Data Analytics Project 15
Szeged Weather Report - Humidity Prediction

Tristano Galimberti*
School of Informatics, M.Sc. Artificial Intelligence
Università della Svizzera italiana, Switzerland
Email: *galimt@usi.ch

***Abstract*—In this report I outline the problem of predicting humidity based on the features provided in the dataset of weather conditions in Szeget and the process taken in order to find the most suitable data from which to base such a prediction. The model that achieved the best results was a polynomial regression on the Temperature-Humidity data, with a mean-square error of 0.02 and an R^2 value of 0.47.**

I. Exploration and Description of the Data

The dataset provided is an hourly record of historical weather conditions in Szeget, Hungary, with a time-span of approximately a decade (April 2006 - September 2016). The attributes in the order that they appear in the dataset, with their respective datatypes and units (if applicable) are as follows:

- **Date/Time** (Year-Month-Day hour), minutes/seconds are irrelevant due to frequency of weather records
- **Summary** (27 types of weather), discrete
- **Precipitation Type** (2 types of precipitation), discrete
- **Temperature outside** (Celsius), continuous
- **Apparent Temperature** (Celsius), continuous
- **Humidity** (0.00 - 1.00), continuous
- **Wind Speed** (km/h), continuous
- **Wind Bearing** (degrees), continuous
- **Visibility** (Km), continuous
- **Cloud Cover** N/A
- **Pressure** (Millibars), continuous
- **Daily Summary** (214 unique summaries of weather), discrete.

With elementary knowledge about weather systems, one can have a reasonable idea about which attributes are more likely to have a greater correlatory link to humidity. For example, we know that the higher the temperature gets, the more likely we are to have drier conditions and as a result, a lower humidity. It is also the case that there are multiple interdependent factors that can influence a single variable like humidity, due to the fact that the weather is an open system; for example the statement made before is generally true, however can be severely false for a location that is not near to the seaside or a significantly large body of water, given the correct wind conditions.

The location in question here is in a landlocked country several hundreds of kilometres away from the sea and not near any significantly large bodies of water so we can discount this

Figure 1. Correlation table for all relevant numerical attributes.

from having a considerable effect on the humidity-temperature correlation. The correlation table of the numerical values in Figure 1 affirms this, showing little to no correlation between wind speed/direction and humidity. For this reason I excluded those 2 attributes relating to wind as they were unlikely to provide any meaningful insight into the task at hand.

From this we can observe that both **Temperature** and **Apparent Temperature** have a moderate, negative correlation with humidity, as defined in [1], with values of -0.63 and 0.6 respectively. We also see that there is an almost perfect correlation between **Temperature** and **Apparent Temperature** so that either can be considered. There is also a weak, negative correlation between **Visibility** and **Humidity**, with a value of -0.37. Another attribute that can be removed from the humidity model is air pressure, given that the correlation table 1 shows almost no correlation to humidity (-0.05).

Figure 2 is a table that shows the relevant information taken from the call of `df.describe()` after the pre-processing step outlined in the next section.

Displaying the more relevant quantities from the table in Figure 2 as Histograms reveals some more details of their respective statistics

	Temperature (C)	Humidity	Visibility
count	96453.000000	96453.000000	96453.000000
mean	11.932678	0.735067	10.395826
std	9.551546	0.195157	4.131831
min	-21.822222	0.120000	0.016100
25%	4.688889	0.600000	8.468600
50%	12.000000	0.780000	10.046400
75%	18.838889	0.890000	14.812000
max	39.905556	1.000000	16.100000

