ESCUELA POLITECNICA NACIONAL
DATA MINING Y MACHINE LEARNING

Proyecto 2 1er Bimestre

Alexis Vera

Introducción:

En este informe vamos a explicar la realización de un modelo predictivo de redes neuronales ANN para la calidad del aire en china que se ha convertido en un problema creciente. La contaminación del aire, causada principalmente por las emisiones de gases y partículas de la industria, el tráfico y otras actividades humanas, tiene un impacto significativo en la salud pública y el medio ambiente.

Un modelo predictivo es una representación matemática o estadística que hace predicciones futuras basadas en datos históricos y otros factores relevantes. En el caso de la contaminación del aire en China, los modelos pueden usar datos recopilados de estaciones de monitoreo ubicadas en todo el país, así como información sobre las condiciones climáticas, la actividad industrial y otros factores influyentes.

Definición del problema:

La contaminación del aire en China es un problema alarmante que ha alcanzado niveles graves en las últimas décadas. El país tiene altos niveles de contaminantes del aire que afectan negativamente la calidad del aire, la salud pública y el medio ambiente. Algunos de los principales problemas relacionados con la contaminación del aire en China son:

- Altos niveles de partículas y material particulado
- Emisiones de gases contaminantes
- Impacto en la salud pública
- Impacto ambiental

En este proyecto tomaremos la medida de varios elementos presentes en el aire para calcular el AQI final es la puntuación más alta de esos cinco contaminantes.

Preprocesamiento de datos:

Para entrenar un modelo predictivo necesitamos primero realizar un análisis de los datos para identificar aquellos campos que puedan tener datos nulos, poco relevantes o necesiten una conversión para poder ver su relevancia con la variable que buscamos predecir.

```
[ ] #Comprobar si existen valores null en alguna columna del dataset
    df.isnull().sum()
```

```
No            0
year          0
month         0
day           0
hour          0
PM2.5       925
PM10        718
SO2         935
NO2        1023
CO         1776
O3         1719
TEMP         20
PRES         20
DEWP         20
RAIN         20
wd           81
WSPM         14
station       0
dtype: int64
```

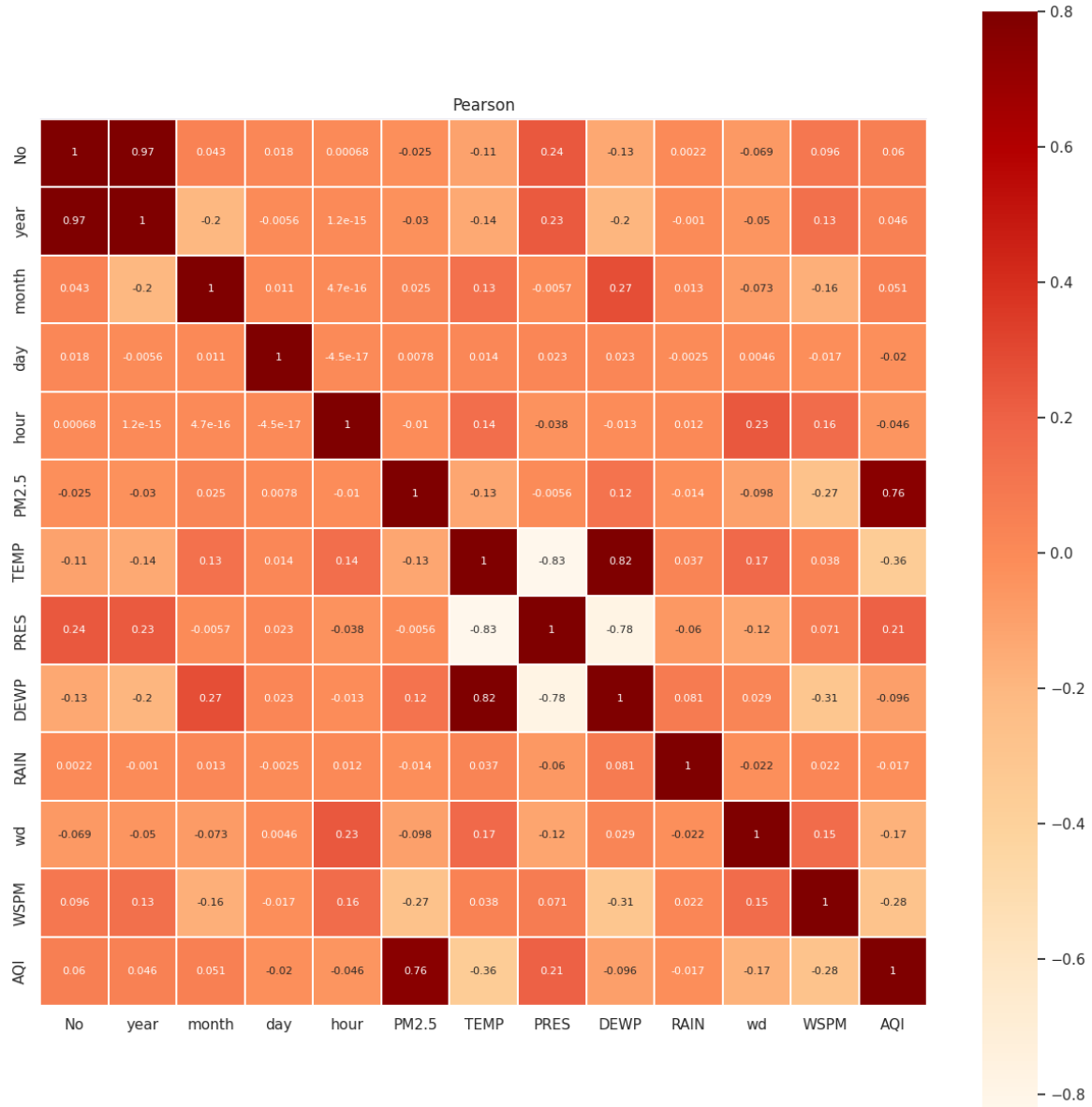
```
[ ] #Informacion de las columnas del dataset
    df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35064 entries, 0 to 35063
Data columns (total 18 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   No       35064 non-null  int64  
 1   year     35064 non-null  int64  
 2   month    35064 non-null  int64  
 3   day      35064 non-null  int64  
 4   hour     35064 non-null  int64  
 5   PM2.5    34139 non-null  float64
 6   PM10     34346 non-null  float64
 7   SO2      34129 non-null  float64
 8   NO2      34041 non-null  float64
 9   CO       33288 non-null  float64
 10  O3       33345 non-null  float64
 11  TEMP     35044 non-null  float64
 12  PRES     35044 non-null  float64
 13  DEWP     35044 non-null  float64
 14  RAIN     35044 non-null  float64
 15  wd       34983 non-null  object 
 16  WSPM     35050 non-null  float64
 17  station  35064 non-null  object 
dtypes: float64(11), int64(5), object(2)
memory usage: 4.8+ MB
```

