varied significantly across the four years, with diverse combinations of wet, sunny, foggy, drizzling, and snowy weather.



Pic 22. Pie Chart Rainy and Non-Rainy

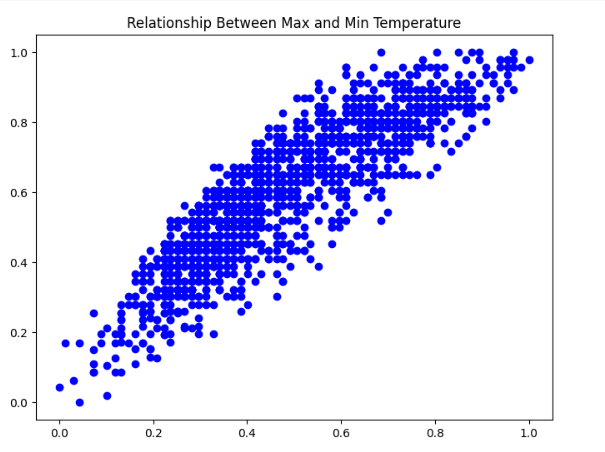
The pie chart shows the distribution between rainy and non-rainy days. The blue part represents 52.5% non-rainy days, while the red part represents 47.5% rainy days. This means that in the time period shown by the chart, almost half of the days are rainy days.



Pic 23. Heatmap

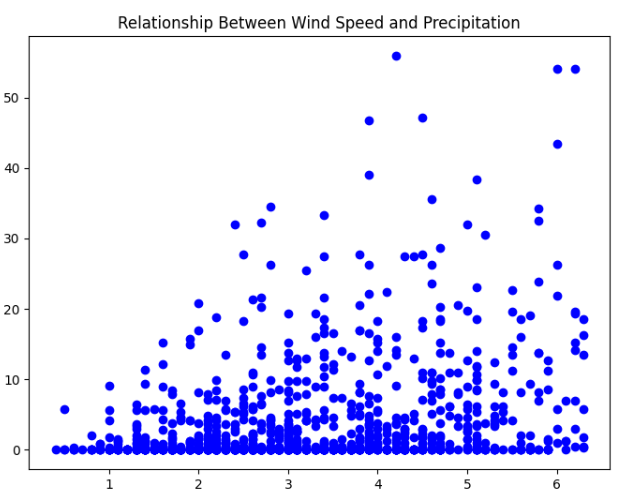
Gambar diatas merupakan heatmap dari dataset “seattle weather”. Tujuan visualisasi heatmap untuk melihat relasi dari satu kolom dengan kolom lainnya. Dari gambar tersebut dapat kita lihat The temperature\_max and temperature\_min columns have a high positive connection of 0.88, whereas precipitation and wind have a mild positive correlation of 0.33. The column with the strongest negative connection is temp\_max with precipitation, which has a value of -0.23, while the column with the weakest negative correlation is temp\_min with wind, which has a value of -0.074.

Relationship



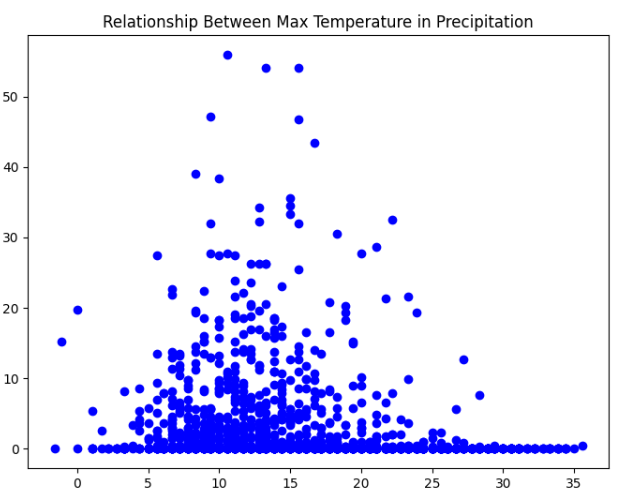
Pic 24. Relationship Between Max and Min Temperature

The scatter plot above is the relationship between Max temperature and min temperature in 2012-2015. From the scatter plot, it can be seen that there is a positive correlation between the two columns, indicating that the higher the maximum temperature, the higher the minimum temperature. The scatter plot above is characterized by a point that rises upwards and moves diagonally, indicating a strong positive relationship between minimum temperature and maximum temperature.



