Will It Rain Tomorrow? Application of Data Mining Techniques for Forecasting Rainfall in Yerevan

Group Members: Maria Davoodian, Silva Arakelyan, Naira Matosyan,
Anna Baghumyan, Arusyak Hakobyan

Instructor: Sean Reynolds

1. Introduction

Our daily lives are significantly influenced by the weather. Therefore, forecasting it can be very informative and beneficial, since it prepares people for the forthcoming climate changes, decreases the uncertainty in weather behavior and the damages it may cause. And the biggest challenge is to make the forecast as accurate as possible. To that end, meteorologists have tried to predict the weather using different methods, some being more accurate than the others. And nowadays, one of the most commonly used techniques for forecasting weather is data mining.

Data mining offers many techniques for analyzing historical data and extracting the significant patterns from it. Then, these extracted patterns can be used to predict future data trends. Hence, data mining allows us to observe changes in weather patterns from previous years and use them to predict the weather conditions of the upcoming years.

This paper analyzes the weather data of Yerevan (Zvartnots International Airport) from 2017 and uses it to predict the rainfall in 2018. For that purpose, it first studies the relationships between the environmental factors and extracts the initial patterns from the data. Then it applies three data mining techniques - Naive Bayes, Decision Trees and k-Nearest Neighbors methods, to make the predictions.

2. Methodology

2.1. Data Preparation

Data used in this project was collected from Zvartnots International Airport. The case data covered a period of one year (03/31/17 - 03/30/18) and included attributes such as `Temperature` (high, average, low), `Dew point` (high, average, low), `Humidity` (high, average, low), `Sea Level Pressure` (high, average, low), `visibility` (high, average, low), `Wind` (high, average), `Gust Wind` (high), `Precipitation, Events` (rain, thunderstorm, etc.) and the `Date`.

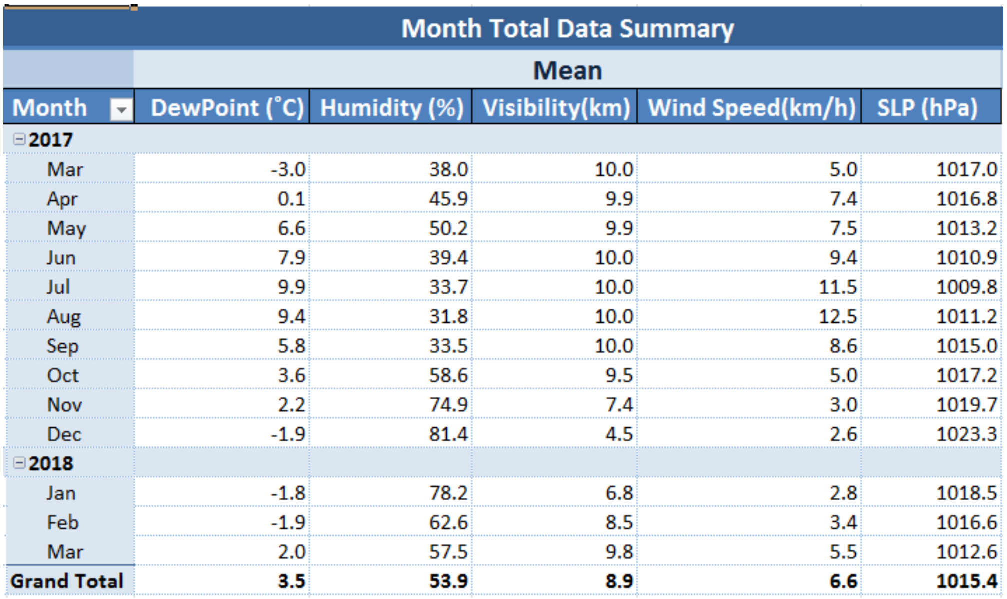
After the retrieval of the data, it was cleaned and formatted for improving the quality of the data and ensuring its ease of use. For data cleaning and formatting, MS Excel was used. First, the extraneous headers, hyperlinks and all other unnecessary information were removed from the data. Then the label rows and columns were frozen, all cells were given proper formatting and names were given to variables. Furthermore, for identification of missing data and/or all zero columns, conditional formatting was used. This step revealed that `Precipitation` contained all zeros and `Gust Wind` was filled with noisy data, so these two attributes were overlooked when doing a prediction.

When calculating the correlations between the variables, it was important to have their correlation with rainfall as well. However, since `Precipitation` was filled with only zeros, the number of rainy days was used instead. And, in order to obtain the number of rainy days, Excel IF function was used and it returned 1 in the case "Rain" was contained in the cell, and 0 otherwise. As the result, the categorical variable was converted to a numeric variable, which, in turn, made it possible to calculate correlations.

After the data cleaning and formatting, the Excel file was converted into CSV format as well, in order to later use it in R. For having a convenient CSV file for several experiments in weather prediction, the average columns were chosen as features. Moreover, the `Events` variable was transformed into ”Rain”, if the cell contained ”Rain”, and ”No Rain”, if it did not.

