**Weather Forecasting**

**Mümine Kaya Keleş, Alize Serra Açar,**

1*Adana Alparslan Türkeş Bilim ve Teknoloji Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Adana*

2 *Adana Alparslan Türkeş Bilim ve Teknoloji Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Adana*

**Özet**

Bu çalışmada, Adana’nın bir yıllık hava durumu analizi ve 2021 yılına dair tahmini göstermiştir.

**Anahtar Kelimeler:** Adana, Weather Forecasting

**Weather Forecasting**

**Abstract**

In this study, Adana’s one year weather analysis and 2021 forecast milk condition.

**Keywords:** Adana,Weather Forecasting

**1. INTRODUCTION**

The aim of the study is :Weather forecasting; has evolved significantly from the late 1800s, when telegraph and telephone technology first allowed multiple weather stations to share observations and develop a synoptic view of weather systems as they moved across the country. In the late

20th century, satellites enabled another leap forward, providing an **eye in the sky** to

monitor Earth system phenomena that include tropical storms evolving into hurricanes.

Now, computer models integrate huge volumes of data, producing everything from

simulations of long-term climate trends to nowcasts that predict small weather events just

a few minutes or hours into the future. And the progress continues: Nowcasting and

weather forecasting are on the brink of a major paradigm shift. New approaches are

required to enable full Earth system prediction and to make cost-effective use of the

dramatic increase in volume, diversity, and capabilities of observations (particularly

satellite observations) and environmental products. We predict that machine learning

(ML) and other data science techniques for predict. Weather Forecasting systems can be used in: air traffic,marine,agriculture,forestry,military and navy etc.

**1.1 Advantages**

* The primary advantage of forecasting is that it provides the business with valuable information that the business can use to make decisions about the future of the organization.

**1.2 Disadvantages**

* Weather forecast by the system is not very certain.
* Previous data is required by the system to forecast weather.

**How accuretely we predict the weather?**

The process of predicting weather patterns is a very complicated science. It

requires the need to analyze and decode massive data sets gathered from

thousands of sensors and weather satellites every day.

Identifying patterns in collected data to predict the future is a very strenuous

task. For best results, it also needs to be done in real-time.

But like any kind of forecast, weather forecasting is something of an

educated guess. Since we cannot control the weather, the best meteorologists can

do, is to use past and present data and patterns to attempt to predict the future.

This is especially true to provide information for disaster events. The accuracy of

weather predictions has increased over time, still it is not 100% accurate.

According to some estimates, a seven-day weather forecast is about 80% reliable.

Shorter timescales are more so, with a five-day weather forecast about 90%

correct. Anything longer than seven days, especially ten-day forecasts or longer

tends to be only about 50% accurate. As the atmosphere is constantly changing,

estimates over long periods have proved to be very difficult to model and predict.

Meteorologists achieve this by using computer programs called weather

models to make these forecasts.

### Accurately analyzing weather data

We have now found and imported our weather data into the application so we can perform our data science. Now let’s look at the typical data we can expect within a weather data report.

The typical weather data dataset will include multiple columns such as [temperature, precipitation and wind speed etc](https://www.visualcrossing.com/resources/documentation/weather-data/weather-data-documentation/). If the dataset is using a short time period for each item of data such as an hourly weather forecast or hourly historical observations, then not a lot of post processing will occur on the dataset that is imported. However if the data is aggregated to summarize a day, month or year then multiple weather data observations are aggregated into a single report. There are different ways for to happen and it’s important to understand what a particular aggregated weather variable value is obtained.

Temperature is typically aggregated in three ways – the maximum temperature, minimum temperature and the arithmetic mean of the temperature. The mean temperature can be the mean of all hourly values or the average of the maximum and minimum values. In a typical day, the maximum temperature often occurs in the afternoon and therefore simply reporting the maximum temperature is a good substitute for the overall temperature of day. In some circumstances, maximum temperature does not occur in the afternoon such as when an colder or warmer air masses are moving through a location. In these cases, using the maximum daily temperature to compare against business performance may not produce accurate analysis and results. It is generally better to analyze temperature at the hourly level.

