Temperature prediction for Szeged

Project Overview:

Szeged is the third largest city of Hungary, a country in Europe.

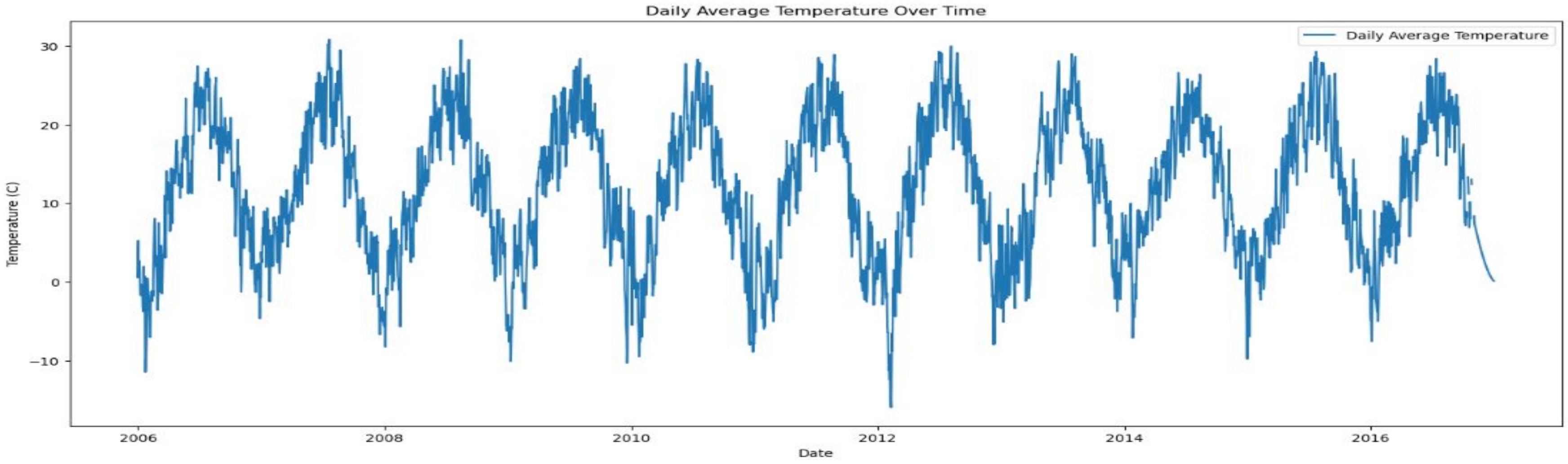
The city played a crucial role in Hungarian history, especially after the Great Flood of 1879, which destroyed much of Szeged. It was rebuilt with a modern city plan, giving it the wide boulevards and elegant architecture that is seen today.

It is known for its Iconic Architecture such as the **Votive Church of Szeged (Szegedi Dóm)**; a neo-Romanesque cathedral that dominates the city skyline, **Reök Palace**; A stunning example of Hungarian Art Nouveau architecture and the **Dóm Square**; where one of the city's most famous festivals **Szeged Open-Air** Festival is hosted. Szeged is often called the *City of Sunshine* due to its high number of sunny days compared to other Hungarian cities.

For this project, we want to investigate what influences the weather in Szeged. We have historical weather data from 2006 to 2016 and we want to find out what factors affect the weather in Szeged.

EDA: Daily Average Temperature over time

The weather in Szeged over the 10 year period (2006-2010) ranges between -15.96°C to 30.81°C. This means that the although the summer season tends to be quite warm, it can get freezingly cold (up to -15.96°C) during the winter season.