2.2. Summary and Visualization

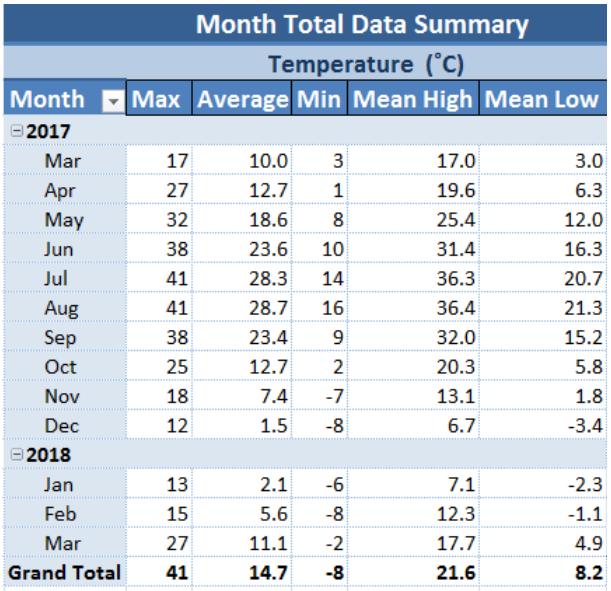
In order to see the initial patterns, the data was summarized and visualized. In particular, Month Total Data and Year Total Data summaries were made using pivot tables. For the variables `Dew Point, Humidity, Visibility, Wind Speed` and `Sea Level Pressure`, the summaries were made using the average value only (see Figure 1 and Figure 3). For the Temperature, however, besides the average value, maximum, minimum, mean high and mean low values were also needed for further analysis (see Figure 2 and Figure 4).



Month Total Data Summary

	Mean				
Month	**DewPoint (˚C)**	**Humidity (%)**	**Visibility(km)**	**Wind Speed(km/h)**	**SLP (hPa)**
⊟2017					
Mar	-3.0	38.0	10.0	5.0	1017.0
Apr	0.1	45.9	9.9	7.4	1016.8
May	6.6	50.2	9.9	7.5	1013.2
Jun	7.9	39.4	10.0	9.4	1010.9
Jul	9.9	33.7	10.0	11.5	1009.8
Aug	9.4	31.8	10.0	12.5	1011.2
Sep	5.8	33.5	10.0	8.6	1015.0
Oct	3.6	58.6	9.5	5.0	1017.2
Nov	2.2	74.9	7.4	3.0	1019.7
Dec	-1.9	81.4	4.5	2.6	1023.3
⊟2018					
Jan	-1.8	78.2	6.8	2.8	1018.5
Feb	-1.9	62.6	8.5	3.4	1016.6
Mar	2.0	57.5	9.8	5.5	1012.6
Grand Total	**3.5**	**53.9**	**8.9**	**6.6**	**1015.4**

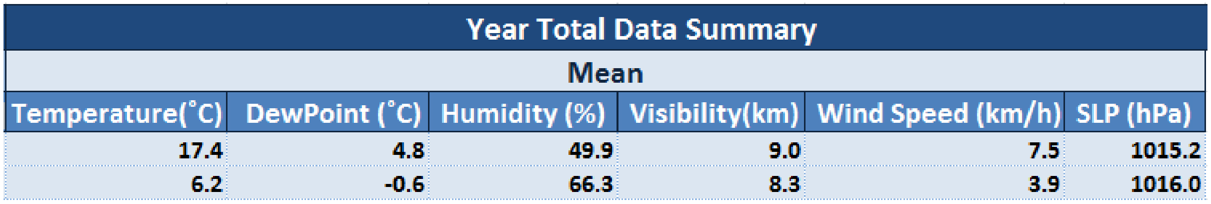
Figure 1: Month Total Data Summary for Dew Point, Humidity, Visibility, Wind and Sea Level Pressure



Month Total Data Summary

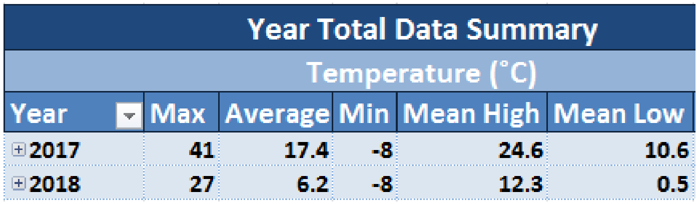
	Temperature (˚C)				
Month	**Max**	**Average**	**Min**	**Mean High**	**Mean Low**
⊟2017					
Mar	17	10.0	3	17.0	3.0
Apr	27	12.7	1	19.6	6.3
May	32	18.6	8	25.4	12.0
Jun	38	23.6	10	31.4	16.3
Jul	41	28.3	14	36.3	20.7
Aug	41	28.7	16	36.4	21.3
Sep	38	23.4	9	32.0	15.2
Oct	25	12.7	2	20.3	5.8
Nov	18	7.4	-7	13.1	1.8
Dec	12	1.5	-8	6.7	-3.4
⊟2018					
Jan	13	2.1	-6	7.1	-2.3
Feb	15	5.6	-8	12.3	-1.1
Mar	27	11.1	-2	17.7	4.9
Grand Total	**41**	**14.7**	**-8**	**21.6**	**8.2**

Figure 2: Month Total Data Summary for Temperature



Year Total Data Summary					
Mean					
Temperature(°C)	DewPoint (°C)	Humidity (%)	Visibility(km)	Wind Speed (km/h)	SLP (hPa)
17.4	4.8	49.9	9.0	7.5	1015.2
6.2	-0.6	66.3	8.3	3.9	1016.0