For some applications it is necessary understand more about the maximum and minimum temperature when investigating the typical weather for a time and location. For example, consider the normal temperature for a location in January. We would like to understand the normal maximum temperature (mean maximum temperature) plus also the possible variability. What temperature range do 80% of the days fall between? What is the maximum maximum or minimum maximum temperature possible at this location? The maximum temperature is a good guide to the typical weather at a location, particularly when additional statistical values are considered so a full understanding of the typical temperature and the variability is understood.

Rainfall is typically summed over the aggregation period. The [precipitation coverage](https://www.visualcrossing.com/blog/what-is-precipitation-coverage), the amount of time the rain fell for, is often as important a driver for business metrics as the amount of rain that falls. A short, sharp but heavy thunderstorm at the end of the day in Miami, Florida may well have less impact on tourist activities than a longer but lighter all day light rain. However the former may well produce significantly more rainfall and therefore look worse in the daily weather observation data.

### Sources of Weather Data

Visual Crossing Microsoft Data provides historical weather data and weather forecast data available for download, [web service](https://www.visualcrossing.com/weather-api) and ODATA access.

### 2. LITERATURE REVIEW

**3. MATERIAL AND METHOD**

In this project we have two modules:

* Data gathering and preprocessing
* Applying Algorithm for data set.

**Step by step Project:**

1. In this module first we gather the data for prediction model. Data comes in csv file with visual crossing data. It cames with every month for seperately. We join the every month for result dataset of 2020.
2. In data preprocessing we should make all the attributes nominal values for regression. Also we have 17 attributes in total, but two of them date and name of the city, they are not effect the result because it is all same citys year plan so first we remove this attributes. Our classify attributes are:

* Maximum Temperature
* Minimum Temperature
* Temperature
* Wind Chill
* Heat Index
* Precipatition
* Snow

NUMERIC

* Snow Depth
* Wind Speed
* Wind Direction
* Wind Gust
* Visibility
* Cloud Cover
* Relative Humidity
* Conditions

NOMINAL

We will use the Weka tool & Python language to process the dataset.

**3.1. Material**

Our material is the computer, WEKA tool of university of waikato and Adana’s 2020 daily weather datas.

We used for Data visualization Python Language on Spyder.

**3.1.1. Data Set**

Adana’s 2020 DAILY weather values.

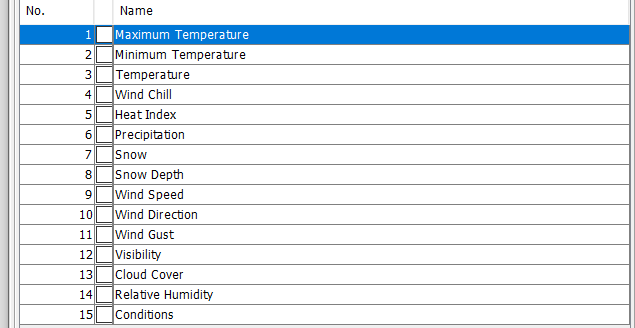
Data set gathered with visual crossing data.

Datas gathered monthly, some days are missing but there has no missing value.

We have total 330 day of the year 2020.

Also it was examined with monthly and all year sets and compared.

