

Introduction (Business Problem):

Saudi Arabia plans to harness 9.5 GW from renewable energy resources according to the 2030 vision. Solar PV and wind farms constitute a major part of this initiative. Hence, this huge investment in solar PV and wind farms projects requires an accurate assessment and analysis to select the appropriate sites. The main objective of this study is to answer this question (Where to invest for solar and wind energy?).

Electrical load profile in Saudi Arabia is unique. According to World Energy Council statistics, in 2014, an average household in Saudi Arabia consumed 23.81 MWh of electricity, which is the third highest consumption rate in the world, while the overall average in the world is 3.35 MWh. Saudi Arabia along with other 25 countries account for three-quarters of the global energy demand. Residential and commercial buildings in Saudi Arabia represent 50% of the total load. The electrical load in Saudi Arabia is high in summer season and low in winter season. Peak load in the summer season is two times higher than the peak load in winter season. This unique load profile in the summer season is mainly because of cooling fur, e.g., air conditioners and fans.

Demand of electricity in Saudi Arabia is rising every year where the increase in the number of subscribers has an average rate of 5.2%. Furthermore, the demand in the industrial sector has an annual growth of 6.9%. This increasing demand has resulted in more combustion of fossil fuels, e.g., oil and natural gas, which increased the amount of carbon dioxide emissions in the atmosphere. In 2013, due to combustion of fossil fuels, Saudi Arabia had produced 458.8 million tons of carbon dioxide emissions into the atmosphere compared to 429.8 million tons in 2012. The annual growth of carbon dioxide emissions between 1971 and 2013 was 5.8%. The annual rate of increase in electricity demand was 7.5% in the last decade. The increasing demand is one of the major problems being faced by power companies in Saudi Arabia. So, if renewable energy systems are not adopted, the demand of fossil fuels for electric power generation is expected to increase from 3.4 mb/d in 2010 to 7.3 mb/d in 2028. Consequently, the country's export revenue from fossil fuels will encounter significant reduction.

Hence, to overcome energy crisis and long-term environmental impact in future, Saudi Arabia has plans to diversify its power generation by including renewable energies. Along with massive oil reserves, Saudi Arabia has other resources like solar energy that could contribute to solve the future crisis of energy. Saudi Arabia has an annual solar irradiance energy that is around 2000-2450 kWh/m² and vast empty space in the deserts. Germany, which has the second highest installed solar PV capacity in the world, has maximum solar irradiation of 1200 kWh/m² which is less than the minimum solar irradiation in Saudi Arabia, i.e., 2000 kWh/m². Hence, this makes Saudi Arabia an ideal place for solar photovoltaic and concentrated solar power generation.

Data Review:

The dataset contains 3-year of historical weather data for 13 cities in Saudi Arabia for 2017, 2018 and 2019. The historical weather data include temperature, wind speed, pressure, visibility, and relative humidity.

The data is mostly numerical, but it has one categorical column which is about the weather status, e.g., sunny and cloudy.

This dataset was selected for our project because solar and wind energy systems depend mainly on weather conditions. So, wind energy systems depend mainly on wind speed data, and solar PV systems

depend mainly on solar irradiance data. Hence, considering weather is an important factor for doing a feasibility study for certain locations to install a renewable energy system, i.e., solar or wind,.

For the data source, the data was retrieved from Kaggle.

Data Preprocessing:

The dataset that is used does not contain null values, however, for some cities, there are some missing rows, i.e., entire rows for some days are missing. Those cities that contain so many missing values such as Assir were removed completely from the dataset.

Furthermore, we needed to clean the weather column to reduce the number of categories in that column by assigning the rare categories to the common categories, i.e., 0.001% Sunny with some clouds = Sunny.

For training the machine learning models, we have tried two different approaches:

- The first approach is using a sliding window technique to convert the time series data into supervised learning format by introducing lagged columns of temperature and wind speed. So, basically, the input was the previous values of temperature and wind speed, and the output was the future value of temperature and wind speed. For example, Table 1 shows a sequence of measurements of wind speed and temperature, and Table 2 shows how the data is transformed into a supervised learning format using a sliding window technique.

| t | Temperature | Wind |
|---|-------------|------|
| 1 | 26 | 15 |
| 2 | 28 | 12 |
| 3 | 29 | 10 |
| 4 | 27 | 19 |
| 5 | 25 | 11 |
| 6 | 23 | 0 |

Table 1: Sequence of Temperature and Wind Speed Measurements.

| Input Features | | | | | | Targets | |
|----------------|-----------|-----------|-----------|-----------|-----------|---------|---------|
| Temp(t-3) | Temp(t-2) | Temp(t-1) | Wind(t-3) | Wind(t-2) | Wind(t-1) | Temp(t) | Wind(t) |
| 26 | 28 | 29 | 15 | 12 | 10 | 27 | 19 |
| 28 | 29 | 27 | 12 | 10 | 19 | 25 | 11 |
| 29 | 27 | 25 | 10 | 19 | 11 | 23 | 0 |