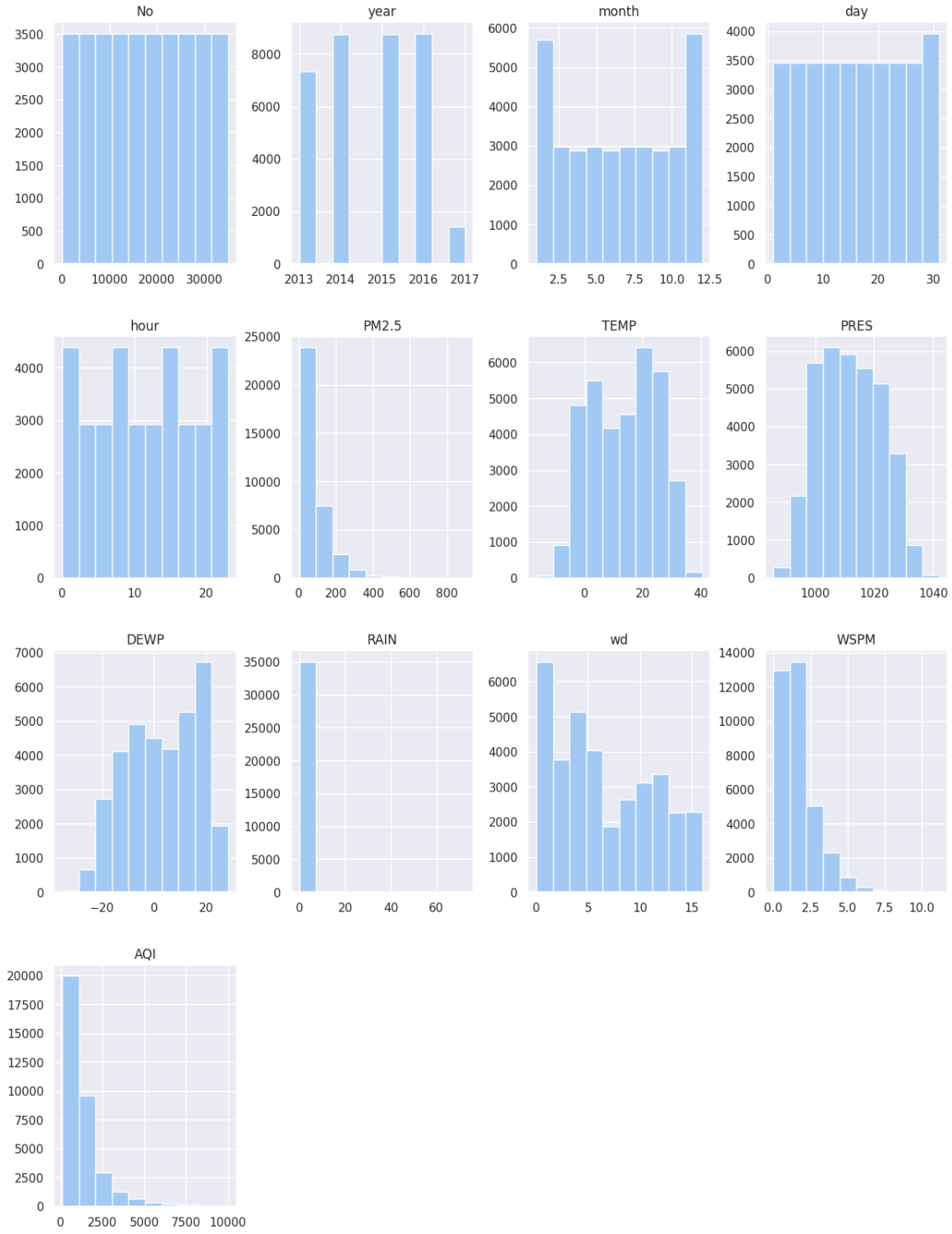
```
[ ]  #Aplicar LabelEncoder para convertir variables categoricas a numericas
     le = preprocessing.LabelEncoder()
     df['wd'] = le.fit_transform(df['wd'])

[ ]  #Eliminar la columna "station" ya que no aporta informacion relevante
     #debido a que toda la data proviene de esa estacion
     df = df.drop('station',axis=1)
```

