Reporte 1

Juan David Gómez & Juan David Rengifo Mera & Francisco Suárez
Proces. de Grandes Volúmenes
Pontificia Universidad Javeriana Cali

Sep 23, 2021

1 Descripción del dataset

1.1 Descripción general

El dataset contiene 9358 instancias que corresponden a medidas tomadas cada hora de 5 sensores que miden la calidad del aire en una ciudad italiana. Estos sensores se encontraban a la altura de la carretera y recolectaron datos entre marzo de 2004 y febrero de 2005(un año).

1.2 Atributos

- Date: (DD/MM/YYYY)
- Time: (HH.MM.SS)
- Promedio de concentración de CO en mg/m^3 por hora(analizador de referencia).
- PT08.S1: (Oxido de estaño) promedio de respuesta del sensor(nominalmente dirigido a CO).
- Promedio de concentración de hidrocarburos no metálicos en $microg/m^3$ por hora(analizador de referencia).
- Promedio de concentración de benceno en $microg/m^3$ por hora(analizador de referencia).
- PT08.S2: (titanio) promedio de respuesta del sensor(nomialmente dirigido a NMHC).
- Promedio de concentración de NOX en ppb(partículas por billón) por hora(analizador de referencia).
- PT08.S3: (Oxido de tungsteno) promedio de respuesta del sensor(nominalmente dirigido a NOx).
- Promedio de concentración de NO2 en $microg/m^3$ por hora(analizador de referencia).
- PT08.S4: (Oxido de tungsteno) promedio de respuesta del sensor(nominalmente dirigido a NO2).

- PT08.S5: (Oxido de indio) promedio de respuesta del sensor(nominalmente dirigido a O3).
- Temperatura en centígrados
- Humedad absoluta

1.3 Variable objetivo

Se busca encontrar el atributo PT08_S5(O3), ...

2 Análisis de los datos

Al hacer análisis del conjunto de datos se pudo observar lo siguiente:

- El conjunto de datos posee datos nulos.
- El conjunto de datos posee una cantidad considerable de entradas.
- La cantidad de atributos permite realizar una regresión por medio de árboles de decisión.
- Todas las entradas (a excepción de la hora y fecha) son de tipo numérico.
- La matriz de correlación (Fig 1.)muestra una fuerte relación entre las moléculas (a excepción de NOx) y muestra que las moléculas tienden a tener una baja relación con las medidas de temperatura y humedad. Esto hace que no sea necesario eliminar atributos del modelo.
- La gráfica de dispersión (Fig 2.) muestra que la variable objetivo tiene un comportamiento ajustable para la realización de una regresión lineal.

3 Transformaciones realizadas

Al ser tan grande la cantidad de entradas en el conjunto de datos, cómodamente se procedió a eliminar todas las entradas que tuvieran datos nulos. Al final se obtuvo una precisión bastante decente, así que no se consideró realizar otros cambios a los datos.

4 Análisis de resultados

La variable a predecir es PT08_S5(O3). Esta variable representa...