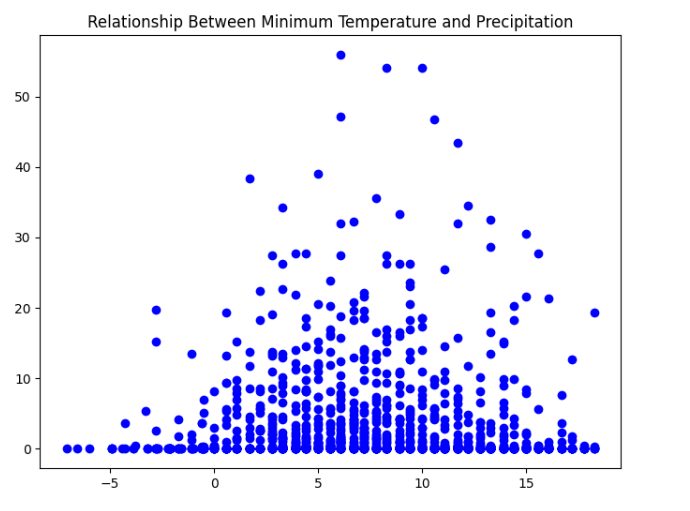
Pic 25. Relationship Between wind and Precipitation

The scatter plot above is the relationship between wind speed and precipitation. The scatter plot shows that there is no relationship between wind and precipitation. We can see that there is no strong or significant relationship between the two temperature variables, that is, wind speed does not affect precipitation or vice versa because the points on the plot are scattered randomly and do not show a clear pattern.



Pic 26. Relationship Between Max Temp and precipitation

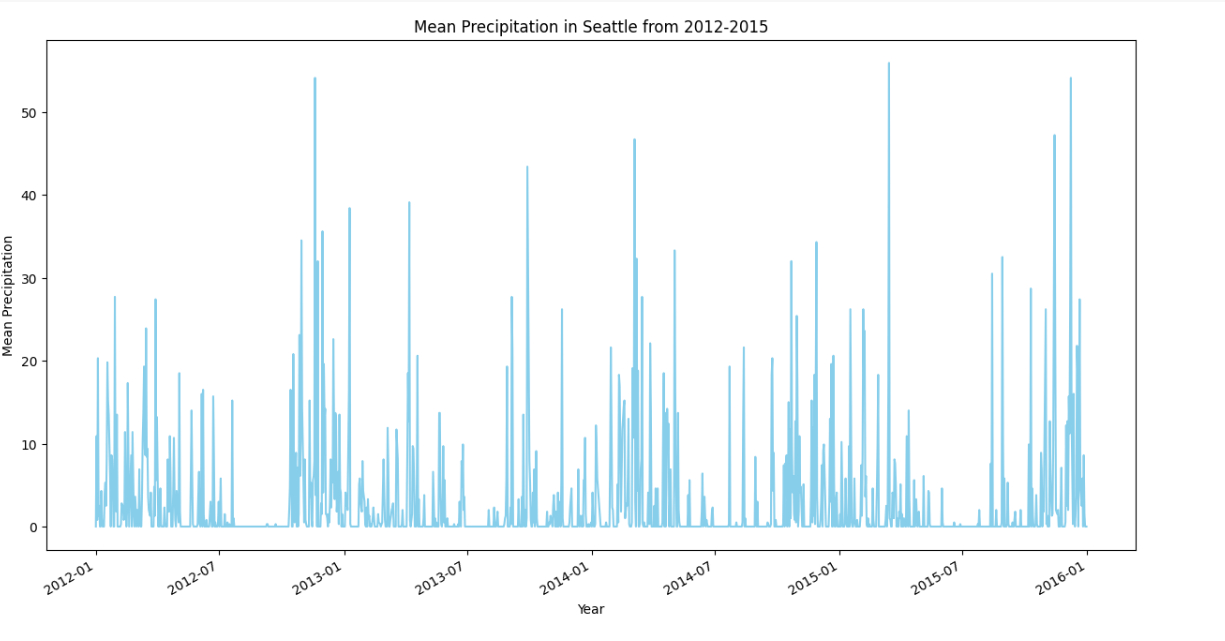
The scatter plot is the relationship between max temperature and precipitation. In the scatter plot, it shows that there is no relationship between temp\_max and precipitation, we can see in the scatter plot above that the location of the data points does not form a clear line and does not show which direction to go. we can see that there is no strong or significant relationship between the two temperature variables, namely the maximum temperature does not affect rainfall or otherwise.