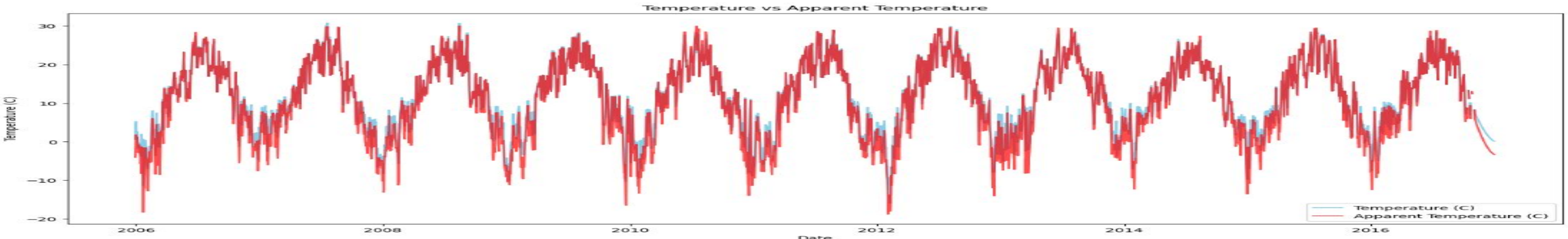


EDA: Temperature versus Apparent Temperature over time

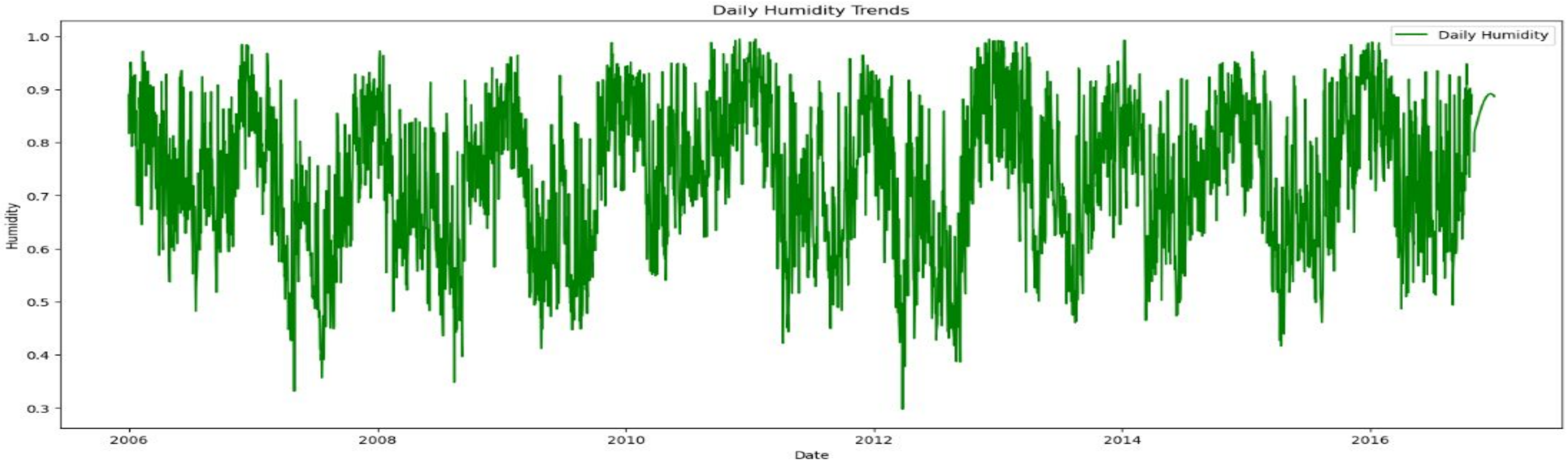
Temperature: is the measure of the warmth or coldness of the air around us, typically measured with a thermometer. It refers to the actual ambient air temperature at a given location and time, regardless of other factors like wind or humidity.

Apparent Temperature: is how the temperature feels to the human body when wind and humidity are factored in. On windy days, the wind can make the air feel cooler than the actual air temperature. Wind removes heat from the body more quickly, which makes it feel colder.

From the plot below, we can see the there is very little difference between the temperature and the apparent temperature. This means that 9 times out of 10, the temperature you feel on your body whilst ouside is the actual temeparature outside when you are in Szeged.