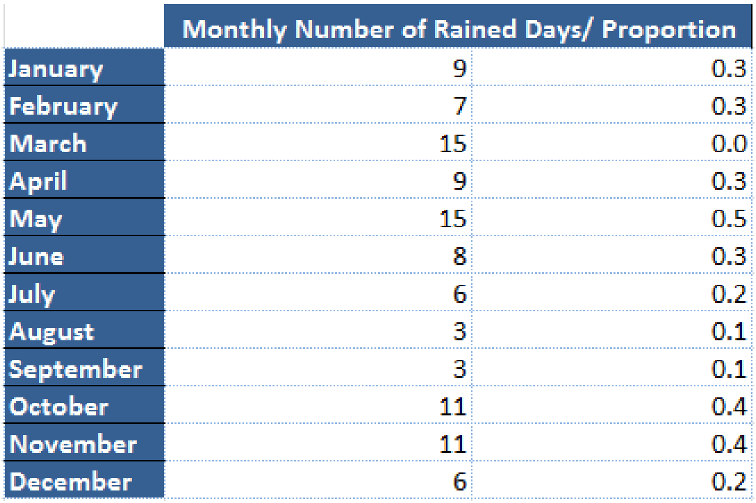
Figure 3: Year Total Data Summary for Dew Point, Humidity, Visibility, Wind and Sea Level Pressure



Year Total Data Summary					
	Temperature (°C)				
Year	Max	Average	Min	Mean High	Mean Low
⊞2017	41	17.4	-8	24.6	10.6
⊞2018	27	6.2	-8	12.3	0.5

Figure 4: Year Total Data Summary for Temperature

Another important information which is necessary to have for further analysis was the number/proportion of rainy days per month. For this reason, the Monthly Rainy Days summary was calculated using the COUNTIF formula (see Figure 5).



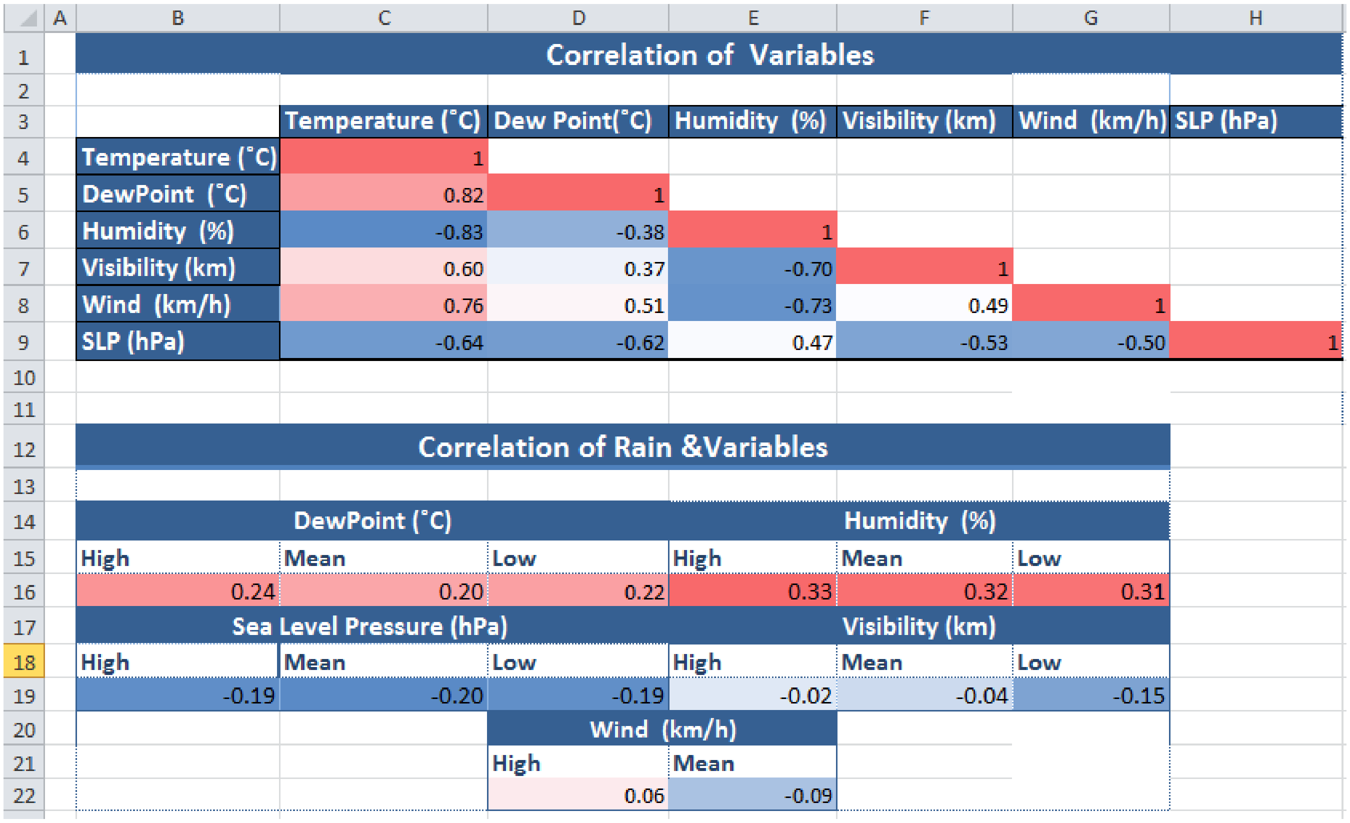
	Monthly Number of Rained Days/ Proportion	
January	9	0.3
February	7	0.3
March	15	0.0
April	9	0.3
May	15	0.5
June	8	0.3
July	6	0.2
August	3	0.1
September	3	0.1
October	11	0.4
November	11	0.4
December	6	0.2

Figure 5: The Summary Table For Monthly Rainy Days

One of the most important steps in pattern analysis was to consider the correlation between the variables. For this purpose, two correlation matrices were calculated:

- For calculating correlations among all variables except rain, Data Analysis Toolbox was used
- For calculating the correlations between all variables' high/mean/low values and Rain, CORREL function was used.