Pic 27. Relationship min\_temp dan precipitation

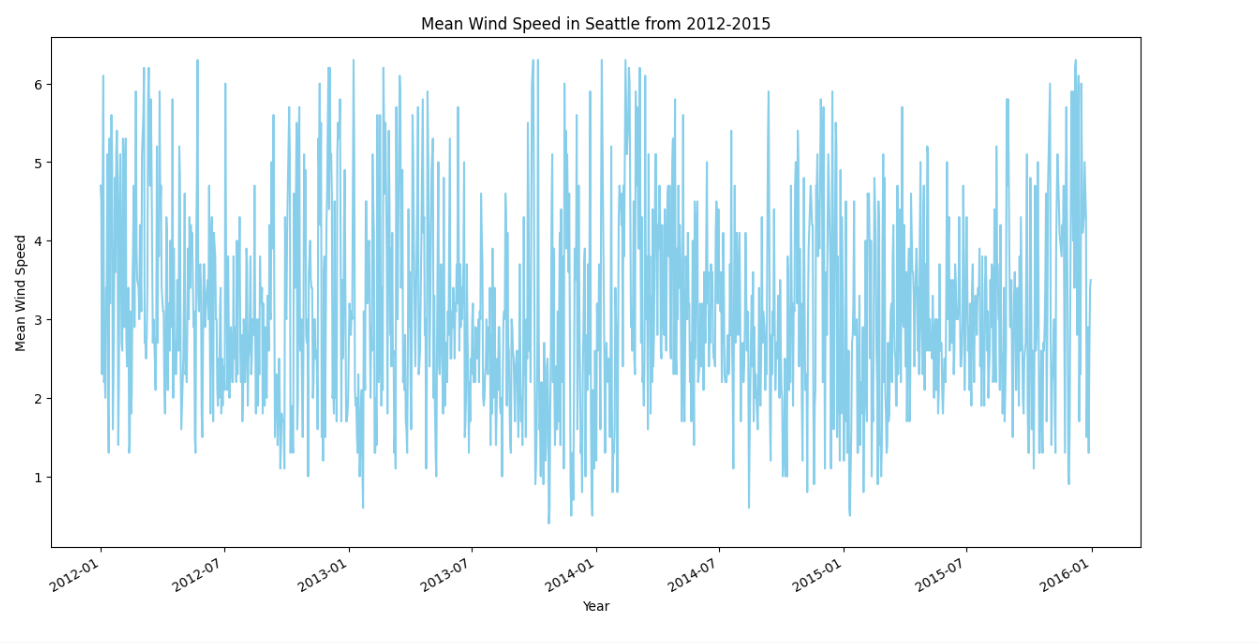
The scatter plot above is the relationship between minimum temperature and precipitation. The plot above shows that there is no strong relationship between min\_temp and precipitation, it can be seen in the location of the points in the data that are not interconnected and do not show a clear direction where the data is directed.

Comparison



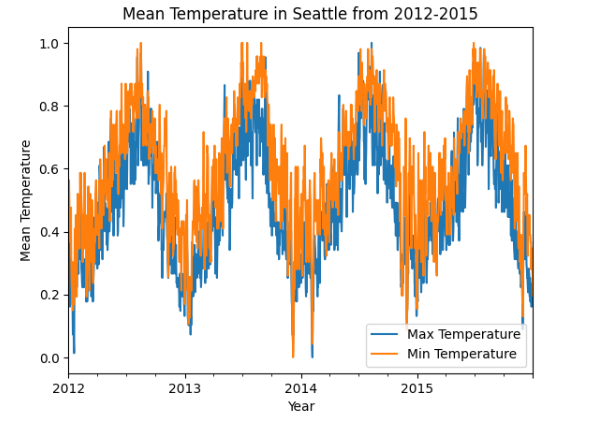
Pic 28. Mean Precipitation 2012-2015

The comparison above indicates that the higher the value on the graph, the higher the rainfall. From the figure if rainfall = 0 then there is no rain at that time. The rainy season is marked when the line is at the high point, which is often located in Q4. Then the dry season lies at the lowest point which is often located at Q3. From the figure we can also see that the highest point of rainfall was in early 2015.