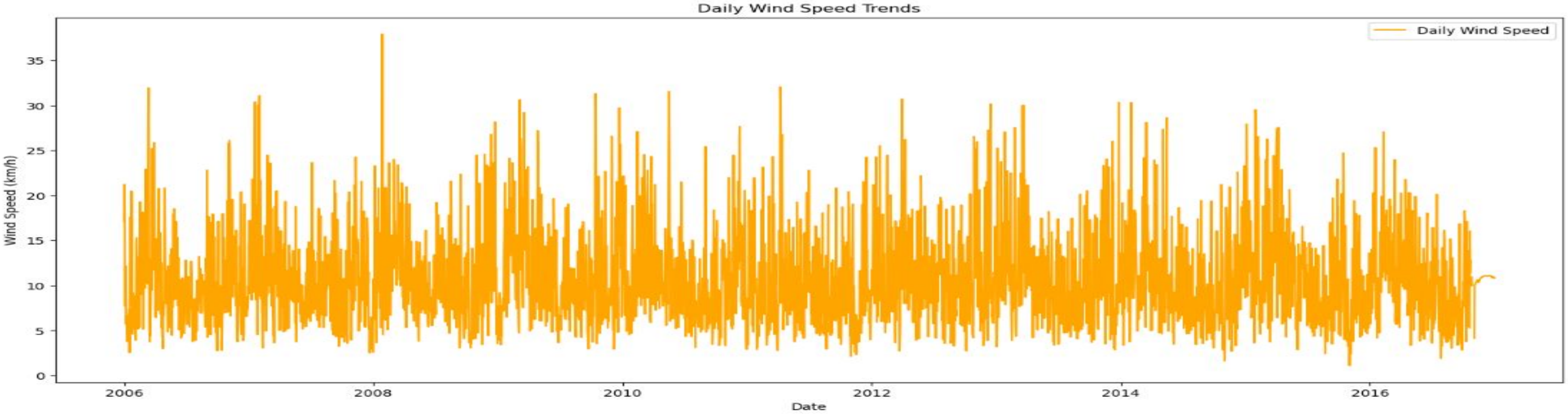


EDA: Daily Humidity Trends



There's a moderate negative correlation, this indicates that as temperature increases humidity tends to decrease.

EDA: Daily Wind Speed Trends



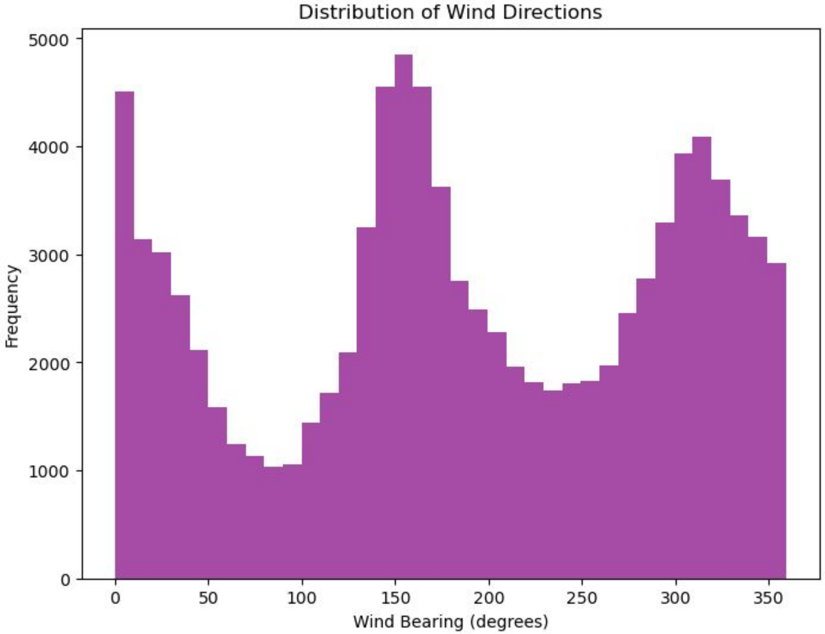
There is very weak negative correlation, there is a slight tendency for higher temperatures to be associated with lower wind speeds, but the relationship is very weak and almost negligible. Wind speed does not play a significant role in explaining variations in temperature in this dataset.

EDA: Distribution of Wind Directions

There are noticeable peaks around 0° (north) and 180° (south), indicating that winds from the north and south are more frequent in Szeged. Similarly, there is another smaller peak around 270° (west