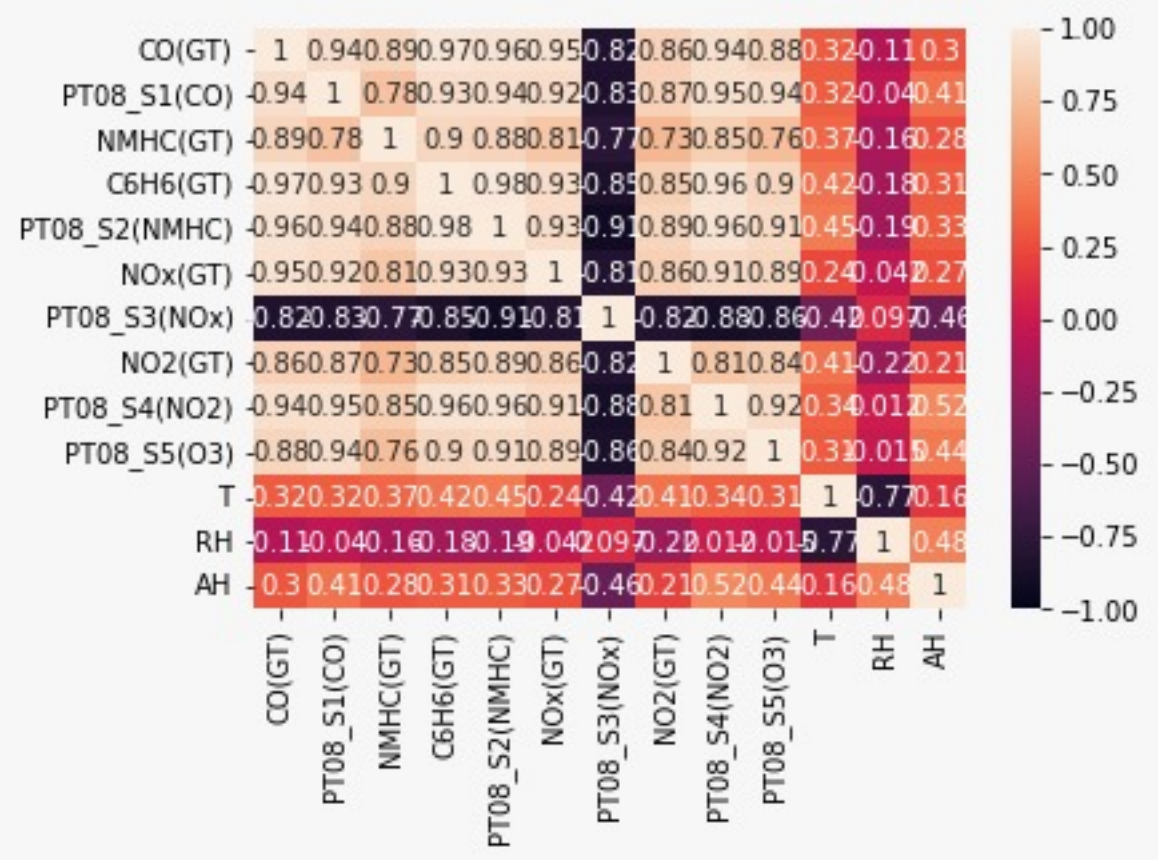

Figure 1: Correlación entre atributos

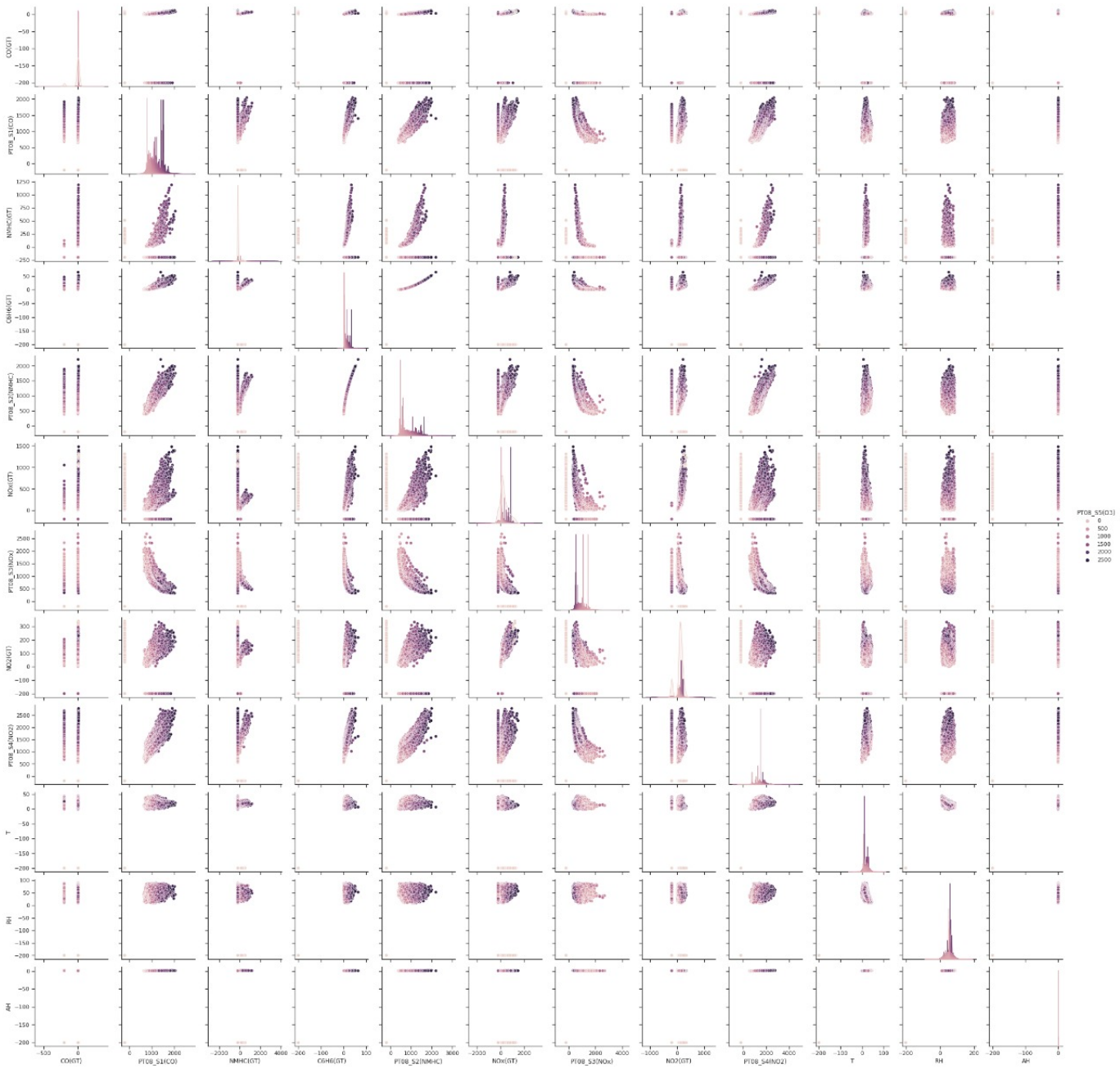

Figure 2: Gráfico de dispersión del atributo objetivo PT08_S5(O3)

4.1 Regresión lineal

Los resultados obtenidos con esta técnica son bastante buenos. A continuación se muestra en detalle las métricas obtenidas del análisis:

- **R-squared:** Es una medida estadística que nos muestra que tan cerca están los datos de la linea de regresión ajustada, para este modelo obtuvimos un R-squared 0.8759, lo que indica que el modelo explica una buena porción de la variabilidad de los datos de respuesta en torno a su media.

- **Explained Variance:** Esta métrica se usa para medir la discrepancia entre los datos reales y las perdiciones, representa que tan fuerte es la asociación dentro del modelo, para esta métrica obtuvimos un valor de 118550.26 lo que implica una fuerte asociación y una excelente predicción.

- **Mean Squared Error:** Esta métrica nos indica que tan cercanos están los puntos a nuestra linea de regresión, la distancia entre el punto y la linea es el error (se eleva al cuadrado para quitar los valores negativo) para esta métrico obtuvimos un valor de 1587.

- **Root Mean Squared Error:** Esta métrica es la Raíz de el Mean Squared Erro, para este modelo obtuvimos un valor de 125.99 .

- **Mean Absolute Error:** Esta métrica nos muestra promedio del error absoluto de nuestro modelo, es decir el promedio de la diferencia entre todas las predicciones con el valor real,para este modelo obtuvimos un Mean Absolute Error de 102.16.

  Las métricas de R-squared y Explained Variance nos indican que nuestro modelo es confiable y es capaz de hacer unas buenas predicciones debido a la cercanía de los puntos respecto a la recta y a la fuerte asociación dentro del modelo.Finalmente, a pesar de que los errores parecen ser grande realmente no lo son, esa es la magnitud esperable si tenemos en cuenta que la media de nuestro dato objetivo es 1045.81 y la desviación estándar es de 400.3.