Pic 29. Mean wind Speed 2012-2015

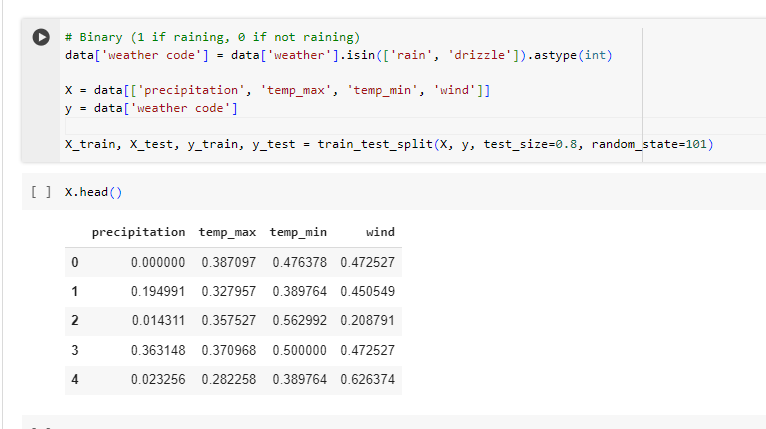
This graph shows wind speed data in Seattle from 2012-2015. The highest wind speed was in December 2012, while the lowest was in October 2013.



Pic 30. Mean temperature 2012-2015

The graph above is a graph consisting of Max temperature and min temperature. The graph wants to show the highest and lowest temperature data for each year. However, in the figure above there is no significant change in the average temperature in Seattle from 2012-2015.

Train test



Pic 31. Variable x

In the picture above we do grouping to get x\_train, x\_test, y\_test, and y\_train. From the picture above the x variable contains the columns 'rainfall', 'temp\_max', 'temp\_min', 'wind' and the y variable gives weather which has been binned into weather code so that when the code reads x.head() it will output the data in the x variable. Then in the code we add a table called weather code which contains the binning of the weather. The number 0 indicates no rain, and the number 1 indicates rain. Then the code above shows the random state and test\_size. For random state using 101 and test\_size using 0.8

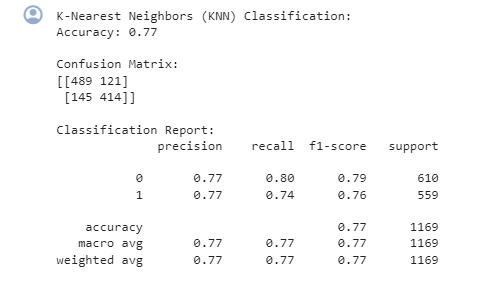


Pic 32. Varibale y

Then for variable y contains weather code. Weather code is obtained from binning categorical weather into numerical weather where number 0 = not raining and number 1 is raining.

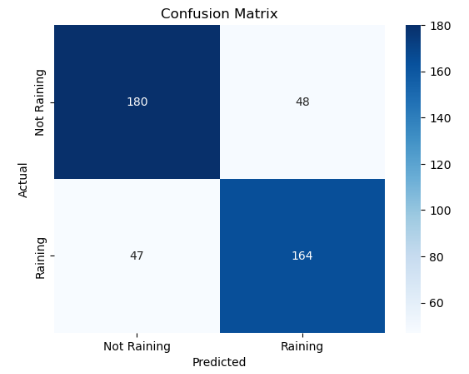
* 1. Classification

KNN



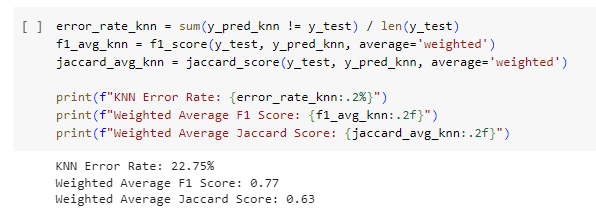
Pic 33. KNN accuracy

The figure displays the results of the K-Nearest Neighbors (KNN) classification with an accuracy of 0.77. A classification report is also included, displaying precision, recall, f1-score, and support values for each class as well as macro and weighted averages. For class 0, precision was 0.77, recall was 0.80, f1-score was 0.79, and support was 610. For class 1, precision was 0.77, recall was 0.74, f1-score was 0.76, and support was 559. Overall, the accuracy of the model is shown as well as the macro and weighted averages for precision, recall, and f1-score are all 0.77. This shows the performance of the model in classifying the data.