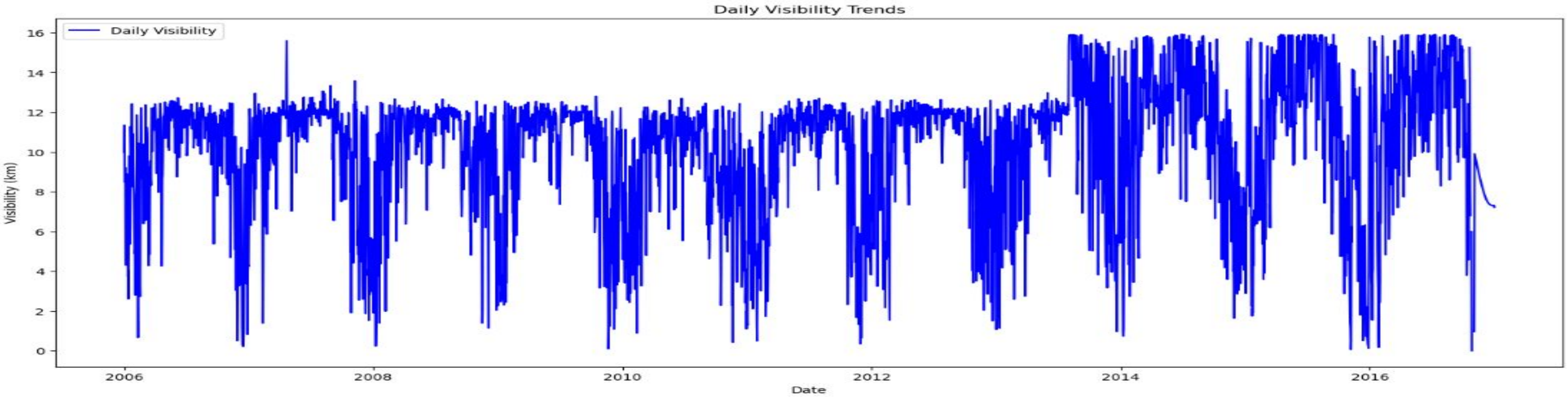
The directions between the main cardinal directions (e.g., 45°, 135° etc.) have relatively lower frequencies.

The histogram appears to show a somewhat symmetric pattern, wh suggests that winds come from opposite directions (north-south an east-west) fairly evenly.



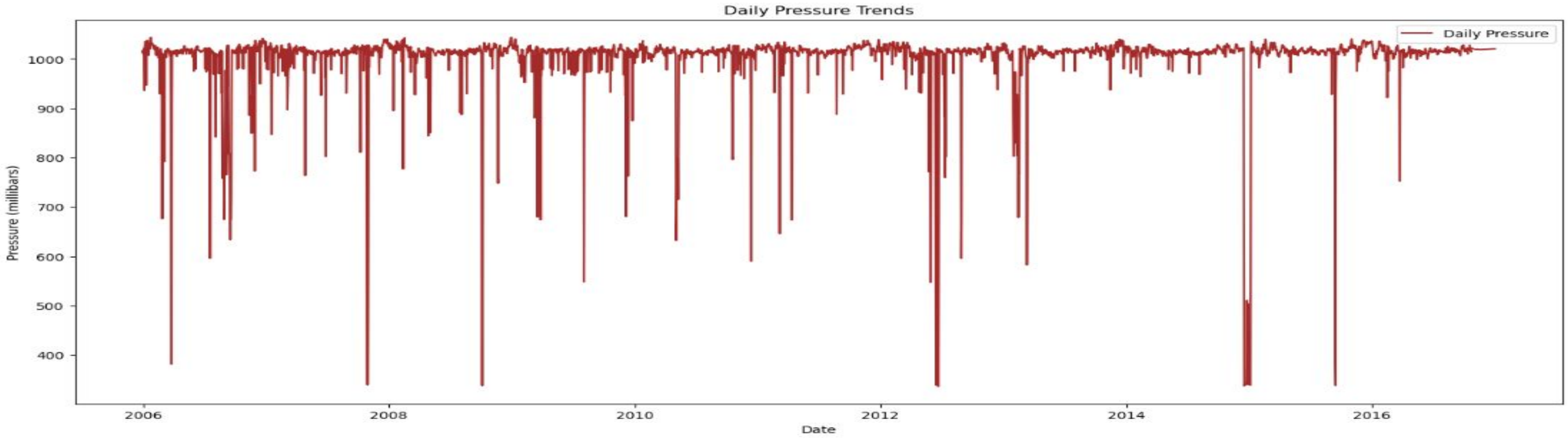
EDA: Daily Visibility Trends

There is a moderate positive correlation, indicating that warmer days are generally clearer, with less haze or obstruction to visibility. As temperature increases, visibility tends to improve.



