UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Diplomado Técnicas Estadísticas y Minería de Datos

Proyecto Final módulo IV

PREDICCIÓN DE RADIACIÓN DNI

Autor

Andrés Moreno Moreno

Febrero 2022

Análisis de Negocio

De acuerdo con el CSM la radiación es la emisión, propagación y transferencia de energía en cualquier medio en forma de ondas electromagnéticas o partículas, en al caso de radiación solar se analizan 3 componentes para evaluar el nivel de radiación que llega a un punto de la tierra, estos componentes son:

- Radiación normal directa (DNI).
- Radiación horizontal difusa (DHI).
- Radiación horizontal global (GHI).

Analizar estos componentes resulta sustancial para la elaboración de proyectos de plantas solares de concentración de potencia CSP o algún otro proyecto relacionado con obtención de energía solar.

Si bien los 3 componentes de radiación son importantes, el presente proyecto estudia únicamente la radiación DNI para desarrollar un modelo de regresión.

Para realizar el trabajo los datos fueron extraídos de la base de datos nacional de radiación solar NSRDB para el año 2015 a una latitud y longitud de 46.57° y -104.78°, las variables extraídas para el estudio fueron las siguientes:

Variable	Descripción	Tipo de Variable
DNI	Radiación solar directa medida en watts por metro cuadrado	De razón.
Month	Mes de la medición	Nominal
Cloud Type	Tipo de nube de acuerdo a PATMOS-X	Nominal
Dew Point	Punto de Rocío medido en °C	De Intervalo
Solar Zenith Angle	Ángulo de elevación solar medido en grados	De Intervalo
Wind Speed	Velocidad del viento medida en metros por segundo	De razón.
Relative Humidity	Relación entre la presión parcial del vapor de agua y la presión de vapor de equilibrio del agua medida en °C	De Intervalo
Temperature	Temperatura medida en °C	De Intervalo
Pressure	Presión medida en milibares	De razón

Cuadro 1: Descripción de Variables

Los valores medidos con cero o cercanos a cero en DNI y "Wind Speed" fueron determinados así por poseer lecturas de medida muy bajas, por ello no representan valores erróneos o nulos y se les da una escala de razón.

Entendimiento de Negocio

1) ETL

Los datos fueron recibidos por correo en un archivo CSV, se eliminaron las dos primeras columnas porque contenían detalles sobre los instrumentos y especificaciones de donde se realizó la medición (posicionamiento, elevación, fecha, hora, etc.). El archivo fue procesado con Python por medio de la librería de Pandas, obteniendo un conjunto de datos con 17520 registros y 9 atributos. Los datos y el código usado pueden encontrarse en: https://github.com/RobertDalton/Proyecto_DNI.git