4.2 Decision Tree Regressor

Al igual que con el modelo de regresión lineal, los resultados de este modelo son bastante prometedores. A continuacion se muestra en detalle las métricas obtenidas del análisis,

- **R-squared:** Es una medida estadística que nos muestra que tan cerca están los datos de la linea de regresión ajustada, para este modelo obtuvimos un R-squared 0.8769, lo que indica que el modelo explica una buena porción de la variabilidad de los datos de respuesta en torno a su media.

- **Mean Squared Error:** Esta métrica nos indica que tan cercanos están los puntos a nuestra linea de regresión, la distancia entre el punto y la linea es el error (se eleva al cuadrado para quitar los valores negativo) para esta métrico obtuvimos un valor de 1548.

- **Root Mean Squared Error:** Esta métrica es la Raiz de el Mean Squared Erro, para este modelo obtuvimos un valor de 125.49

  Al igual que con el modelo de regresión lineal, la métrica de R-squared al ser tan cercana a 1, nos indica que el modelo es capaz de hacer predicciones bastante buenas. Las métricas de error siguen siendo de una magnitud aceptable al tener en cuenta aspectos como la media y la varianza de el dato objetivo.

5 Comparación de técnicas de aprendizaje automático

Como se puede observar en las métricas ambos modelos obtienen resultados similarmente buenos, lo que nos indica que ambos modelos son viables para realizar la predicción de este dataset.

También se evidencia que para este uso en particular las diferencias entre ambos modelos son poco relevantes, por lo que en otros datasets puede que la diferencia entre las métricas de los modelos sea significativa.

6 Descripción de streaming y aplicación a grafos

7 Referencias:

- `https://archive-beta.ics.uci.edu/ml/datasets/auto+mpg`