Figure 2. Standard Statistics

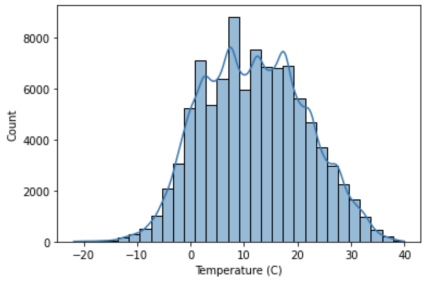

Figure 3. Temperature Histogram

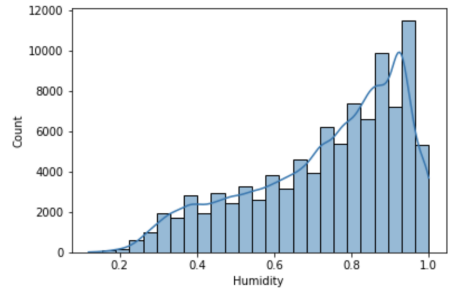

Figure 4. Humidity Histogram

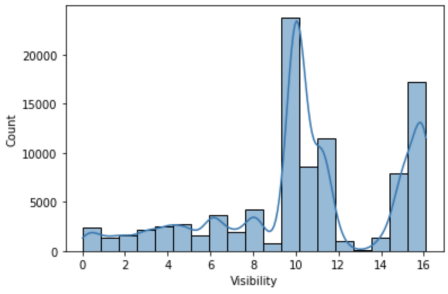

Figure 5. Visibility Histogram

where again, all of these were obtained after the cleaning in the next section.

We can see that the data in the Temperature histogram 3 follow an approximately Gaussian distribution, which is expected given the cyclic nature of the seasons and the resultant changes in temperature. The humidity histogram 4 follows the pattern of a negatively skewed normal distribution with a visual artifact of every other bin having an offset related to the unique numbers of humidity in the dataset and the width of the histogram bins, but it doesn't change the overall distribution curve. Finally, the visibility histogram shows very little uniformity, one reason for this could be that regions in the line of sight of the visbility measurement have certain qualities that could cause fog/mist/traffic fumes to accumulate. This would explain the apparently randomly distributed peaks in the graph, however the dataset didn't provide any details about the particular line of sight used.

I decided to also omit the **Summary** and **Daily Summary** fields in the predictor as there were too many types within each field to sort through (27 and 214 respectively). For another time it could be helpful to group the summaries into distinct categories (e.g. cloudy/clear) to improve the predictive power of the humidity predictor model, although a lot of the conditions can be inferred from the numerical statistics we're already using; for example humidity, visibility, air pressure etc.

This leaves **Temperature** and **Visibility**, which we can plot against Humidity in a scatter plot to obtain a more meaningful representation of each of the attributes with respect to temperature.

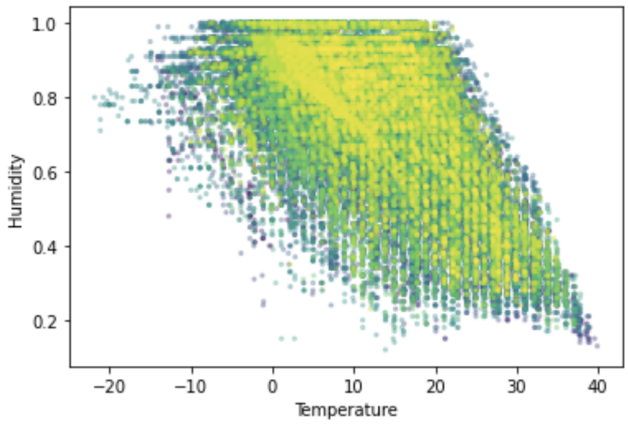

Figure 6. Humidity vs. Temperature with a coloured gradient related to frequency

Given the number of datapoints in figure 6 (approximately 90,000), to make this plot more interpretable, I decided to couple transparency with a colour gradient (generated from a Gaussian distribution of all the points) to more easily delineate an approximation of a best-fit line and also potential outliers. The vertical boxplot 7 suggests a 'range' for such a best-fit line, along with the outliers.