2) Análisis de Datos

a) Análisis General

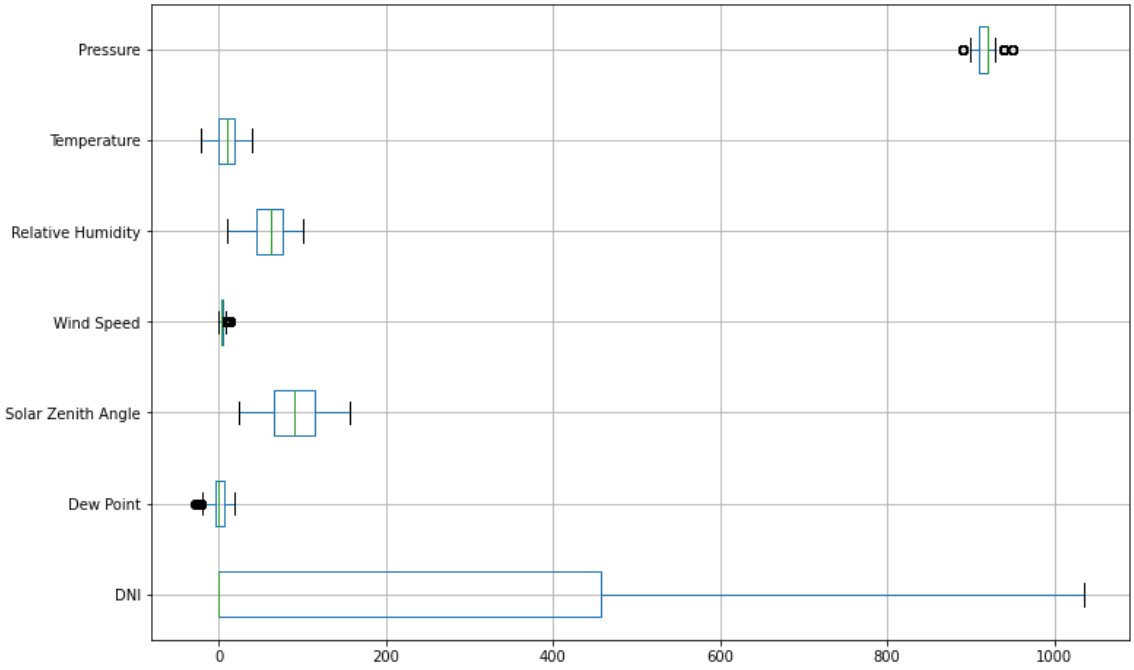
	DNI	Dew Point	Solar Zenith Angle	Wind Speed	Relative Humidity	Temperature	Pressure
count	17520.000000	17520.000000	17520.000000	17520.000000	17520.000000	17520.000000	17520.000000
mean	225.353311	0.529053	89.664271	3.443927	60.729293	9.298459	918.647260
std	327.743833	8.649458	32.763676	1.885001	21.429893	11.804974	6.942311
min	0.000000	-28.000000	23.130000	0.100000	9.830000	-22.000000	890.000000
25%	0.000000	-4.000000	64.960000	2.000000	44.920000	0.000000	910.000000
50%	0.000000	0.000000	89.590000	3.200000	61.490000	9.000000	920.000000
75%	458.000000	7.000000	114.172500	4.600000	76.892500	18.000000	920.000000
max	1036.000000	18.000000	156.860000	13.300000	100.000000	40.000000	950.000000

	Month	Cloud Type
count	17520	17520
unique	12	10
top	1	0
freq	1488	6109

Comentarios:

- Los datos parecen no contener valores nulos debido a que la cantidad de valores en cada atributo es la misma (17520 registros).
- Las medias son muy diferentes entre las variables numéricas de razón sobre todo en la variable DNI, esto sugiere realizar un escalamiento.
- La desviación estándar para la variable “Dew Point” es muy alta respecto a su media, esto sugiere valores outliers y una dispersión muy alta en esta variable.
- La mitad de los valores en la variable DNI es cero, esto nos indica que durante las mediciones de radiación muy pocas veces se contó con cielo despejado.
- Observamos que tenemos 10 tipos de nubes y mediciones para cada mes.

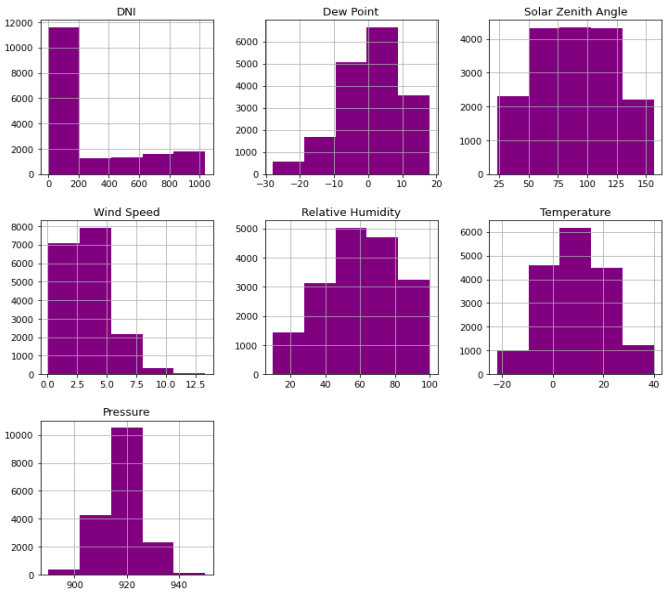
b) Análisis de Outliers



Comentarios:

- Se observan outliers para las variables “Pressure”, “Wind Speed” y “Dew point”, si alguna de estas variables tiene una asociación alta con la variable DNI, se deberán remover los valores atípicos.
- Se confirma la necesidad de escalar los datos para reducir la dispersión de los datos.