**3.1.2. WEKA**

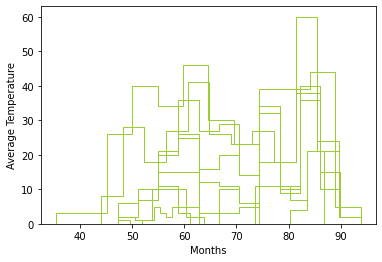


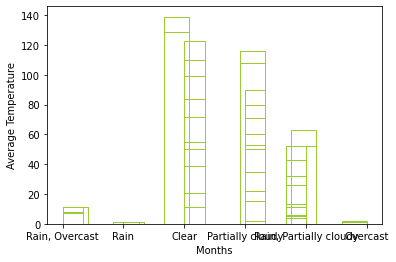
**Figure 1.** WEKA Tool Attributes

**3.2. Method**

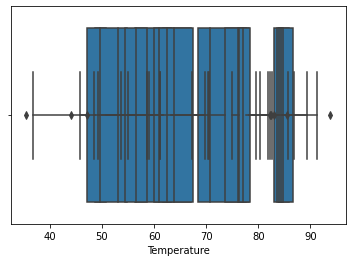
**3.2.1 Data Visualization**

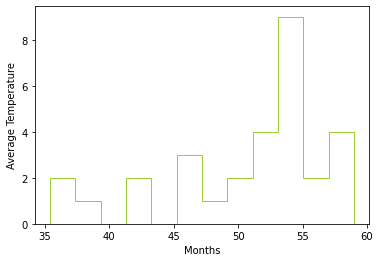
**ATTRIBUTE VISUALIZATION**





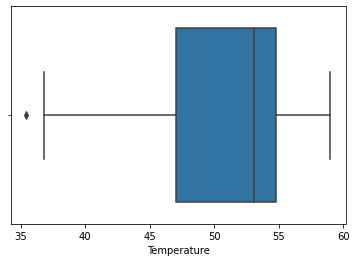
**1.Total of 12 months boxplot graphic**