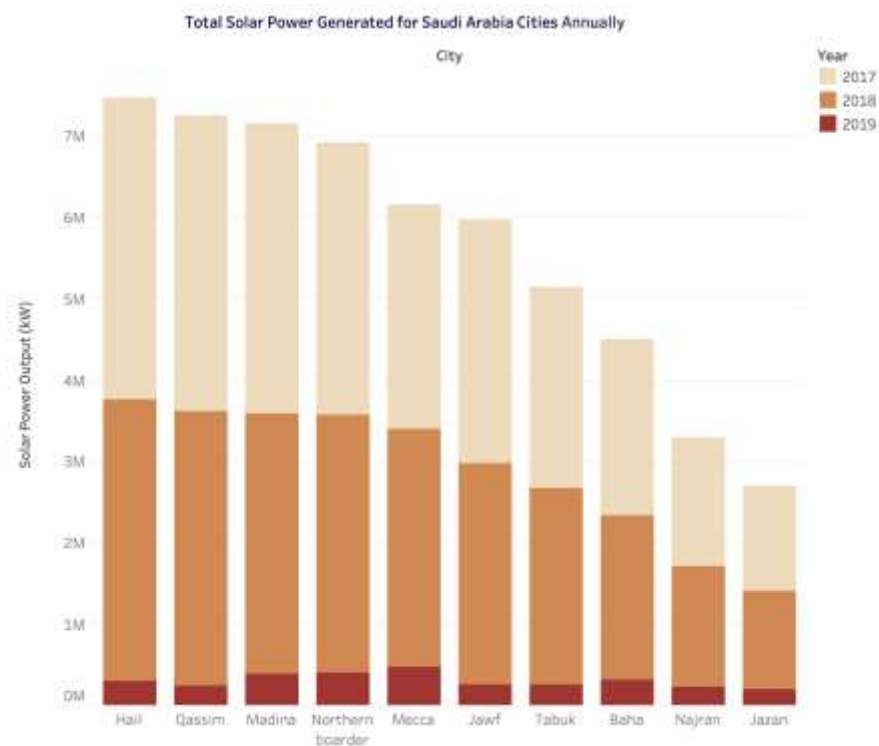
Table 2: The time series sequence after converting it into supervised learning format.

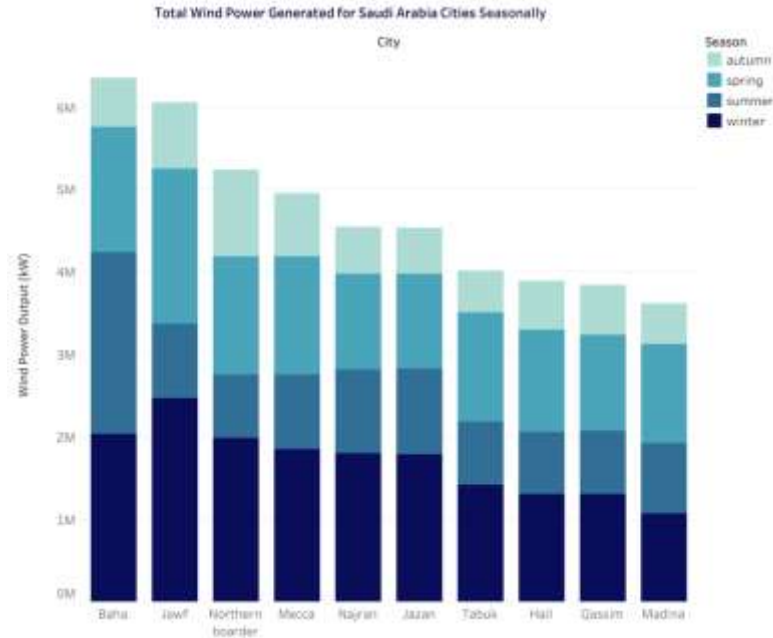
- The second approach is by organizing the data such that the target is temperature, and the input features are other columns, e.g., pressure and wind speed. We repeated this approach to predict the wind column as well, so, in this case, the target is wind speed and the input features are temperature, pressure, relative humidity and weather.

Furthermore, we have done power calculations for solar and wind energies for all cities. We used mathematical formulas to calculate the energy generated by wind turbines using wind speed, temperature, and pressure data. Also, we used solar irradiance and temperature data to calculate the energy generated by solar panels.