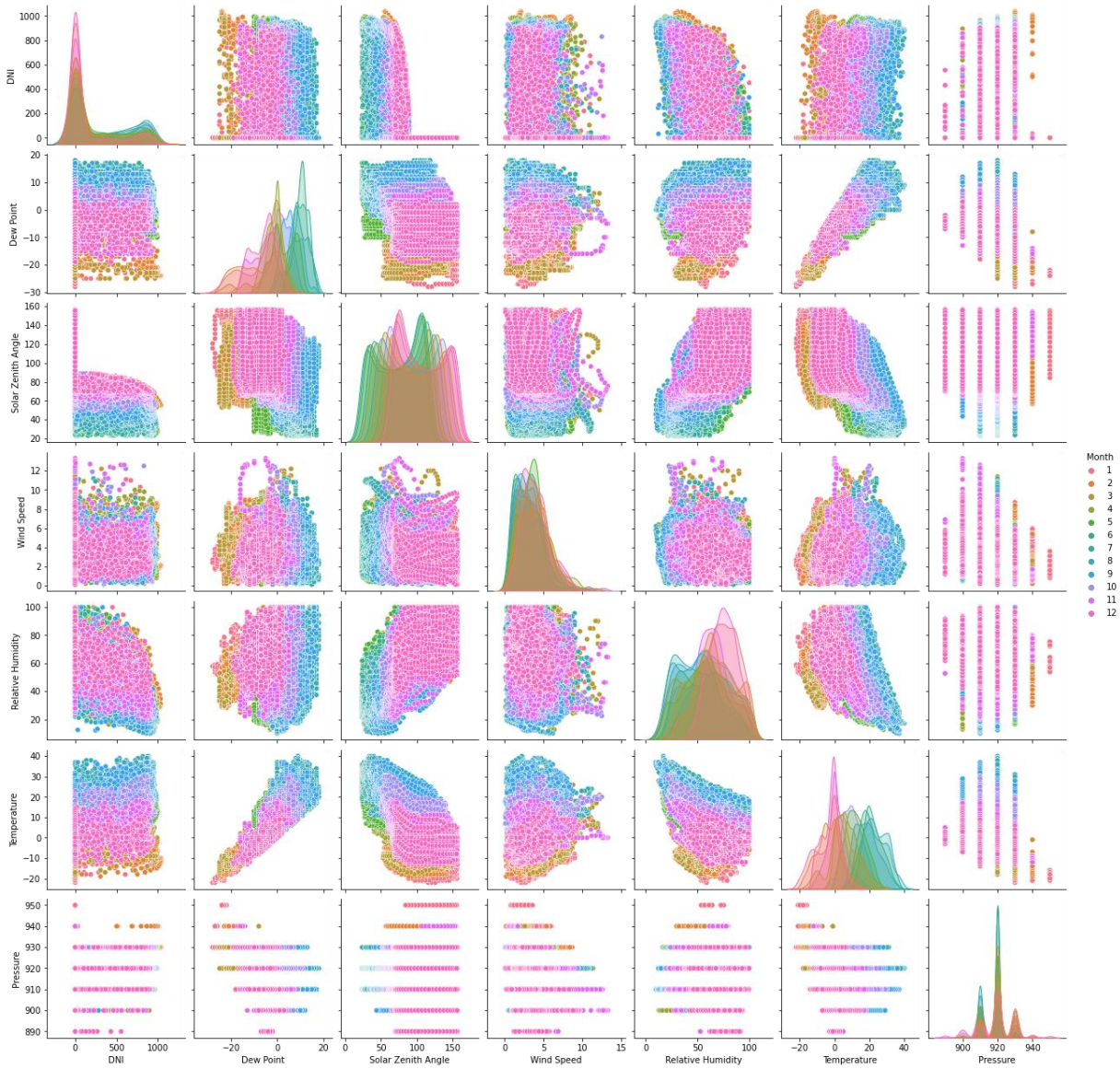
c) Análisis de Simetría



Comentarios:

- Sólo las variables “temperatura”,” solar zenith angle” y “pressure” parecen seguir una distribución normal.
- El resto de las variables tienen sesgo a la izquierda o a la derecha.
- La variable DNI y Wind Speed parecen ajustar bien a una distribución exponencial