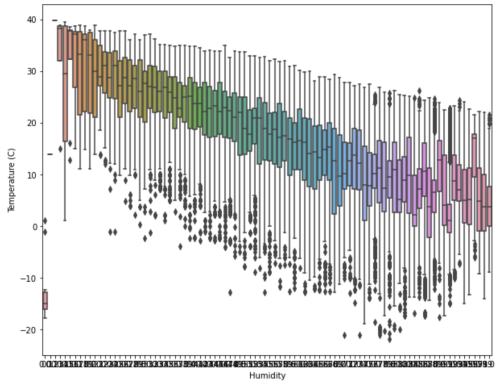

Figure 7. Temperature vs. Humidity box plot

Figure 8 displays the behaviour of visibility with humidity in the same format as Figure 6. It can be observed that visibility is only helpful in enabling us to predict humidity when the visibility is very low as there's a range of roughly 0.15, but for longer visibility distances (10km onwards), there is little to no predictability.

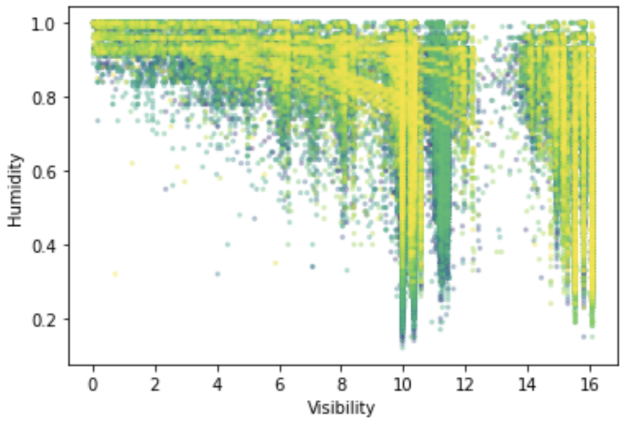

Figure 8. Humidity vs. Visibility (km) with a coloured gradient related to frequency

II. Data Preparation & Pre-processing

Several steps of pre-processing had to be taken prior to determining that the data was to create the prediction model. A good indicator for this was the initial call of `df.describe()` which outlined minimum values of zero which were very unlikely/impossible.

Firstly, for **Humidity** I noticed that the lowest non-zero value was 12%, however there were 22 values which had a humidity of 0. This is physically impossible for atmospheric conditions. Only 22 values out of 96,000 had an invalid (zero) value, so I replaced those values with the mean of the humidity, which was calculated without using those values. The maximum humidity in 2 reads 1.00, which I initially assumed to be another error, however after some research I found there is a meteorogical phenomenon called 'supersaturation' [2] which is common in most parts of the world.

There were also some other suspicious minimum values of exactly zero which are very unlikely. An example of this was a Wind Speed of exactly zero, given that the readings were to 2 decimal places of accuracy.

Another example was the 1288 pressure values of 0, which again, is physically impossible given that this would imply the atmosphere is mimicking a perfect vacuum. This represented 1.5% of the total data which would've been a considerable loss so I adopted the same tactic with replacing the zero values with the mean pressure, considering that at the time I wasn't convinced I would need to use the pressure in the model anyway, given the correlation table 1. I also did the same for the 0 values in the visibility field for the same reason, given that I had already generated the plot in Figure 8 and concluded that it wouldn't be much help in the model for larger values of visibility.

As outlined in the notebook, the attribute **Cloud Cover** was a column full of zeros so it's safe to conclude that this needed to be omitted from the subsequent analysis.

There were also approximately 500 values missing from the **Precipitation Type** field, which I left as they didn't have any influence on the model that I chose to go with.

III. Prediction Model

The prediction model I chose was based on taking a linear, then polynomial regression of the Temperature-Humidity plots, using scikit-learn [3]. I opted for this approach primarily due to the trends between temperature and humidity exhibited in Figures 6 and 7, which were supported by the correlation table.

The results of the linear regression are shown in Figure 9, where we can see that this follows the same trend approximated in the box plot 7 (albeit inverted). As an algebraic function, this reads

$$y = -0.01304842x + 0.8905350, \tag{1}$$

where y is the humidity and x is the temperature.

