Detecció i predicció d'anomalies en dispositius IoT en l'Edge computing

Antoni Llussà i Sala
Treball Final de Màster
Màster Universitari en Ciència de Dades

Índex

1. Introducció: Motivació

- Especial interès en l'àmbit de la Internet de les coses (IoT).
- En general, el Machine Learning.
- Especial interès en la detecció d'anomalies.
- Aplicar els coneixements adquirits durant el transcurs del Màster.

1. Introducció: Objectius

- Analitzar l'estat de l'art de treballs relacionats de Machine Learning, que tinguin models per la detecció i predicció d'anomalies en les dades en l'Edge computing.
- Realitzar diferents models de Machine Learning per la detecció i predicció d'anomalies en series temporals.
- Desenvolupar l'aplicació de l'Edge computing per poder realitzar les prediccions.
- Implementar els models predictius pel microcontrolador.
- Realitzar deteccions d'anomalies en temps real.

1. Introducció: Metodologia

S'utilitzarà la metodologia CRISP-DM (Cross Industry Standard Process for Data Mining)

És una metodologia iterativa, on s'estableixen petits cicles de planificació, execució i revisió.
Consta de sis fases:

- Comprensió del negoci.
- Comprensió de les dades.
- Preparació de les dades.
- Modelat.
- Avaluació del model.
- Desplegament.

2. Estat de l'art

- Jaewon Moon, Seungwoo Kum, and Sangwon Lee. A heterogeneous iot data analysis framework with collaboration of edge-cloud computing: Focusing on indoor pm10 and pm2.5 status prediction. Sensors, 19(14), 2019.
- Joseph Schneible and Alex Lu. Anomaly detection on the edge. In MILCOM 2017-2017 IEEE Military Communications Conference (MILCOM), pages 678–682. IEEE, 2017.
- Yuhuai Peng, Aiping Tan, Jingjing Wu, and Yuanguo Bi. Hierarchical edge computing: A novel multi-source multi-dimensional data anomaly detection scheme for industrial internet of things. IEEE Access, 7:111257–111270, 2019.Adaptive anomaly detection for IoT data in hierarchical edge computing
- B. Hussain, Q. Du, S. Zhang, A. Imran, and M. A. Imran. Mobile edge computing based data-driven deep learning framework for anomaly detection. IEEE Access, 7:137656–137667, 2019Squeezed Convolutional Variational AutoEncoder for unsupervised anomaly detection in edge device industrial Internet of Things.
- Darmawan Utomo and Pao-Ann Hsiung. A multitiered solution for anomaly detection in edge computing for smart meters. Sensors, 20(18):5159, 2020
- Mao V Ngo, Hakima Chaouchi, Tie Luo, and Tony QS Quek. Adaptive anomaly detection for iot data in hierarchical edge computing. arXiv preprint arXiv:2001.03314, 2020.
- D. Kim, H. Yang, M. Chung, S. Cho, H. Kim, M. Kim, K. Kim, and E. Kim. Squeezed convolutional variational autoencoder for unsupervised anomaly detection in edge devi ce industrial internet of things. In 2018 International Conference on Information and Computer Technologies (ICICT), pages 67–71, 2018.

3. Arquitectura

Dispositius

Argon Wi-Fi Develompment Board

- Processador ARM Cortex-M4F de 32 bits a 64 MHZ
- Flash de 1 MB RAM de 256 KB
- Coprocessador Wi-Fi Espressif ESP32-D0WD 2.4 GHx
- Flaix integrada de 4MB pel ESP32

Laser PM2.5 HM-3301 Dust Sensor

- Alta sensibilitat en partícules de pols de 0.3 µm o més
- Admet sortida de sis canals de 0.3 µm, 0.5 µm, 1.0 µm, 2.5 µm, 5 µm i 10 µm.
- Detecció contínua i en temps real de la concentració de pols a l'aire.