EDA: Daily Pressure Trends

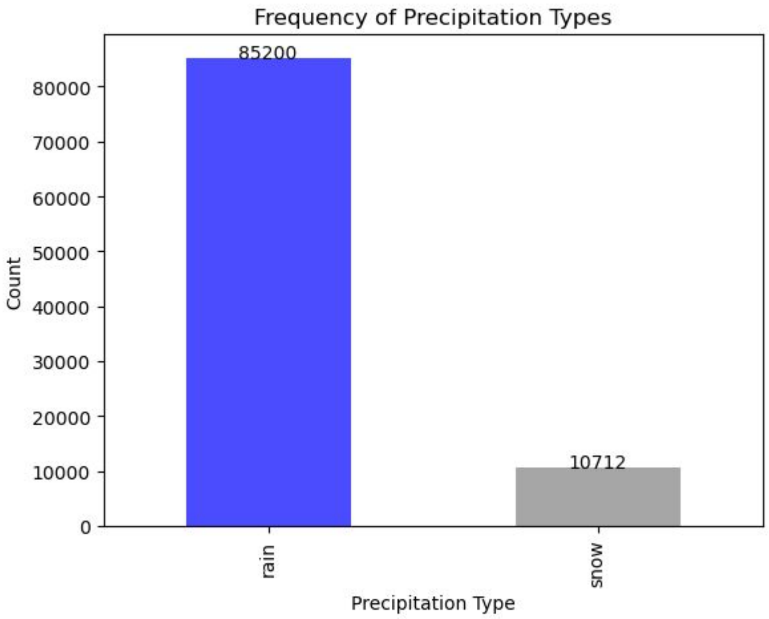
The correlation of 0.00 indicates no linear relationship between temperature and pressure



EDA: Frequency of precipitation types

The high frequency of rain (approximately 89% of precipitation events) indicates that Szeged experiences a climate where rain is the dominant form of precipitation throughout the year.

Snow occurs much less frequently (about 11%), suggesting that Szeged has milder winters compared to regions where snow is more common.



Model Building: Features:

The following features have been selected for their relationship with temperature (correlation) or, in the case of pressure, based off scientific evidence of relationship. The features are the following:

Apparent Temperature

Humidity

Wind Speed (km/h)

Wind Bearing (degrees)

Visibility (km)

Cloud Cover

Pressure (millibars)

Year

Month

Day

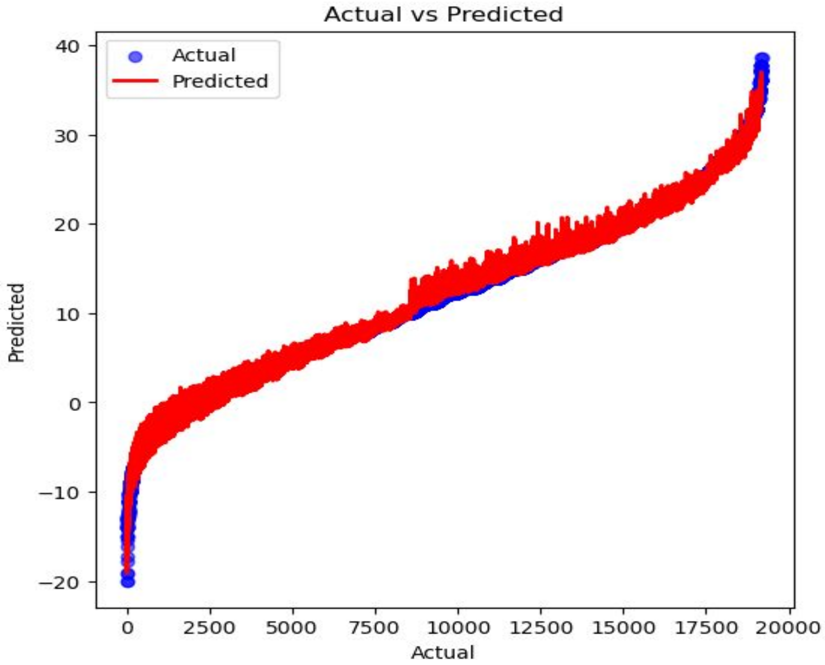
Hour

Model Building: Standardisation:

We are standardising the data using StandardScaler as we believe that standardizing weather data will allow the different variables to have comparable scales, allowing for accurate model training and prediction. The normalisation process improves the model's ability to identify patterns and relationships within the data, leading to more reliable temperature predictions.

Model 1: Ridge Regression:

MSE = 0.8960
RMSE = 0.9466



Model 1 insights:

It has been determined that the best ridge regression model had an alpha of 1.099. Which gave an MSE of 0.8961 and thus having an RMSE of 0.9466 degrees. This means on average the model is off by around 0.95 degrees ie +- 1 degree C which can be argued as relatively accurate.