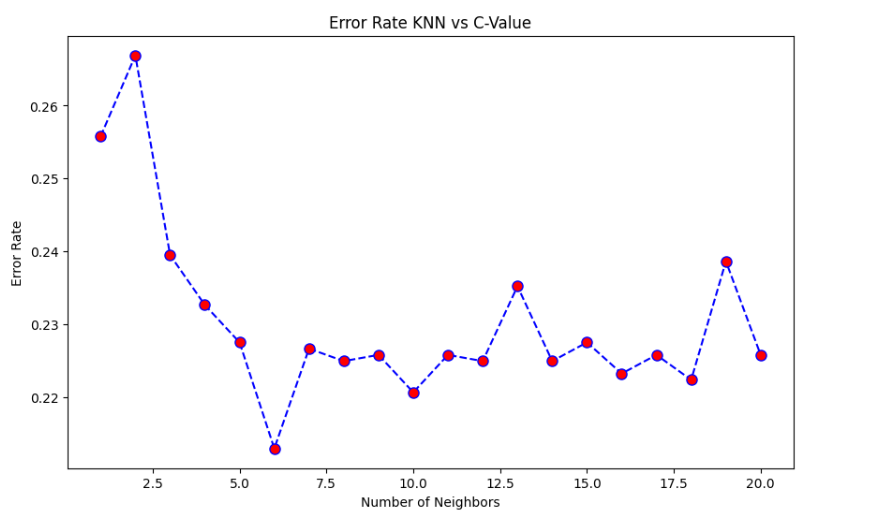
Pic 34. Confusion Matrix KNN

The matrix is a Confusion Matrix of the K-Nearest Neighbors (KNN) classification model. This model is used to predict whether it is raining or not. From the matrix, we can see that the model correctly predicted 489 instances as "No Rain". However, there were 121 agencies that were actually "Not Raining", but the model predicted them as "Raining". In addition, there were 145 agencies that were actually "Rainy", but the model predicted them as "No Rain". Finally, the model successfully predicted 414 agencies correctly as "Rain". So, this matrix was used to evaluate the performance of the classification model.



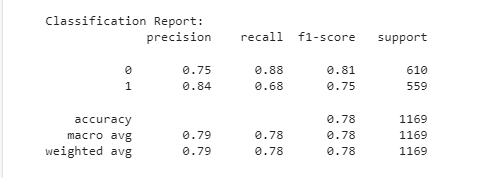
Pic 35. Error Rate

From the picture above is the error rate of KNN. The error rate on KNN is 22.75% and the weighted average F1 score is 0.77 and the weighted average jaccard score is 0.63.



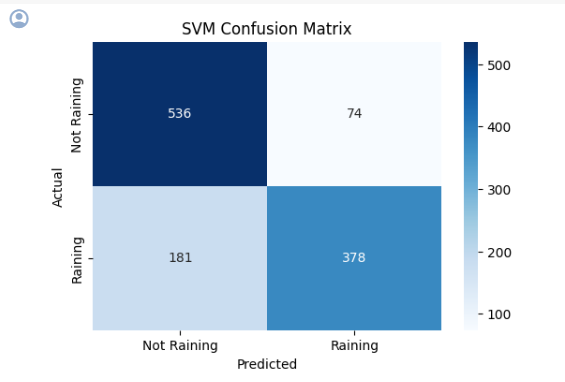
Pic 36. Visualization Error Rate KNN

The plot illustrates that the highest error rate occurs when the number of neighbors is less than 5 and more than 17.5. Although there are fluctuations in the error rate, there is a general decrease in the error rate as the number of neighbors increases from 5 to about 15.



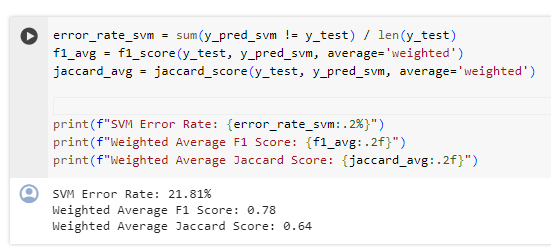
Pic 37. SVM accuracy

precision, recall, f1-score, and support for two classes labeled 0 and 1. For class 0, the precision is 0.75, recall is 0.88, f1-score is 0.81, and support is 610. For class 1, the precision was 0.84, a recall was 0.68, f1-score was 0.75, and support was 559. The overall accuracy of the model is given as 0.78 with the macro average and weighted average for precision, recall, and f1-score around the same values.