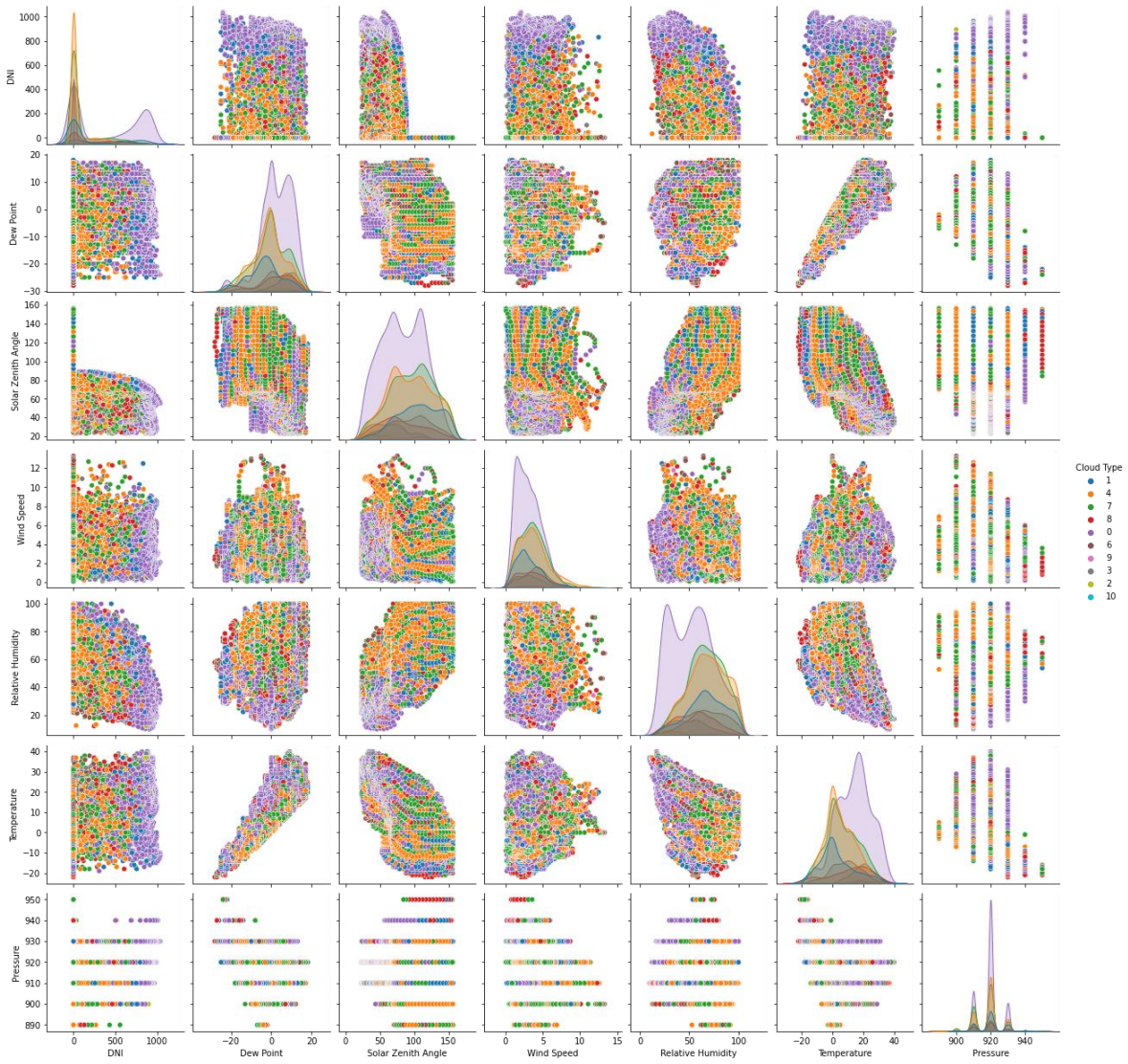
d) Distribución por mes



Comentarios:

- Ninguna de las demás variables parece tener una asociación lineal con la variable DNI, incluso aparentan tener correlación lineal nula dada la forma de los puntos.
- Se puede observar que salvo por la variable “pressure”, los colores de los meses parecen agrupar las mediciones, esto podría sugerir la aplicación de un modelo de regresión por mes o grupos de meses.
- La variable “Dew Point” muestra una relación lineal con “temperatura”, esto sugiere retirar alguna de las dos.