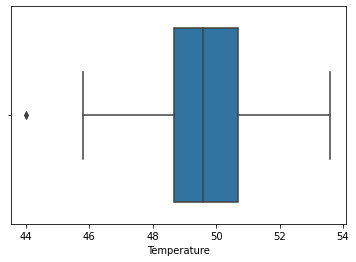
**12 months Histogram Graphic**

**2.Data Frame’s boxplot graphic**

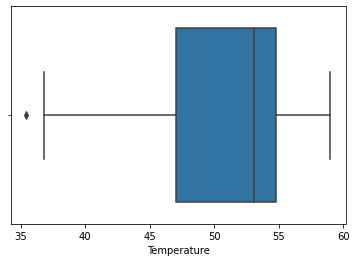


**3.Month by Month Boxplot graphics**

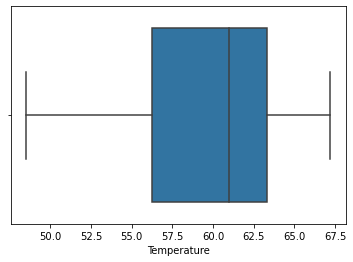
**3.1 January**



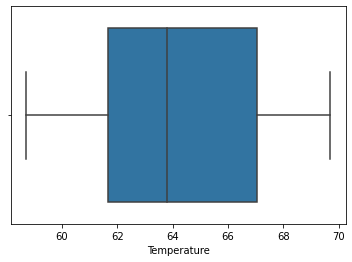
**3.2 February**



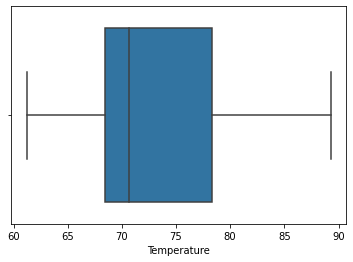
**3.3 March**



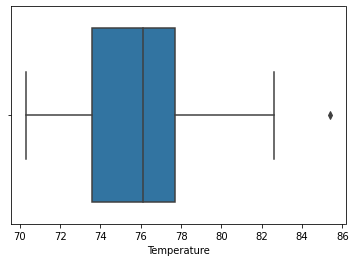
**3.4 April**



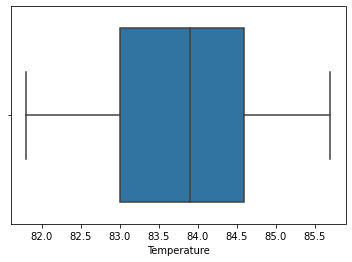
**3.5 May**



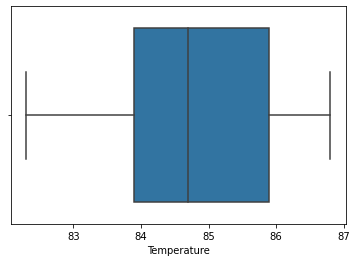
**3.6 June**



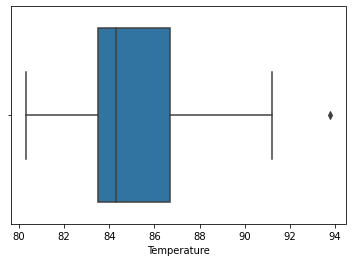
**3.7 July**



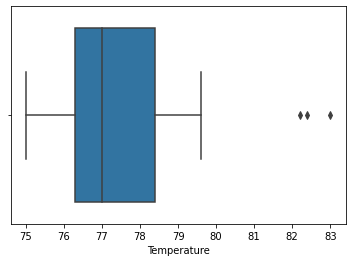
**3.8 August**



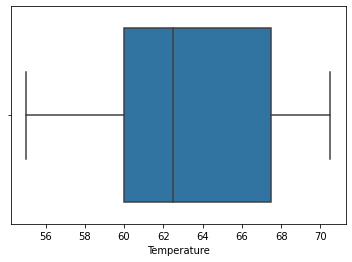
**3.9 September**



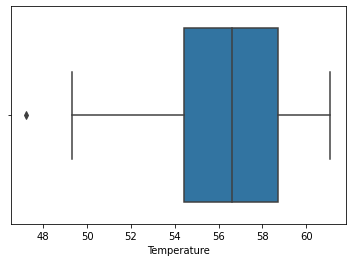
**3.10 October**



**3.11 November**



**3.12 December**



**3.2.3 Preprocessing**

In Weka Preprocessing for Logistic and simple logistic regression we convert the string datas to nominal datas. Because In logistic and Simple Logistic Regression datas can be just nominal or numeric.

What we used in preporcess >unsupervised>StringToNominal>NominalToBinary

In Linear Regression we convert the nominal and string datas to numeric type. Because linear regression work with numeric datas.

we used preporcess >unsupervised>StringToNominal>NominalToBinary for string attributes

NominalToString>StringToWordVector for last condition attribute.

**3.2.1. Data Mining**

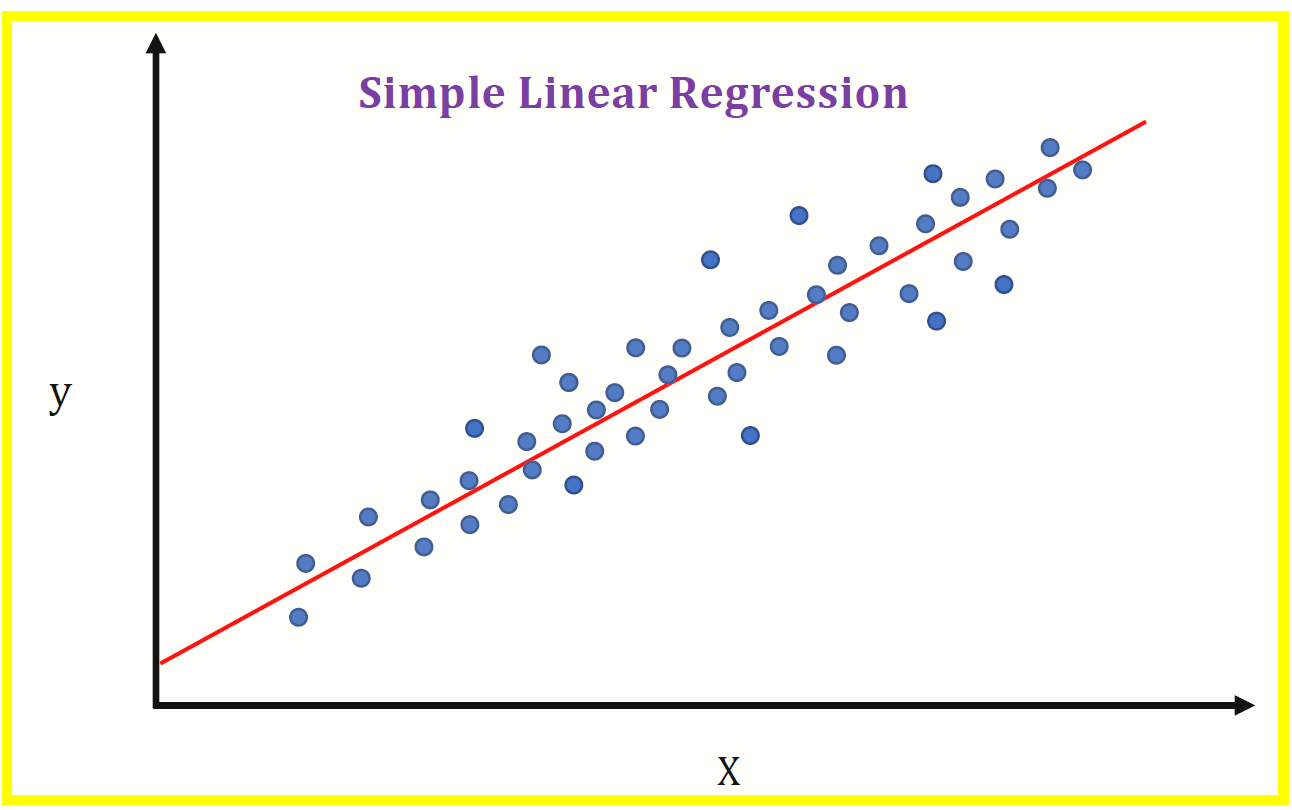
Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data.It is a multi-disciplinary skill that uses machine learning, statistics, AI and database technology.