Análisis exploratorio de datos:

Realizamos un análisis del diagrama de correlación de Pearson para identificar las variables más relevantes para la predicción del AQI.



Además de visualizar el histograma de los datos para reconocer cual sería la mejor función de activación para cada capa del modelo de redes neuronales.

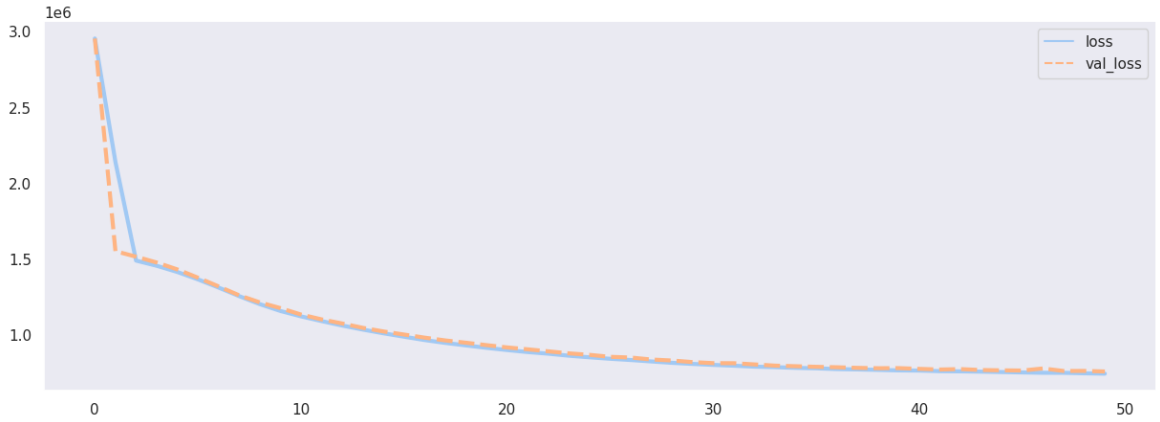


Modelado predictivo:

Usando un modelo de redes neuronales realizamos un entrenamiento según la siguiente configuración.

```
[ ]  #Modelo ANN
     ann = Sequential()
     ann.add(Dense(25,activation='relu'))
     ann.add(Dense(19,activation='relu'))
     ann.add(Dense(14,activation='relu'))
     ann.add(Dense(7,activation='relu'))
     ann.add(Dense(1))

[ ]  #Compilacion del modelo
     ann.compile(optimizer='adam', loss='mse')

[ ]  #Entrenamiento del modelo
     ann.fit(x=X_train,y=Y_train,validation_data=(x_test,y_test),batch_size=200,epochs=50,verbose=1)
```

Luego del entrenamiento revisamos el grafico de los vs val_loss para comprobar si la configuración del modelo a conseguido un entrenamiento aceptable.



Conclusiones:

- El uso de modelos predictivos para elementos contaminantes en el aire de China es una herramienta importante en la lucha contra la contaminación y la protección de la salud pública.
- Los modelos predictivos basados en redes neuronales pueden capturar patrones de datos no lineales complejos, lo que puede proporcionar predicciones más precisas y confiables. Estos modelos pueden analizar grandes conjuntos de datos históricos y capturar relaciones entre variables ambientales, emisiones industriales, condiciones climáticas y niveles de contaminación.