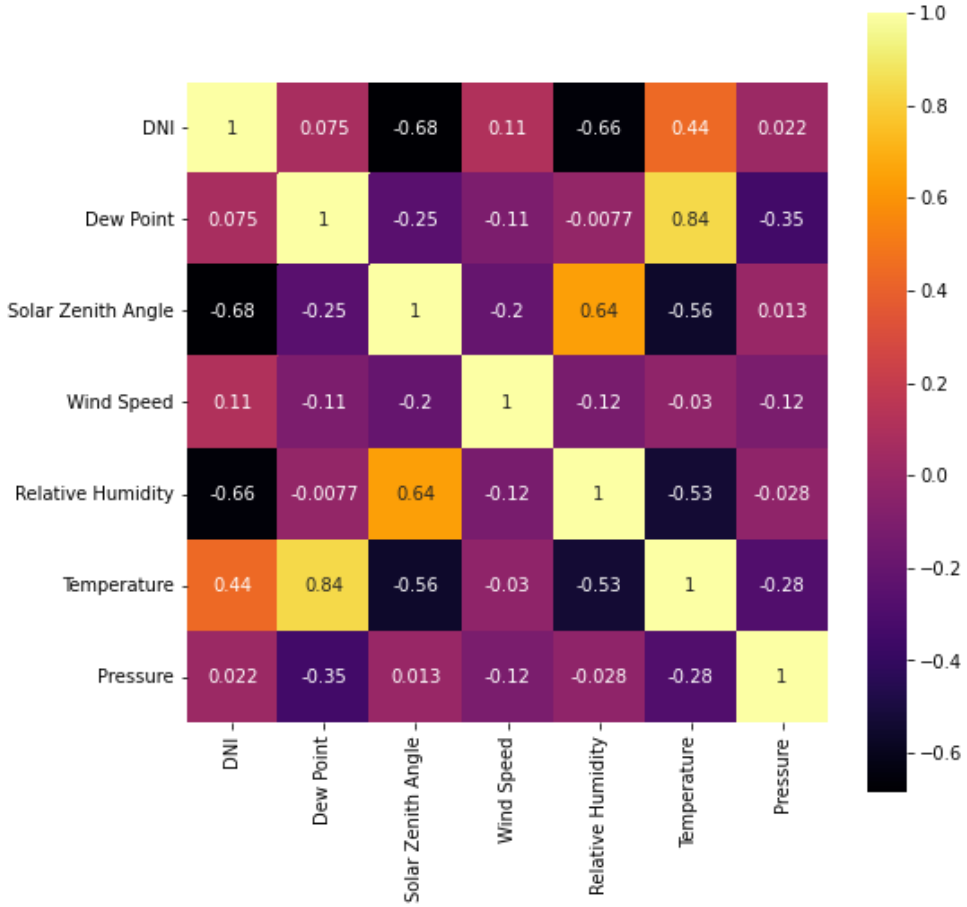
e) Distribución por Tipo de Nube



Comentarios:

- La agrupación es menos clara que al agrupar por mes.
- La nube de tipo 6 (Morada) parece estar relacionada con niveles de radiación más altos, dado que los registros con DNI más alto se presentaron con ese tipo de nube. Esto indica que la radiación atraviesa con mayor facilidad este tipo de nube.

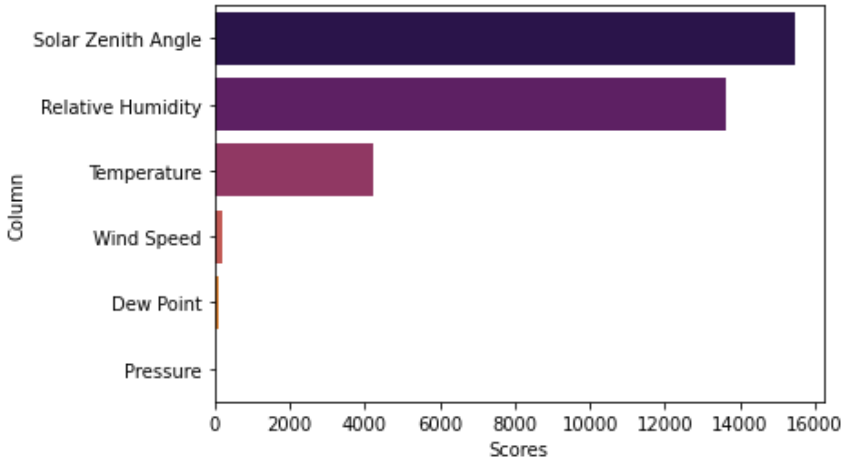
f) Correlación y Mejores Predictores



Comentarios:

Como se observo en la dispersión de los datos, casi todas las variables tienen un nivel bajo de asociación con el DNI, por suerte encontramos correlaciones negativas relativamente altas con “Zolar Zenith Angle” y “Relative Humidity”, además hay una ligera correlación positiva con “temperature”.

Realizando un feature selection para evaluar el poder predictivo se confirma que las variables que deben incluirse en el modelo son: “Solar Zenith Angle”, “Relative Humidity” y “Temperature”.



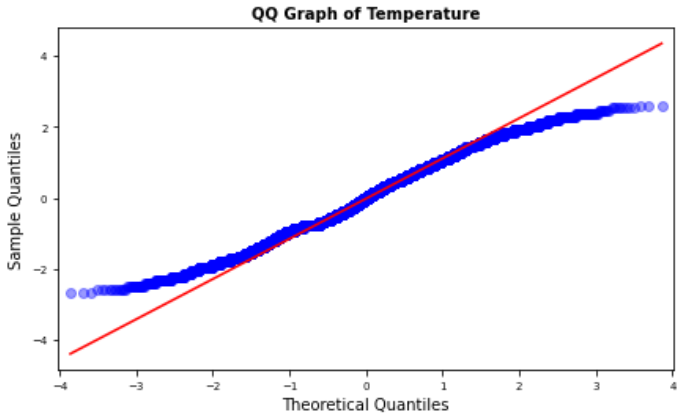
Estás variables no presentaban outliers, por lo que podemos proceder sin remover registros.

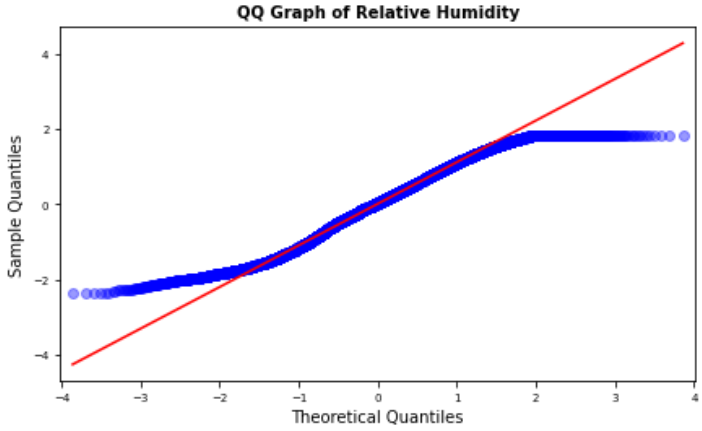
3) Limpieza y Transformación.

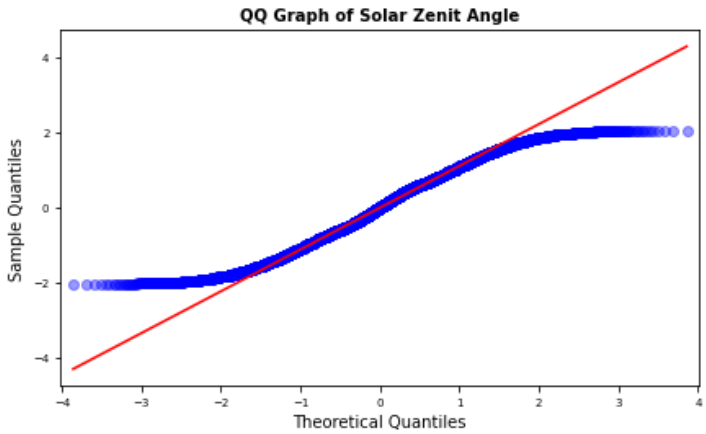
Como se observó en el análisis de datos, no se encontraron valores nulos.

```
DNI                  0
Month                0
Cloud Type           0
Dew Point            0
Solar Zenith Angle   0
Wind Speed           0
Relative Humidity    0
Temperature          0
Pressure             0
dtype: int64
```