Then, the correlation matrices were given conditional formatting to mark correlation coefficients from low to high by using the coloring from bright to dark (see Figure below).



Correlation of Variables

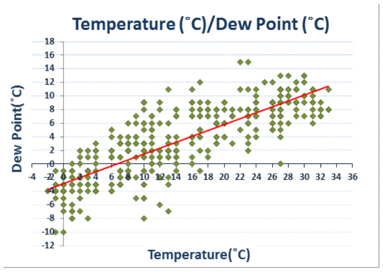
	Temperature (˚C)	Dew Point(˚C)	Humidity (%)	Visibility (km)	Wind (km/h)	SLP (hPa)
Temperature (˚C)	1					
DewPoint (˚C)	0.82	1				
Humidity (%)	-0.83	-0.38	1			
Visibility (km)	0.60	0.37	-0.70	1		
Wind (km/h)	0.76	0.51	-0.73	0.49	1	
SLP (hPa)	-0.64	-0.62	0.47	-0.53	-0.50	1

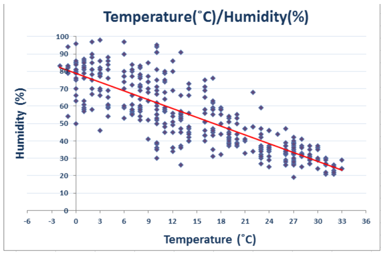
Correlation of Rain &Variables

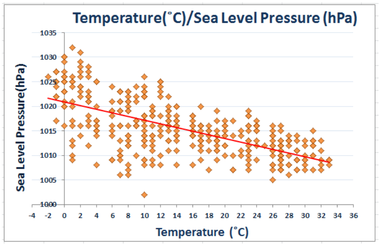
DewPoint (˚C)			Humidity (%)		
High	Mean	Low	High	Mean	Low
0.24	0.20	0.22	0.33	0.32	0.31
Sea Level Pressure (hPa)			Visibility (km)		
High	Mean	Low	High	Mean	Low
-0.19	-0.20	-0.19	-0.02	-0.04	-0.15
		Wind (km/h)			
		High	Mean		
		0.06	-0.09		

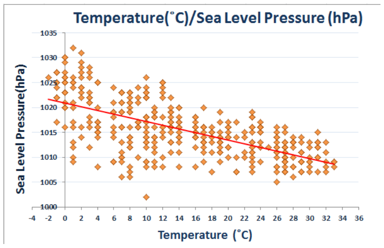
After constructing the summary tables, the following visualizations were done:

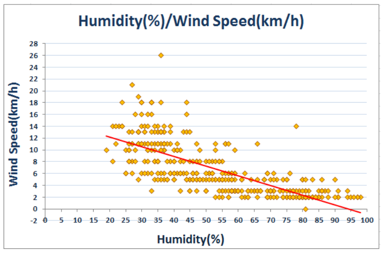
- Correlation Graphs - After the calculation of the correlation coefficients among all variables, the relationships between those variables that had very high or very low correlations were visualized with the help of scatterplots.



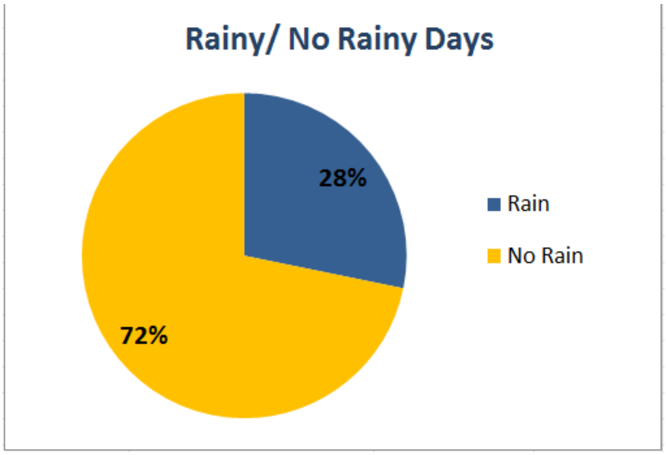








- The Graph demonstrating the Proportion of Rained and Not Rained days during the whole year. For plotting the pie chart, the number of rained and not rained days were needed. To that end, the function COUNTIF was used.



- The Graph demonstrating monthly proportion of rained days and maximum, minimum and mean temperatures - For the creation of this graph, all calculations which were needed (monthly max/mean/min temperatures) were already done in the summaries. This graph is analyzed in the next section.

- The Graph demonstrating monthly mean wind speed, sea level pressure, humidity, dew point and the proportion of rainy days - These graphs were also plotted by using the summaries. All these graphs are demonstrated in next section.

2.3. Initial Pattern Analysis

2.3.1. The Geographic Location of Yerevan

When analyzing the initial climate patterns, it is important to take into account the geographic location of Yerevan. To that end, this paper used a system for climate classification called *Kppen climate classification.*

The Kppen climate classification system divides climates into five major climate groups: A (tropical), B (arid), C (temperate), D (continental), and E (polar). These groups are subgrouped further by second and third letter, which represent the seasonal precipitation type and level of heat, respectively.