From the graph, we can see that there are certain areas where the model predicts particularly well. From about -20 to -10, there are quite accurate predictions as well as around the 5-10 degree C mark and also in the high temperatures. The rest seem to have a larger range and don't predict as well.

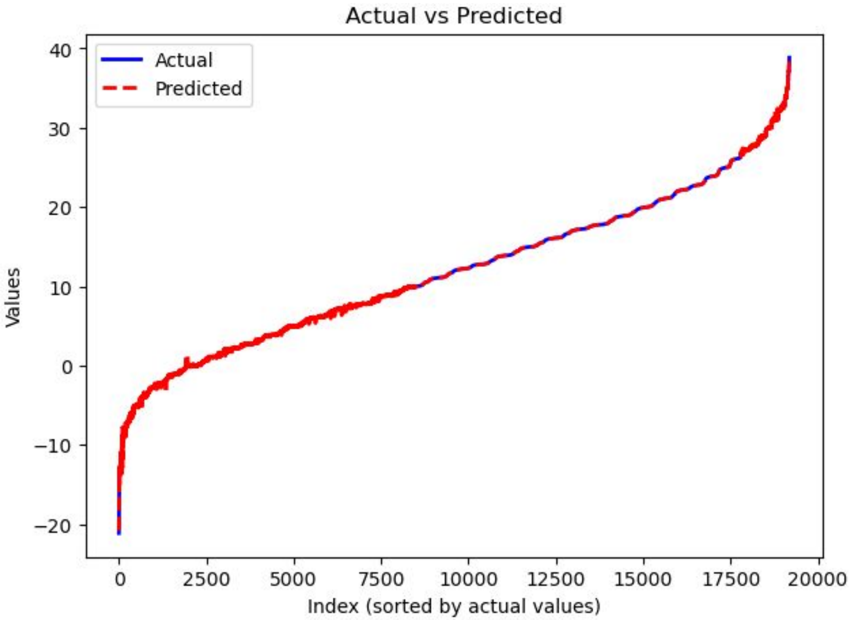
From inspection of the coefficients we see that the apparent temperature has the largest influence on temperature and this is to be expected where a 1 unit increase in the scaled apparent temp leads to on average 9.34 degrees C increase in temperature.

Model 2: Decision Tree:

R2 = 0.9999

MSE = 0.0068

RMSE = 0.0822



Model 2 Insights:

With an R2 of 0.9999, an RMSE of 0.0822 and an MSE of 0.0068 we can say that this model seems highly accurate. On average the model is 0.0822 degrees wrong when it comes to predicting which is highly accurate. Couple that with 99.99% of the variation in temperature being explained by the model it is highly accurate.

From the graph we can see that the actual and predicted values are extremely close thus showing a highly accurate model. Caution should be expressed though because decision trees are known for overfitting and so a comparison between training rmse and test rmse is extremely important.

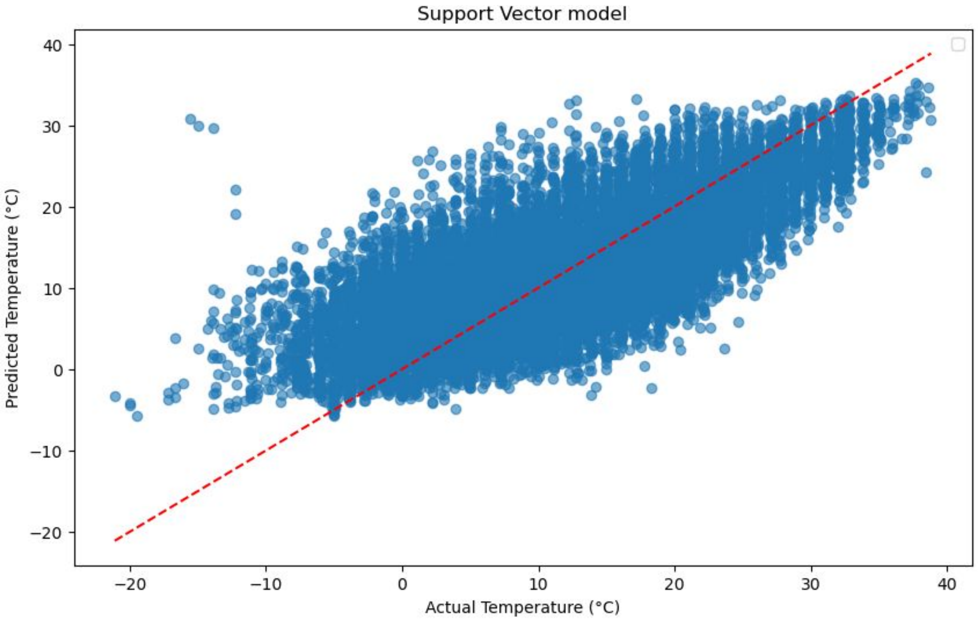
However, these models are known for overfitting. Thus precautions must be taken when using this model to predict.