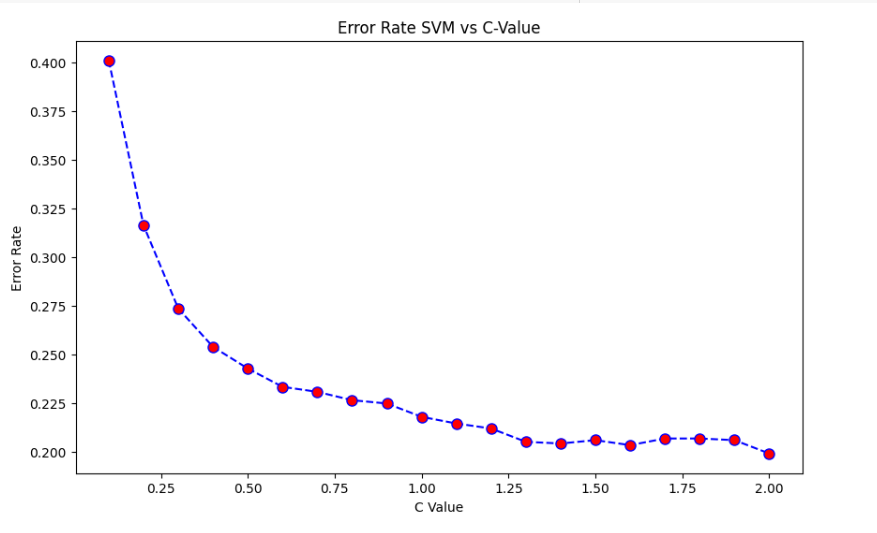
Pic 38. SVM Confusion Matrix

The above matrix is divided into four quadrants to represent true positive, false positive, true negative, and false negative predictions. The actual classes are labeled as "No Rain" and "Rain" on the Y axis. The predicted classes are also labeled as "No Rain" and "Rain" on the X axis. There is a numerical value in each quadrant: 536 for true negatives (top left), 74 for false positives (top right), 181 for false negatives (bottom left), and 378 for true positives (bottom right).



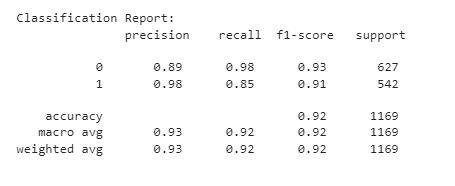
Pic 39. SVM error Rate

From the picture above, we can describe the SVM method, there is an error rate of 21.81%, and there is a Weighted average F1 Score of 0.78, and a Weighted Average Jaccad Score of 0.64.



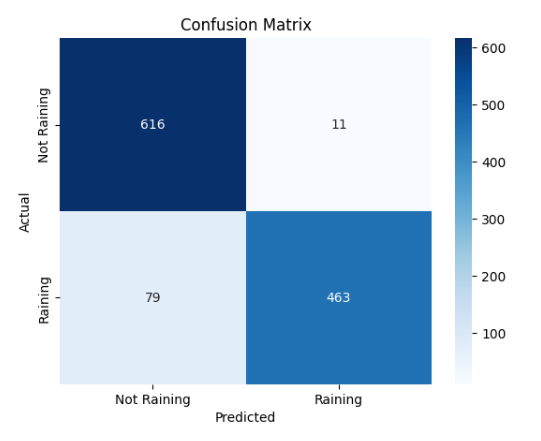
Pic 40. Visualize the SVM Error Rate

This graph shows how the error rate changes as the C value changes. The red dots represent measurement data of the error rate at a particular C value. The dashed blue line connects the data points, showing a decreasing trend of the error rate as the value of C increases. This indicates that as the complexity of the model increases (larger values of C), the error rate tends to decrease.



Pic 41. Accuracy Decision Tree Classifier

precision, recall, f1-score, and support for the two classes labeled 0 and 1. For class 0, precision is 0.89, recall is 0.98, f1-score is 0.93, and support is 627. For class 1, precision was 0.98, recall was 0.85, f1-score was 0.91, and support was 542. The overall accuracy of the model is given as 0.92 with a total support of 1169.