By Kppen climate classification, Yerevan falls into BSk group, that is, has ”cold semi-arid climate”. The distinguishing features of this type of climate are:

- warm to hot, dry summers with the temperature in August reaching up to 40 °C

- cold winters, which carry snowfall and freezing temperatures with January often being as cold as 15 °C and lower

- major daily temperature swings, sometimes by as much as 20 °C

The cold semi-arid climate is typically found in continental interiors some distance from large bodies of water. Yerevan, being on a plain surrounded by mountains and distant from the sea and its effects, has

- long, hot, dry summers

- short, cold, relatively wet winters

- even wetter springs and autumns.

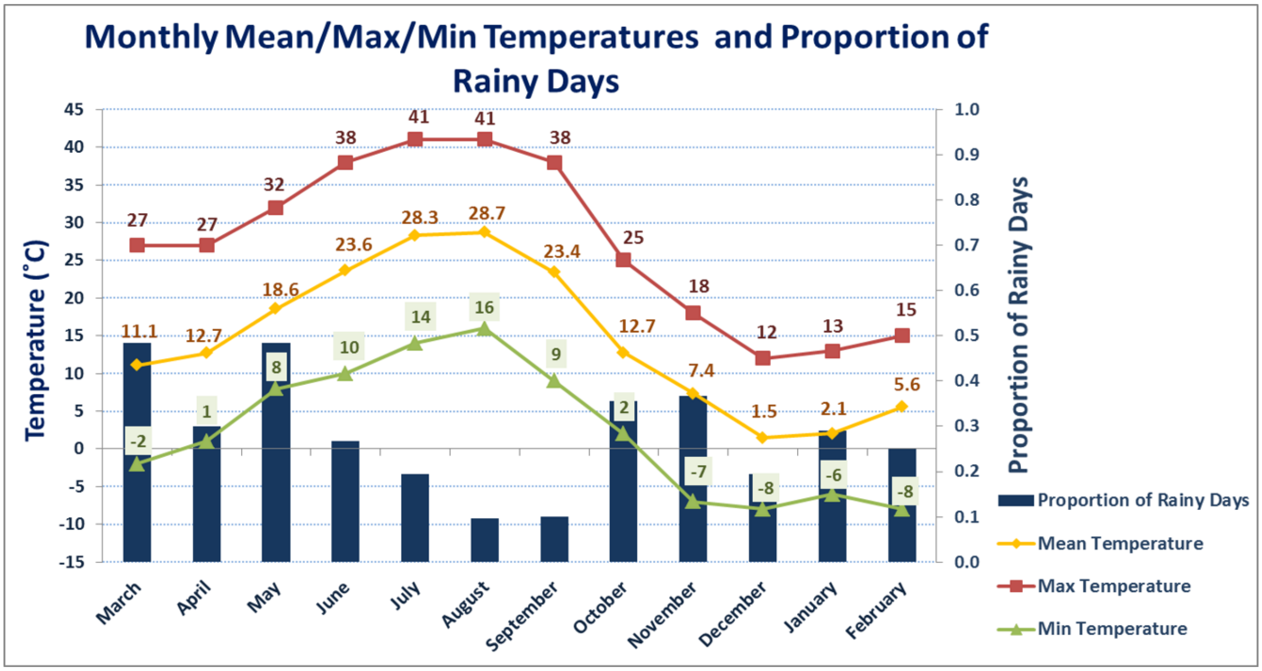


Figure 6: Seasonal weather based on 2017-2018 data: In Spring, Fall and Winter when temperature is low, precipitation is high. In Summer, as the temperature increases, the rate of rainfall decreases.
The highest temperature in August is 41°C(≈ 40°C). The lowest temperature in winter is -8°C, which is expected as last year was unusual one.