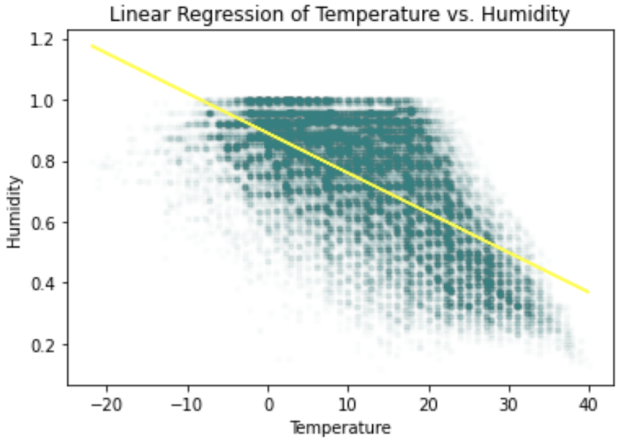

Figure 9. Linear Regression of Temperature vs. Humidity

We can also see from comparing the linear regression result to the underlying scatter plot that we could do a bit better with a polynomial curve, i.e. we could follow the path of the yellow points in Figure 6 more closely. After experimentation with different orders of polynomials, I decided to go for $n = 4$ as there were sharply diminishing returns for higher order curves.

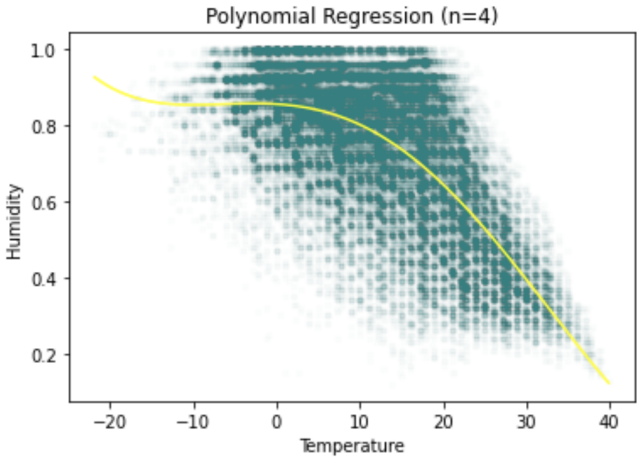

Figure 10. Polynomial Regression of Temperature vs. Humidity

As an algebraic function, this reads

$$y = 2.393 * 10^{-7}x^4 \quad (2)$$
$$-1.2655 * 10^{-5}x^3 \quad (3)$$
$$-3.0324 * 10^{-4}x^2 \quad (4)$$
$$-1.2360 * 10^{-3}x \quad (5)$$
$$+0.856073 \quad (6)$$

This polynomial more closely follows the majority of the points in the plot.

IV. Evaluation and Model Accuracy

When using a train:test split of 70:30, the linear regression equation generated from the training set that resulted in the graph of Figure 9, had the following results when subjected to the test set:

- Mean Square Error: 0.022942
- R^2 value: 0.39586

This was improved upon by the choice of polynomial regression of order 4 as represented in the graph of Figure **??**, that was trained on the same split of data as the linear version. The results for this model when subjected to the test set were:

- Mean Square Error: 0.0202755
- R^2 value: 0.466087

This result is by no means a perfect correlation and I believe it could be further improved upon in the future by incorporating the non-continuous attributes such as the weather summary. By seeing which types of weather correspond to certain ranges in humidity, e.g. if there's a certain weather condition that produces a very specific level of humidity, it could inform us better when combined with the regression polynomial that was generated solely from the temperature-humidity values.

References

[1] "Scatterplots and Correlation Lecture Notes," 2021. [Online]. Available: https://www.westga.edu/academics/research/vrc/assets/docs/scatterplots_and_correlation_notes.pdf
[2] "Glossary of Meteorology," 2012. [Online]. Available: https://glossary.ametsoc.org/wiki/Supersaturation
[3] "Scikit-learn website," 2021. [Online]. Available: https://scikit-learn.org/stable/