Model 3: Support Vector Regression:

MSE = 37.4544

RMSE = 6.17

R2 = 0.58



Model 3 Insights:

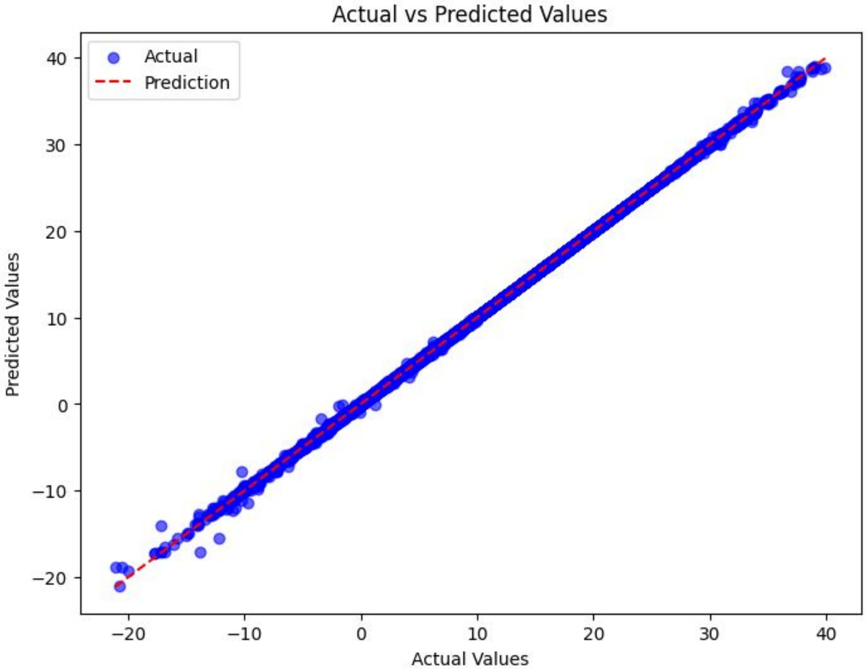
With an R2 of 0.58 and an RMSE of 6.17 it does not seem like this model is an accurate one. An R2 value of 0.58 means that only 58% of the variance in temperature is explained by the model while an RMSE of 6.17 means that on average the prediction is off by 6.17 degrees which is highly inaccurate.

Judging from the SVR graph, it does not look very accurate at all and thus this would not be a great model to make predictions on. There do not seem to be any accurate areas for predictions in this model.

The top 3 features are Visibility, Humidity and Wind Speed in predicting the temperature.

Model 4: Stacking Ensemble:

RMSE = 0.0854
MSE = 0.0073



Model 4 Insights:

This model looks to be highly accurate.

With an RMSE of 0.0854 meaning that on average the predictions from this model are off by 0.0854 degrees. It can be concluded that the ensemble method of the three other models is a highly accurate one.

From the graph we can see that the actual and predicted values are matched almost 1-1. With the most accurate temperature range being between 10 and 25 degrees C (as it is tightly centered around the 45 degree line). The rest while still highly accurate does deviate slightly indicating that there are slight discrepancies for predictions in those temperature ranges.

Conclusion:

Clearly Models 2 and 4 are the best performing both with very low RMSE's and high R2 values.

However because the Decision Tree model is overfitting, to put all eggs in one basket and only use a decision tree to make predictions might not be a clever idea. It is for this reason that we recommend the Stacking Ensemble model. The ensemble model will combine the the best characteristics about all three models (like the SVR's fitting) which reduces the possible risk of overfitting.

This allows the model to not overfit as much but to retain the predicting power of the best models.