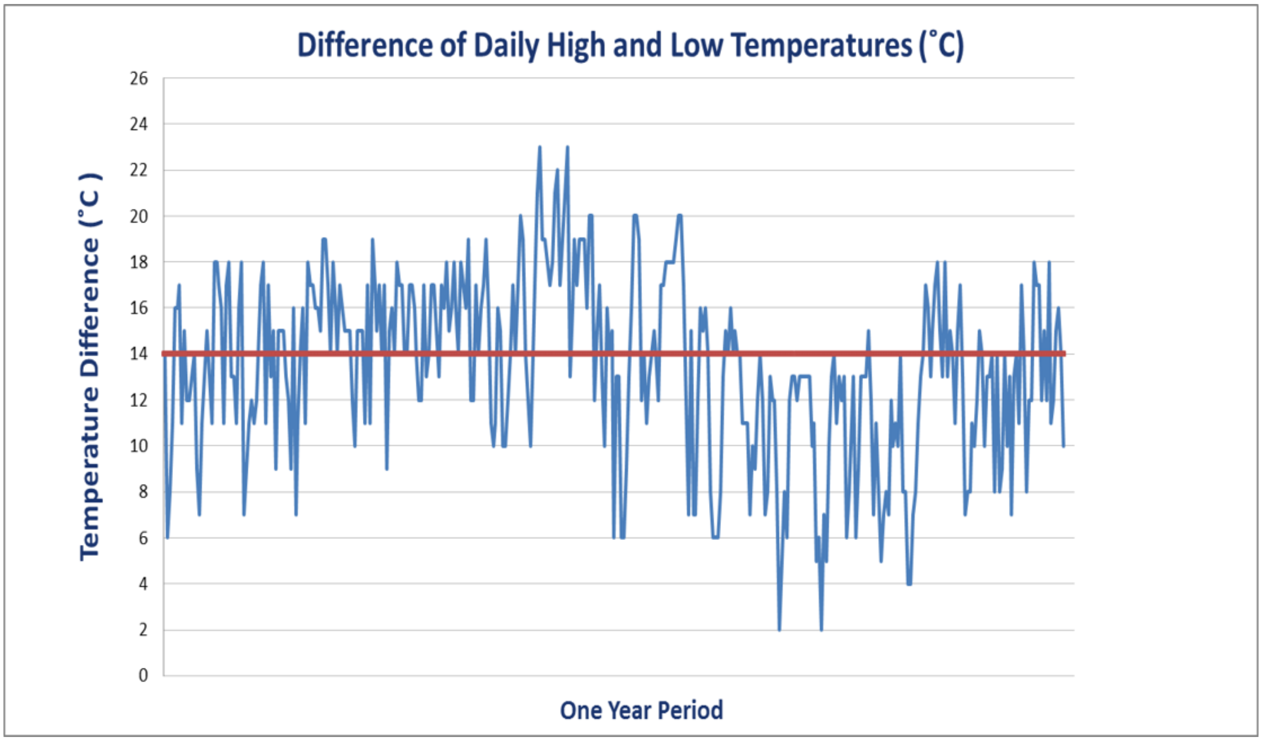


Figure 7: Average daily temperature difference, based on 2017-2018 data, is close to 14. This implies that Yerevan has major temperature swings between day and night.

2.3.2. Relationships Between Variables

Many environmental factors affect rainfall, although the major influencing factors are atmospheric temperature, wind speed, pressure and humidity. The mentioned determine the type of the rainfall, its occurrence and velocity as well as the amount of precipitation.

Temperature/Dew Point and Rainfall: The main relationship between temperature and precipitation is that at higher temperatures, the precipitation amounts are low, while at lower temperatures, the precipitation amounts are high. Therefore, over land, The dominating correlations are negative as higher temperature implies more sunshine, thus less evaporative cooling. Positive correlations are recorded in colder conditions as the precipitation amounts are restricted by the capacity of the atmosphere to hold water. Relative humidity is affected directly by air temperature as relative humidity drops as the temperature rises, the reason why is because for a certain amount of moister in the atmosphere, higher temperature indicates low humidity rates. The dew point indicates the water vapor concentration in the air and temperature varies more then dew point does. Although rainfall is likely when temperature is equal to the dew point as at that point condensation occurs since humidity reaches hundred percent. Figure 8 is the visual demonstration plotted from the data taken from 2017 from Yerevan showing the relationship between Temperature and Dew Point.

Wind Speed and Rainfall: Winds carry certain amounts of water, thus they greatly influence the perception of an area. They consume moisture as they pass over a body of water, thus the larger the source is, the more moisture and therefore, precipitation will they bring. Rainfall loss rate is directly associated with wind speed as when the latter and/or the intensity of rainfalls decrease, rainfall loss rates increase. Faster winds are linked with more precipitation, while calmer winds tend to not carry large amount of moisture. Depending on the geographical location this link changes and its increase is dependent on the atmospheric moisture. Figure 9 demonstrates the relationship between the intensity of wind speed and rainy days.

