Datathon

No sabemos Python

20/12/2022

1. Introducción

El reto presentado trata de crear una inteligencia artificial basada en machine learning para poder aplicar en la industria 4.0 y así poder lograr el objetivo de cero defectos. Esta AI evaluará procesos industriales en una cadena de producción para poder predecir el nivel de fiabilidad de las piezas fabricadas.

Para nuestra solución, se han generado varias inteligencias artificiales que emplean distintos algoritmos de machine learning para poder determinar qué modelo es el óptimo.

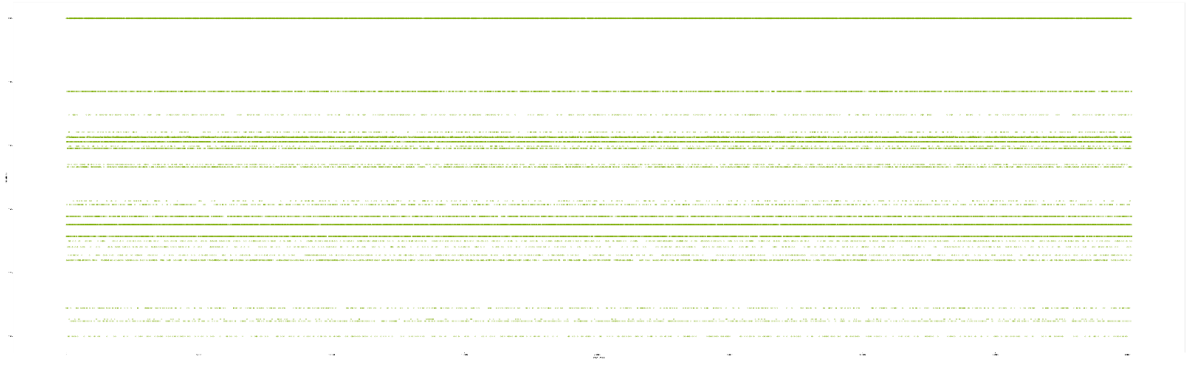
2. Análisis exploratorio de datos

El análisis exploratorio de datos (Exploratory Data Analysis, o EDA) se trata de un método empleado para analizar e investigar conjuntos de datos, así como resumir sus principales características. En este proyecto se ha utilizado el EDA para evaluar datos de industrias, tanto para entrenar los distintos modelos de

inteligencia artificial como para predecir los resultados de fiabilidad en el modelo final.

3. Tareas de preprocesamiento

Los datos proporcionados eran datos brutos, por lo cual era necesario tratarlos (o pre procesarlos) para poder emplearlos como input para los modelos de



inteligencia artificial.

4. Comparación de modelos predictivos

Para poder comparar los distintos algoritmos de machine learning, se tiene en cuenta el valor de r^2. Este valor (también llamado coeficiente de determinación) indica cuánto se acerca el valor de fiabilidad predecido al valor real. Se han obtenido los valores de r^2 de todos los modelos. También se ha obtenido el error cuadrático medio (mean squared error, o MSE) para cada modelo, el cual indica el promedio del error al cuadrado. Es decir, el MSE mide cuánto se aleja el valor

medido del valor predecido.

	Method	Training MSE	Training R2	Test MSE	Test R2
0	Linear regression	0.026269	0.653809	45963637328.191605	-609552859335.952515
1	Random forest	0.054894	0.276561	0.053715	0.287655
2	K-Nearest Neighbors	0.000056	0.999266	0.014994	0.801152
3	S-Vector Machine	0.04492	0.408009	0.043429	0.424063

5. Selección del mejor modelo

Se han probado cuatro modelos predictivos: Regresión linear, Bosques aleatorios, KNN y SVM. Tras evaluar la r^2 y el MSE de todos los modelos, se han agrupado los resultados en la siguiente tabla:

```
#
#!              Method  Training MSE     Training R2           Test MSE              Test R2
#! 0  Linear regression      0.026466        0.651201    2603143596.200458   -34521933304.644592 ------> Worst model
#! 1      Random forest      0.054894        0.276561            0.053715              0.287655
#! 2        KNNeighbors      0.000056        0.999266            0.014994              0.801152 ------> Best model
#! 3                SVM      0.044920        0.408009            0.043429              0.424063
#
```

Ya que buscamos el valor de r^2 más alto y el MSE más bajo, se puede afirmar que el modelo de KNN (k-nearest neighbors) es el mejor, en el sentido de que es el más preciso a la hora de predecir la fiabilidad.