Regression is a [data mining](https://www.lifewire.com/what-is-data-mining-4784169) technique used to predict a range of numeric values given a particular dataset. For example, regression might be used to predict the cost of a product or service, given other variables.

Regression and classification are data mining techniques used to solve similar problems, but they are frequently confused. Both are used in prediction analysis, but regression is used to predict a numeric or continuous value while classification assigns data into discrete categories.

**1.Linear Regression**

**Linear Regression**is a machine learning algorithm based on **supervised learning.** It performs a **regression task.** Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are consideringand the number of independent variables being used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x).

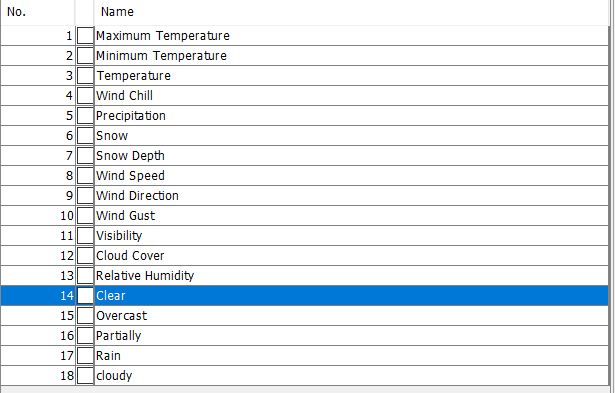
So, this regression technique findsout a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

Generally y is the output and x is the input.

**y=mx+c**

linear regression hypothesis function

Advanced techniques, such as multiple regression, predict a relationship between multiple variables — for example, is there a correlation between income, education and where one chooses to live? The addition of more variables considerably increases the complexity of the prediction. There are several types of multiple regression techniques including standard, hierarchical, setwise and stepwise, each with its own application.



Linear Regression’s Numeric Attributes

**2.Standard Multiple Regression**

**it** considers all predictor variables at the same time.

1) what is the relationship between income and education (predictors) and choice of neighborhood (predicted)

2) to what degree do each of the individual predictors contribute to that relationship?

**3**. **Stepwise Multiple Regression**

**It a**nswers an entirely different question. A stepwise regression algorithm will analyze which predictors are best used to predict the choice of neighborhood — meaning that the stepwise model evaluates the order of importance of the predictor variables and then selects a relevant subset. This type of regression problem uses "steps" to develop the regression equation. Given this type of regression, all predictors may not even appear in the final regression equation.