Data Visualization:

The following plots ranks 13 Saudi Arabia cities/areas to their renewable energy power generation based on calculation of a various factors over the period between 2017 to 2019. From the statics data Hail considered as one of the fastest-growing Solar energy for 7M kW. While in Baha wind farms harness the energy are more efficient, and stronger for 6M kW.





Model:

This prediction study is a regression problem. There are many machine learning models for regression tasks such as Linear Regression (LR), Artificial Neural Networks (ANNs), Support Vector Regression (SVR), Random Forest (RF) regressor, XGBoost regressor, K-Nearest-Neighbor (KNN) regressor, and many others.

For the first approach, Random Forest (RF) regressor, K-Nearest-Neighbor (KNN) regressor, and Support Vector Regression (SVR) models were considered to do the one hour ahead forecasting task.

For the second approach, Random Forest (RF) regressor, XGBoost regressor, K-Nearest-Neighbor (KNN) regressor, and Support Vector Regression (SVR) models were considered to do the prediction task.

Results:

For the first approach, three models were considered with different number of step size and number of lags for both wind and temperature. Figure 1 shows the results for forecasting one hour ahead of the temperature, and Figure 2 shows the results for forecasting one hour ahead of the wind. In Figure 3, results as the best model for both wind and temperature, Support Vector Regression, shows. From the results of all models for both wind and temperature, it can be seen that step size 1 has the best results. Furthermore, since the improvement of increasing the number of lags is very small, then, three lags could be the best compromise between model complexity and accuracy because lowering the number of features reduces the model complexity.

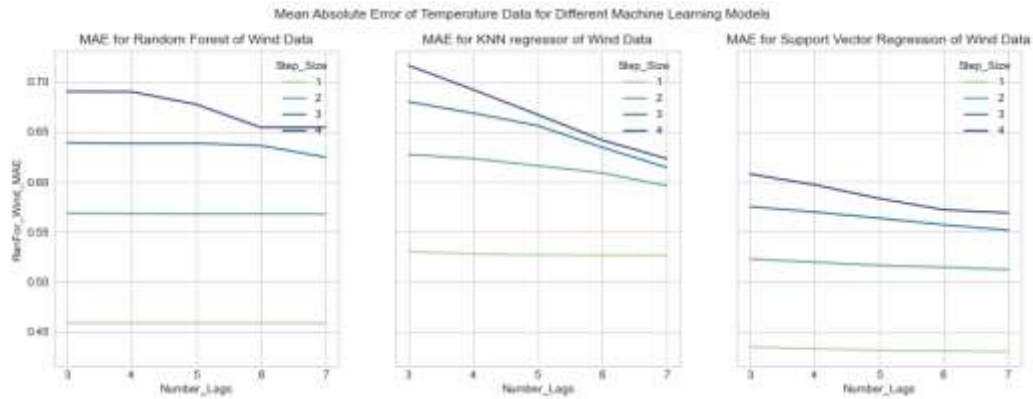


Figure 1: Mean Absolute Error of Hourly Prediction of Temperature Data Using Three Machine Learning Models.

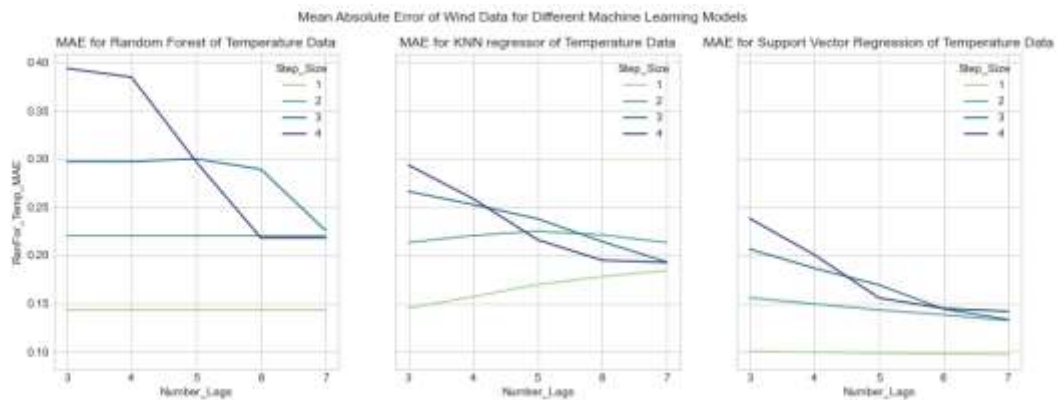


Figure 2: Mean Absolute Error of Hourly Prediction of Wind Data Using Three Machine Learning Models.