Se aplico el escalamiento StandardScaler y la transformación de Yeo-Jhonson para conseguir normalidad, dado que al aplicar la prueba de Normalidad se obtuvieron los siguientes resultados, indicando una transformación necesaria:







Esto se realizó con un pipeline de preprocesamiento.

```
pipeline = Pipeline([
      ("Scaler", StandardScaler()),
      ("Transformation", PowerTransformer(method='yeo-johnson'))
])
```

Notas:

- En un inicio se intento la transformación de MinMax escaler dando resultados de menor rendimiento, por ello se opto por la transformación StandardScaler.
- A pesar de que la variable DNI presentaba un sesgo positivo muy notable, al transformar la variable se obtenían resultados de menor desempeño en las predicciones, por ello no se le aplico la transformación.

Modelado Analítico

1) Algoritmo a Aplicar

En un inicio se tenía la intención de aplicar solo los modelos de regresión lineal y regresión lineal Ridge, pero al desarrollar estos modelos el desempeño fue muy bajo y por ello se optó por implementar también una red neuronal de regresión.

2) Datos de Entrenamiento y Prueba

Los datos fueron distribuidos con 80% de entrenamiento y 20% de prueba, con 14016 registros de entrenamiento y 3504 de prueba.

3) Resultados del Modelo

- **Regresión Lineal**

```
Mean squared error (MSE): 49025.86
Root Mean squared error (RMSE): 221.42
Coefficient of determination (R^2): 0.55
```

- **Regresión Lineal Ridge**

```
Mean squared error (MSE): 49025.86
Root Mean squared error (RMSE): 221.42
Coefficient of determination (R^2): 0.55
```

- **Red Neuronal de Regresión**

```
Mean squared error (MSE): 31821.85
Root Mean squared error (RMSE): 178.39
Coefficient of determination (R^2): 0.71
```

El modelo de regresión lineal tiene un porcentaje de efectividad muy bajo, sólo un poco mayor a el lanzamiento de un volado, este resultado era esperado debido a que las correlaciones no eran muy altas. Aún así se intentó una regresión Ridge con diferentes valores de alfa para intentar disminuir el error del modelo, pero los resultados no fueron diferentes. En conclusión, los modelos de regresión lineal explican el 55% de la variabilidad de DNI y tienen un error de 221.42 watts por metro cuadrado.

Por otro lado, la red neuronal es notablemente superior, ya que se explica un 71% de la variabilidad de DNI y tiene un error de 178.39 watts por metro cuadrado.

Aunque lo ideal sería tener entre 80% y 85% en el coeficiente de determinación, se considera que el resultado de la red neuronal es aceptable debido a la naturaleza de la variable a predecir.

Aplicación del Modelo

1) Beneficios

Los resultados obtenidos pueden ser usados para obtener estimaciones de radiación DNI y como una base para la mejora futura del modelo, además se ha descubierto que el tipo de nube 6 y algunos meses del año también influyen para obtener mayor o menor radiación DNI.

Por otro lado, se observa que la variable DNI parece seguir una distribución exponencial, por lo que este problema de predicción podría abordarse con un modelo probabilístico de tipo exponencial.

2) Limitaciones

En el momento que se consultó la base de datos solo estaban disponibles para su descarga los datos para años anteriores o iguales a el 2015, un estudio para un año más reciente hubiera sido mejor.

Por otra parte, el desarrollador del presente proyecto no es experto en radiación solar ni en energías renovables, un investigador con una mejor formación en el área pudo haber determinado mejores coordenadas de latitud y longitud para extraer los datos, así como mejores interpretaciones, sobre todo en los tipos de nube dado que la clasificación de PATMOS-X no es tan clara.

3) Supuestos y Conclusiones

Los supuestos más importantes en el desarrollo del proyecto son la clasificación de las variables DNI y "Wind Speed" de tipo "razón", es claro que el 0 en estás escalas no implica ausencia de radiación o de viento, pero dado que el objetivo del proyecto era realizar predicciones se consideraron de esta manera para tener un indicador de valores muy bajos de radiación y viento.

Para finalizar, se obtuvo un modelo aceptable con amplias mejoras a futuro considerando: la distribución de la variable DNI, otras técnicas de regresión y añadiendo más datos.