**4.Hierarchial Regression**

It is like stepwise, is a sequential process, but the predictor variables are entered into the model in a pre-specified order defined in advance, i.e. the algorithm does not contain a built-in set of equations for determining the order in which to enter the predictors. This is used most often when the individual creating the regression equation has expert knowledge of the field.

**5.Setwise Regression**

It is also similar to stepwise but analyzes sets of variables rather than individual variables.

**6.Logistic Regression**

Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables.

So: Logistic regression is the correct type of analysis to use when you’re working with binary data. You know you’re dealing with binary data when the output or dependent variable is dichotomous or categorical in nature; in other words, if it fits into one of two categories (such as “yes” or “no”, “pass” or “fail”, and so on).

However, the independent variables can fall into any of the following categories:

**Continuous**—such as temperature in degrees Celsius or weight in grams. In technical terms, continuous data is categorized as either **interval data**, where the intervals between each value are equally split, or **ratio data**, where the intervals are equally split **and** there is a true or meaningful “zero”. For example, temperature in degrees Celsius would be classified as interval data; the difference between 10 and 11 degrees C is equal to the difference between 30 and 31 degrees C, but there is no true zero—a temperature of zero degrees does not mean there is “no temperature”. On the other hand, weight in grams would be classified as ratio data; it has the equal intervals **and** a true zero. In other words, if something weighs zero grams, it truly weighs nothing.

**Discrete, ordinal**—that is, data which can be placed into some kind of order on a scale. For example, if you are asked to state how happy you are on a scale of 1-5, the points on the scale represent ordinal data. A score of 1 indicates a lower degree of happiness than a score of 5, but there is no way of determining the numerical value between each of the points on the scale. Ordinal data is the kind of data you might get from a customer satisfaction survey.

**Discrete, nominal**—that is, data which fits into named groups which do not represent any kind of order or scale. For example, eye color may fit into the categories “blue”, “brown”, or “green”, but there is no hierarchy to these categories.

So, in order to determine if logistic regression is the correct type of analysis to use, ask yourself the following:

**Is the dependent variable dichotomous?** In other words, does it fit into one of two set categories? Remember: The dependent variable is the outcome; the thing that you’re measuring or predicting.

**Are the independent variables either interval, ratio, or ordinal?** See the examples above for a reminder of what these terms mean. Remember: The independent variables are those which may impact, or be used to predict, the outcome.

In addition to the two criteria mentioned above, there are some further requirements that must be met in order to correctly use logistic regression. These requirements are known as “assumptions”; in other words, when conducting logistic regression, you’re assuming that these criteria have been met. Let’s take a look at those now.

**We chose the linear and logistic regression method to make our weather forecasting.**

**3.2.2. Data Preprocessing Methods**

For linear Regression; values have to be numeric value to apply the datas. We are using weka to preprocesses.

We have to our datas be numeric so, we convert the our string and nominal datas are numeric data.

In preprocessing section, first we used

-StringToNominal command.

(1)

**3.2.3. Classification Algorithms/Clustering Algorithms/Regression Algorithms/Association Rule Mining Algorithms**

**3.2.3.1 adANA’S WEATHER FORECASTING CONCLUSIONS MONTHLY**

**january**

|  |  |
| --- | --- |
| clear | **13** |
| partıally cloudy | **10** |
| raın,partıally cloudy | **9** |

**fEBRUARY**

|  |  |
| --- | --- |
| RAIN,PARTIALLY CLOUDY | **11** |
| PARTIALLY CLOUDY | **8** |
| OVER CAST | **1** |

**march**

|  |  |
| --- | --- |
| raın,partıally cloudy | **13** |
| clear | **12** |
| partıally cloudy | **7** |

**aprıl**

|  |  |
| --- | --- |
| raın,partıally cloudy | **6** |
| clear | **11** |
| partıally cloudy | **15** |

**may**

|  |  |
| --- | --- |
| Raın,partıally cloudy | **6** |
| partıally cloudy | **11** |
| clear | **15** |

**june**

|  |  |
| --- | --- |
| partıally cloudy | **15** |
| clear | **5** |
| raın,partıally cloudy | **5** |