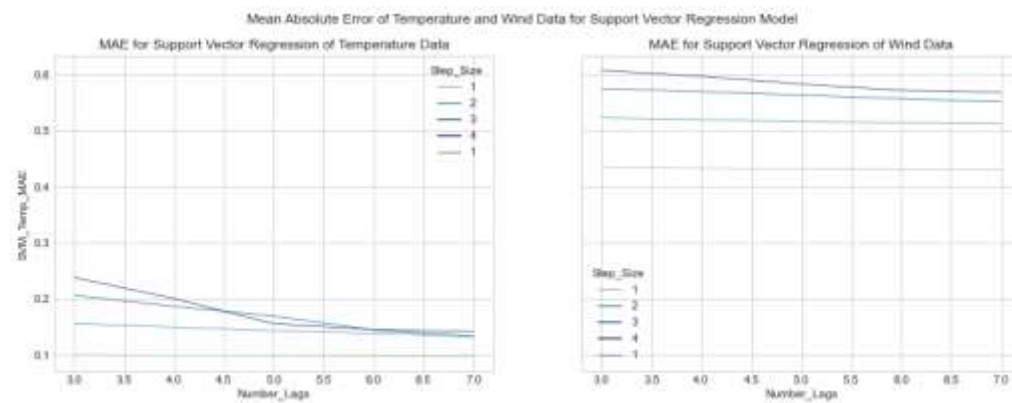


Figure 3: Results of Support Vector Regression (SVR) for Temperature and Wind Data.

| Model | MAE Temperature | MAE Wind |
|-------|-----------------|----------|
|-------|-----------------|----------|

| | | |
|--------------------|-------------|-------------|
| Random Forest | 0.099001048 | 0.484019854 |
| KNN | 0.833380913 | 0.864868315 |
| XGBoost | 0.135522777 | 0.578913635 |
| XGBoost GridSearch | 0.088125963 | 0.501760723 |
| SVM | 0.8224502 | 0.787255783 |

Table 3 : The Performance Results for Models of Predicting Temperature and Wind.

For the second approach, first we split the data into training (80%) and testing (20%). For the data preparation part, we found some columns which have a unique category value that is assigned an integer value, so we applied an Ordinal Encoder for 'season', 'year', 'month', 'day' and 'hour' columns. We used Standard Scaler for 'temp', 'wind', 'humidity', 'barometer' and 'visibility' columns. Also One-Hot Encoding was applied for 'city' and 'weather_category' columns. Then we predicted Temperature and Wind by using four models: Random Forest, KNN, XGBoost and SVM.

From table 3 above, the best model for predicting the Temperature column is XGBoost after using Grid Search to tune its hyperparameters. XGBoost showed the lowest Mean Absolute Error (MAE), which is the metric used to measure model accuracy. The best model for predicting the Wind column is Random Forest as it shows the lowest Mean Absolute Error (MAE).

Conclusion and Future Work:

This study has answered the question “Where to invest?” for both solar and wind energy systems in Saudi Arabia based on weather conditions in each city. From the results in Figure X, it is obvious that Hail is the best place to invest for solar energy in Saudi Arabia because it has the highest total of solar energy generated annually. Also, for wind energy, Jawf is the best city for investment because it has the highest total energy among other cities.

For the machine learning part using the sliding window approach, it is found that using step size of 1 has the best results. Also, using more number of lags, i.e., historical data, improves the model accuracy. However, increasing the number of lags, or number of features, increases model complexity, and this increase in model complexity has insignificant improvement in the accuracy. Hence, the number of lags that will provide us with the best compromise between model accuracy and complexity is three.

In the first approach, the models which were used are Random Forest (RF) regressor, K-Nearest-Neighbor (KNN) regressor, and Support Vector Regression (SVR). Results show that Support Vector Regression (SVR) is superior in terms of model accuracy. Meanwhile, Random Forest (RF) and K-Nearest-Neighbor (KNN) regressors had lower training time with a reasonable accuracy.

For the machine learning part using the second approach, the models which were used are Random Forest (RF) regressor, K-Nearest-Neighbor (KNN) regressor, Extreme Gradient Boosting (XGBoost) regressor and Support Vector Machine (SVM) regressor. Results show that Extreme Gradient Boosting (XGBoost) and Random Forest (RF) regressors have the lowest Mean Absolute Error.

Future Work:

In future, it will be interesting to conduct a return-on investment study to find the payback period of the investment in each city for both solar and wind energy systems.

For the machine learning part, it will be interesting in future to try more machine learning models. There are excellent deep learning models for time series forecasting that are worth trying such as Long Short-Term Memory (LSTM) Recurrent Neural Network and Gated Recurrent Unit (GRU) Neural Network. Furthermore, it will be interesting to try some statistical models and compare their performance with the machine learning models. These statistical models include Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA).

Here is the link for our Medium blog:

<https://medium.com/@halaalmodarra/forecasting-weather-in-saudi-arabia-for-renewable-energy-applications-ac99443f9757>