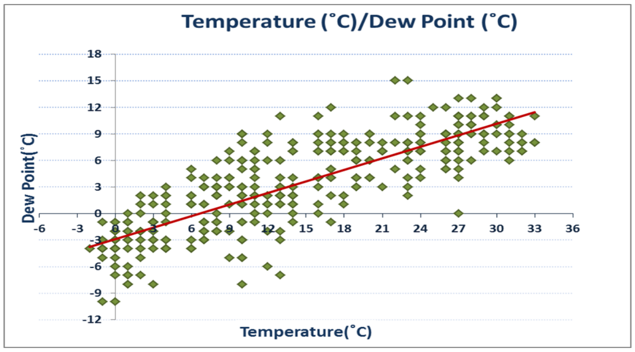


Figure 8

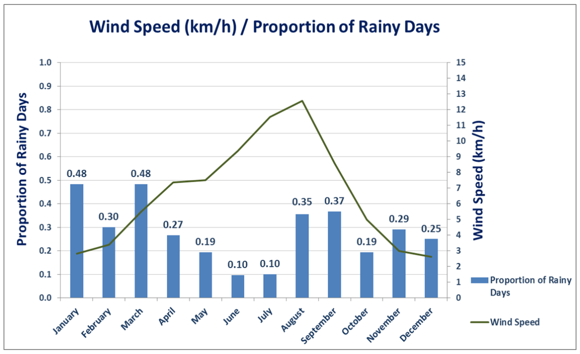


Figure 9

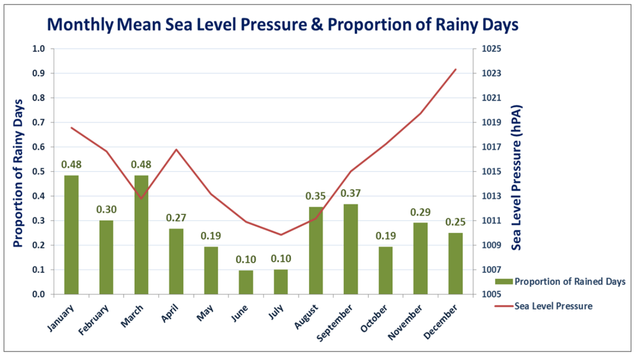


Figure 10

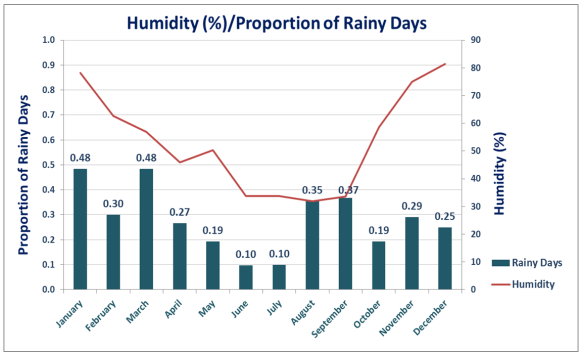


Figure 11

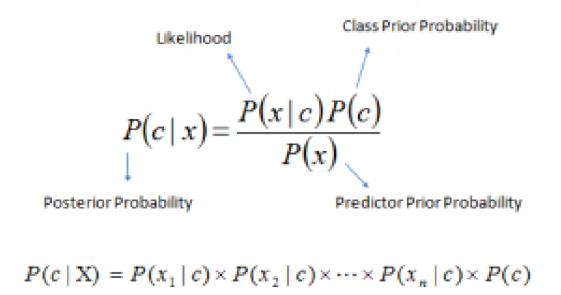
Pressure and Rainfall: Vertical movements of the air cause low and high pressure areas as at lower pressures air rises but drops at higher pressures. Rainfall is significantly associated with low and high pressure. The probability of rainfall increases, when the pressure falls. Intense winds are associated with low barometric pressures, which means that air masses cool moving upwards, thus condensation occurs, clouds form, precipitation increases thus it rains. While at the probability of rainfall decreases, as the pressure rises, since winds are very weak or completely absent at high pressure areas. As the air masses move downwards, evaporation takes place, thus no clouds form resulting in no rainfall. Figure 10 visualizes the relation between SLP and proportion of rainy days of Yerevan from 2018.

Humidity and Rainfall: The environmental factor humidity indicates the amount of water vapor in the air, which depends on the evaporation from different sources of water. This means that the likelihood of rain is a lot more on the coast then it is inland. Low humidity implies reduced rainfall as sometimes the droplets of rain do not reach the surface of the Earth. Although relative humidity increases because of rainfalls, thus if the humidity below a cloud is high or low, evaporation rates vary, increasing and reducing rainfall respectively. Figure 11 clearly shows the dependency of rainfall on humidity.

2.4. Classification Algorithms

2.4.1. Naive Bayes Classifier

Naive Bayes is a probabilistic classifier which is based on the Bayes Theorem (see Figure 12):



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Figure 12: Bayes Rule

It uses the Bayesian equation for calculating the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction. The name Naive comes from the fact that this classifier makes a naive assumption that the data features are independent (uncorrelated).

For categorical variables, Naive Bayes works perfectly, yet for numerical ones, it adds one more assumption: the variable has a normal distribution.

There are three event models for Naive Bayes algorithm:

1. Gaussian Naive Bayes: used for continuous variables having Gaussian (Normal) distribution.

2. Multinomial Naive Bayes: used for discrete variables.

3. Bernoulli Naive Bayes: used when inputs are independent boolean features.

In this project, Naive Bayes was implemented using R, which supports a package called e1071 providing the Naive Bayes training function (Gaussian Naive Bayes).

2.4.2. Decision Tree Classifier

Decision Tree Classifier is another classification algorithm. At first, this algorithm uses the training dataset to extract patterns, then, based on those patterns, it models rules and arranges them in a tree structure. And, when making the prediction for the new observation, it step-by-step checks which rules are satisfied and, according to that, assigns labels to the new observations.

There are three types of decision trees:

1. Univariate tree: uses one attribute for testing a condition at any particular node of the tree.

2. Multivariate tree: uses more than one attribute for testing a condition at the branch while splitting.

3. Hybrid (heterogeneous) tree: uses more than one algorithm to build the tree.

For this project, a univariate decision tree was used.

2.4.3. k-Nearest Neighbors Classifier

This algorithm takes the k nearest neighbors of the new observation, then assigns to this observation the class that is most common among its k nearest neighbors (i.e., it is classified by majority voting).

Here Euclidean distance is used as distance metric,

$$d(Y,Z) = \sqrt{\sum_{i=1}^{n}(y_i - z_i)^2} \tag{1}$$

where Y is an n-dimensional observation from the training dataset and Z is an n-dimensional observation from the testing dataset for which class is to be predicted.

And the number of neighbors (k) depends on the size of the training dataset.

3. Implementation

In the beginning, the data was divided into training (80%) and testing(20%) sets. The algorithms were constructed on the training set, and then, with the use of testing set, their accuracy was assessed. For the assessment of the accuracy, two measures were considered:

- Overall accuracy - the sum of the true positives and true negatives divided by the total number of tested observations

- Area under the ROC curve - the area under the curve when the true positive rates are plotted against false positive rates. The closer the obtained value is to 1, the more accurate is the prediction

Naive Bayes - After running the Naive Bayes algorithm on the training set and constructing confusion matrix based on the testing set, the accuracy of the model was approximately (65%), i.e., the labels of 47 out of 72 observations from the testing set were predicted correctly.

The area under the ROC curve is approximately 0.63, which is not bad, but it is not very close to 1.