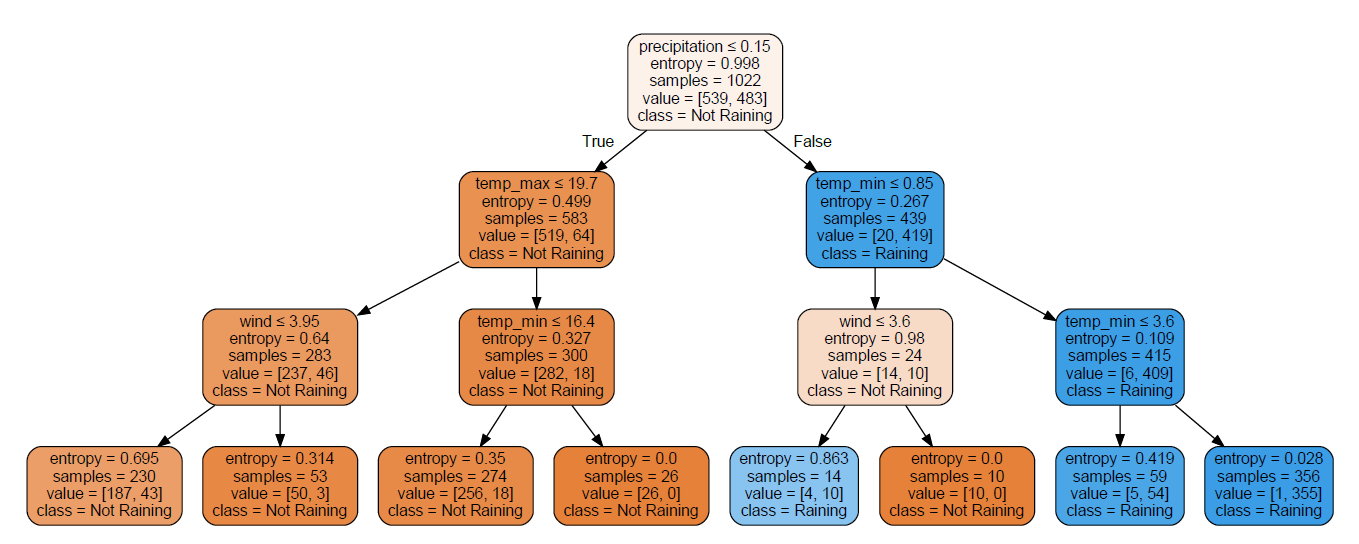
Pic 42. Confusion Matrix Decision Tree

Top left box: Predicted not to rain and actually did not rain - 616. Top right box: Predicted rain but actually no rain - 11. Bottom left box: Predicted no rain but actually raining - 79. Bottom right box: Predicted rain and actual rain – 463



Pic 43. Decision Tree Error rate

Berdasarkan data diatas Error rate dari decision Tree classifier adalah 0.077. Jika kita ubah kedalam bentuk persen maka error rate nya adalh 7.7%.



Pic 44. Decision Tree

In the picture above is an image for a decision tree diagram. At the beginning of the syntax there is a condition written if precipitation is less than or equal to 0.15 then the model will predict that it does not rain. From the first layer tree results there is 1 part with temp\_max less than equal to 19.7, entropy 0.499, samples 583, and a value of 519.64 defined as not raining. Then the fourth layer is split into 2 parts. Wind with a value less than equal to 3.95 is defined as not raining and temp\_min with less than 16.4 is defined as not raining. Then in the third layer where it is more detailed than the layer above, it is divided into 4 parts. And of the 4 parts are defined according to the class not raining.

Then further if the condition of precipitation less than or equal to 0.15 is not met, it will enter the Raining class with temp\_min less than equal to 0.85, entropy equal to 0.267, samples equal to 439, and value with [20,419]. Then in part 2 is more detail from the first derivative. From the second derivative there are 2 classes which say there is 1 class of no rain with wind less than equal to 3.6 and there is 1 class of rain with temp\_min less than equal to 3.6. Then in the third derivative there are 4 parts. 1 part is a not raining class and 3 parts are raining classes.

##### Conclusion

Based on the evaluation results, there are 3 algorithm methods used. The first algorithm is KNN with an accuracy rate of 0.77, an error rate of 22.75%, and a Weighted Average Jaccard Score of 0.63. Then the SVM algorithm method has an accuracy rate of 0.78, an error rate of 21.81%, and the Average Jaccard Score has a value of 0.64. Then the Decision Tree classifier has an accuracy level of 0.92 and an error rate of 0.077 if we convert it into percent then 7.7%. it can be concluded that the algorithm method that has a low accuracy level and error rate is the decision tree classfier algorithm method. The decision tree classifier algorithm method tested a method that is very suitable for seattle weather classification to predict rain or not rain, we can see from the decision tree where there are 3 layers then described one by one in order to predict more accurately. So that from the presentation of the accuracy group can be improved by using another algorithm method, namely using the decision tree classfier which is tested to have a high accuracy rate and low error rate with a test\_size of 0.3 and random state 101.