**july**

|  |  |
| --- | --- |
| partıtally cloudy | **18** |
| clear | **6** |
| overcast | **1** |

**august**

|  |  |
| --- | --- |
| partıtally cloudy | **13** |
| raın,overcast | **1** |
| clear | **10** |
| raın,partıtally cloudy | **1** |

**september**

|  |  |
| --- | --- |
| partıally cloudy | **13** |
| clear | **11** |
| raın,partıally cloudy | **1** |

**october**

|  |  |
| --- | --- |
| partıtally cloudy | **7** |
| cloudy | **18** |

**november**

|  |  |
| --- | --- |
| partıtally cloudy | **3** |
| raın,overcast | **3** |
| clear | **17** |
| raın,partıtally cloudy | **2** |

**DEcember**

|  |  |
| --- | --- |
| partıtally cloudy | **2** |
| raın,overcast | **7** |
| clear | **11** |
| raın,partıtally cloudy | **4** |
| raın | **1** |

### 4. results and dıscussıon

**Table 1.** The Results of Algorithms

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **RMSE** |
| Linear Regression | %98 | - | - | - | - | 0.0913 |
| Logistic | %70 | 0,730 | 0,7 | 0,714 | 0,854 | 0.3837 |
| Simple Logistic | %90 | 0,9 | 0,9 | 0,9 | 0,953 | 0.2337 |

### 5. conclusıon

**5.1.Logistic Regression confusion matrix predicts for**

* January –Clear
* February– Rain,Partially Cloudy
* March– Rain,Partially Cloudy
* April–Rain,Partially Cloudy
* May– Partially Cloudy,Clear are equal
* June– Partially Cloudy
* July– Partially Cloudy
* August– Partially Cloudy
* September– Partially Cloudy
* October– Clear
* November-Clear
* December–Clear
* DataFrame:Partially Cloudy

**5.2 Simple Logistic Regression confusion matrix predicts for**

* January –Clear
* February– Partially Cloudy
* March–Rain,Partically Cloudy or Clear.
* April–Clear
* May– Clear
* June–Partially Clody
* July– Partially Cloudy
* August– Partially Cloudy
* September– Partially Cloudy
* October– Partially Cloudy
* November- Partially Cloudy
* December– Partially Cloudy
* DataFrame:Partially Cloudy

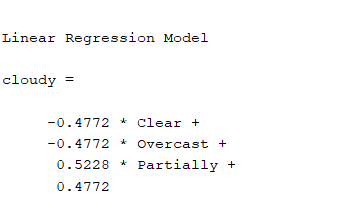
**Dataframe Results**

According to Logistic and Simple Logistics Regression, the most used weather condition in Adana in 2021 is partly cloudy weather, together with the weather results of 2020 by months in Adana.

**5.3. Linear Regression Matrix Predict**

* January –Partially Cloudy
* February– Partially Cloudy
* March–Partically Cloudy
* April–Clear or Partically Cloudy
* May– Partically Cloudy
* June–Partially Cloudy
* July– Partially Cloudy
* August– Partially Cloudy
* September– Partially Cloudy
* October–Clear
* November-Partially Cloudy
* December–Clear

Dataframe:Partially Cloudy



1.2 Dataframe Linear Regression Results on WEKA

**5.4.Final Results**

Weather in Adana is partly cloudy on most days, but it may be clear on some days. While Rain and Overcast weather are not so visible. No snow has been found in the one-year weather history. Therefore, the weather is expected to be mostly partly cloudy in 2021. Not waiting for a snowfall in 2021.

**1.** Coduto, P. D., 2001. Foundation Design: Principal and Practices (Second Edition), Prentice Hall, New Jersey, p. 883.

**2.** Terzaghi, K., Coduto, P. D., 1943. Theoretical Soil Mechanics, Wiley, p. 510-512, New York.

**3.** https://www.visualcrossing.com/resources/blog/how-to-find-weather-data-for-data-science/

**4.** Python and Data Science from Pusula Yayınları

…

**10.**