Desicion Tree - For constructing the model, the Decision tree algorithm first determined the variables that affect rainfall the most. Those variables are humidity, dew point and sea level pressure.

The accuracy of this algorithm is (80%), which means that 60 observation out of 72 were predicted correctly.

The area under the ROC curve is approximately 0.76

kNN - The kNN algorithm is implemented using 19 as the number of neighbors. This number is chosen by using the method of cross-validation (trying several values for it and choosing the best one with minimal error). The accuracy of the model is approximately (77.8%), and the area under the ROC curve was approximately 0.73.

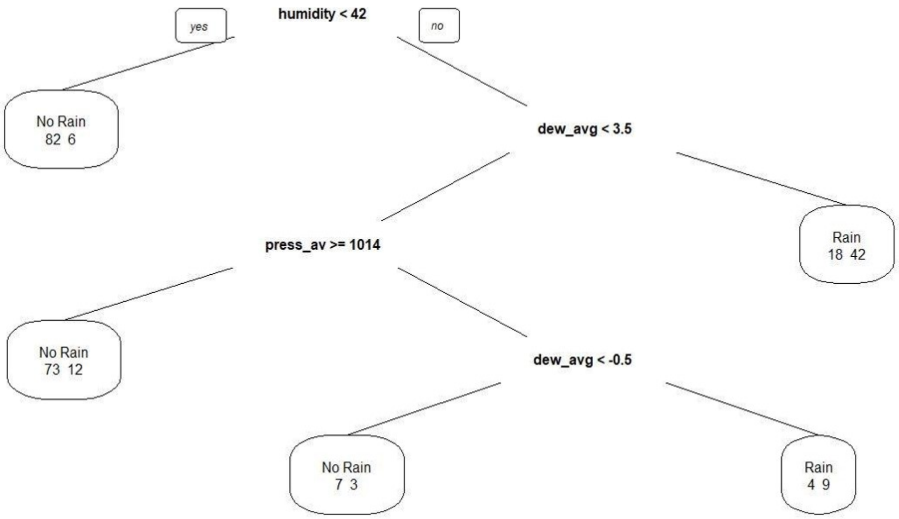


Figure 13: The Decision Tree Generated by R

For the prediction, the following observation was used:
Date: 05/10/2018
Average temperature: 16 $°C$
Average dew point: 9 $°C$
Average humidity: 72 %
Average sea level pressure: 1007 hPa
Average visibility: 10 km
Average wind speed: 5 km/h

Two of the models - Naive Bayes and Decision tree, predicted the label to be ”Rain”, and, since two out of three models predicted that it will rain on this day, and the model with highest overall accuracy and area under the curve was one of them, the final prediction will be that it will rain of 05/10/2018.

4. Conclusion

In conclusion, although there are many other methods for weather prediction starting from observing the sky to computerized methods, data mining provides an accurate alternative. The aim of the proposed project is to present how data mining allows us to predict the rainfall of Yerevan for 2018 using the weather data from 2017 with three different approaches. The concept behind the three classifications is analyzing the weather data from previous years and by extracting patterns and trends predict the rainfall for the upcoming years. This project applies three data mining classifiers - Naive Bayes, Decision Trees and kNN, to build statistical models that can predict the rainfall of Yerevan with an up to (80%) accuracy. Furthermore, it demonstrates how rainfall is influenced by the geographical location of the observed region, by other environmental factors and how their dependency and correlations can help us extract patterns from historic data.

References

[1] Birba, M., Kondilis, T. (2016, December 5). Science in School www.scienceinschool.org. Wind and rain: meteorology in the classroom www.scienceinschool.org Retrieved May 20, 2018 from
http://www.scienceinschool.org/content/wind-and-rain-meteorology-classroom

[2] Back, L.E., Bretherton, C.S. (2205, October 15). AMS Journals. The Relationship between Wind Speed and Precipitation in the Pacific ITCZ: Journal of Climate: Vol 18, No 20. Retrieved May 20, 2018 from http://journals.ametsoc.org/doi/10.1175/JCLI3519.1

[3] Tiwari, P. (n.d.). Geography Notes - Exclusive Notes on Geography. Humidity and Precipitation (Useful Notes). Retrieved May 20, 2018 from http://www.geographynotes.com/articles/humidity-and-precipitation-useful-notes/816

[4] (n.d.). Climates to Travel - world climate guide. Climate in Armenia: temperature, precipitation, when to go, what to pack. Retrieved May 2, 2018, from http://www.climatestotravel.com/climate/armenia

[5] (n.d.). National Geographic Society - National Geographic Society. Climate - National Geographic Society. Retrieved May 2, 2018, from http://www.nationalgeographic.org/encyclopedia/climate/

[6] (n.d.). Analytics Community — Analytics Discussions — Big Data Discussion. 6 Easy Steps to Learn Naive Bayes Algorithm (with code in Python) . Retrieved May 2, 2018,
from http://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[7] (n.d.). R-bloggers — R news and tutorials contributed by (750) R bloggers. Understanding Nave Bayes Classifier Using R — R-bloggers. Retrieved May 2, 2018, from http://www.r-bloggers.com/understanding-naive-bayes-classifier-using-r/

[8] (n.d.). TIBCO Statistica — TIBCO Software. Naive Bayes Classifier. Retrieved May 2, 2018, from http://www.statsoft.com/textbook/naive-bayes-classifier

[9] (n.d.). IEEE Xplore Digital Library. IEEE Xplore Full-Text PDF:. Retrieved May 2, 2018, from http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7359273