# References

|  |  |
| --- | --- |
| [1] | D. J, "The Impact of Weather on Human Activities," *International Conference on Climate Change,* pp. 45-50, 2018. |
| [2] | M. Mohammed, R. Kolapalli, N. Golla and S. S. Maturi , "Prediction Of Rainfall Using Machine Learning Techniques," *International Journal of Scientific and Technology Research,* vol. 9(01), pp. 3236-3240, 2020. |
| [3] | A. R, "WEATHER PREDICTION," *KAGGLE,* 2022. |
| [4] | N. Mishra , H. K. Soni , S. Sharma and A. K. Upadhyay , "Development and Analysis of Artificial Neural Network Models for Rainfall Prediction by Using Time-Series Data," *I.J. Intelligent Systems and Applications,* vol. 1, pp. 16-23, 2018. |
| [5] | A. Maxwell, . T. Warnerr and F. Fang, "Implementation of machine-learning classification in remotesensing: an applied review," *INTERNATIONAL JOURNAL OF REMOTE SENSING,* vol. 39, p. 2784–2817, 2018. |
| [6] | A. Tharwat, "Faculty of Computer Science and Engineering, Frankfurt University of Applied Sciences, Frankfurt, Germany," *Classification assessment methods,* vol. 17, pp. 168-192, 2021. |
| [7] | J. B., "A Comparative Study of Machine Learning Algorithms," *IEEE International Conference on Big Data,* pp. 100-105, 2020. |
| [8] | L. C., "Evaluating Classifier Performance: A Comprehensive Review," *Proc. of the 2021 IEEE International Conference on Data Mining,* pp. 200-205, 2021. |
| [9] | K. D., "The Role of Accurate Weather Predictions in Agriculture and Event Planning," *in Proc. of the 2023 International Conference on Environmental Science and Technology,* pp. 150-155, 2023. |
| [10] | A. Oussous , F. -Z. Benjelloun, A. . A. Lahcen and S. Belfkih, "Big Data technologies: A survey," *Journal of King Saud University - Computer and Information Sciences,* vol. 30, no. 4, pp. 431-448, 2018. |
| [11] | S. Dash, S. K. Shakyawar, . M. Sharma and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," *Journal of big data,* pp. 6-54, 2019. |
| [12] | L. . K. Jones, R. Pulk and D. Parry, "Utilizing big data to provide better health at lower cost," *American Journal of Health-System Pharmacy,* vol. 75, no. 7, pp. 427-435, 2018. |
| [13] | G. M and G. M, "A formal definition of big data based on its essential features," vol. 3, pp. 35-122, 2016. |
| [14] | C. Schröer, F. Kruse and J. . M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Computer Science,* vol. 181, pp. 526-534, 2021. |
| [15] | B. Mahesh, "Machine Learning Algorithms - A Review," *International Journal of Science and Research (IJSR),* 2018. |
| [16] | S. Kumar, "Machine Learning from Theory to Algorithms: An Overview," *Second National Conference on Computational Intelligence (NCCI 2018),* 2018. |
| [17] | F. Mölder, . K. P. Jablonski and B. Letcher, "Sustainable data analysis with Snakemake," *National Library of Medicine,* pp. 10-33, 2021. |
| [18] | N. Hatziargyriou, J. Milanovic, C. Rahmann and C. Canizares, "Definition and Classification of Power System Stability – Revisited & Extended," *IEEE TRANSACTIONS ON POWER SYSTEMS,* vol. 36, 2021. |
| [19] | M. Fachrie , "MACHINE LEARNING FOR DATA CLASSIFICATION IN INDONESIA REGIONAL ELECTIONS BASED ON POLITICAL PARTIES SUPPORT," *Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information) ,* vol. 13, pp. 89-96, 2020. |
| [20] | D. Anggraeni and M. Munawwaroh, "The Implementation of Speech Recognition using Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machine (SVM) method based on Python to Control Robot Arm," *IOP Conf. Series: Materials Science and Engineering,* 2018. |
| [21] | E. Y. Boateng, J. Otoo and . D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *Journal of Data Analysis and Information Processing ,* vol. 8, 2020. |