Recol·lecció de les dades IoT

3. Arquitectura: Tecnologia utilitzada

Machine Learning (ML):

- Anaconda
- TensorFlow
- Numpy
- Pandas
- Sklearn: Llibreries de ML

Particle:

- Web Ide

REST-APIS:

- Django
- TensorFlow
- Numpy
- Pandas
- Sklearn: Llibreries de ML

4. Implementació del treball: Dades utilitzades

Dos conjunts de dades. Dades històriques i dades recopilades durant el transcurs del treball.

Dades Històriques:

- Recopilades en l'estudi: Darmawan Utomo and Pao-Ann Hsiung. A multitiered solution for anomaly detection in edge computing for smart meters. Sensors, 20(18):5159, 2020.
- Tres períodes de recol·lecció entre el 16/09/2019 al 18/11/2019
- Cada període les dades han estat recopilades en situacions diferents.

Dades recopilades durant el treball:

- Un període de recol·lecció entre el 10/11/2020 i el 13/12/2020
- Mateixa situació en les dades recopilades.

4. Implementació del treball: Exploració i preparació de les dades

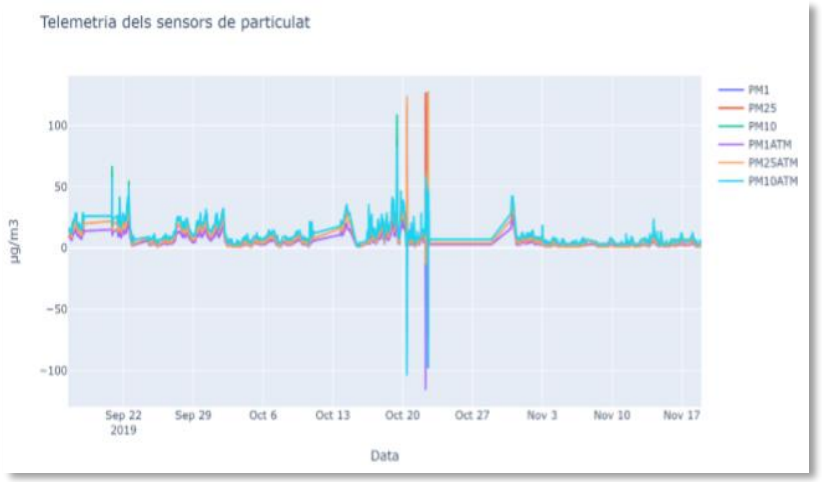
Anàlisis estadístic i exploratori de les dades, i preparació de les dades.

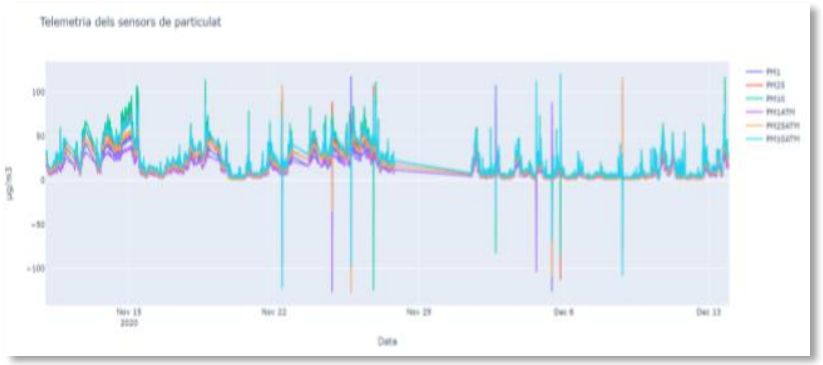
Dades històriques:

- 4989 registres.
- 7 variables: Time, PM1, PM2.5, PM10, PM1 ATM, PM2.5 ATM, PM10 ATM .

Dades treball:

- 3897 registres.
- 7 variables: Time, PM1, PM2.5, PM10, PM1 ATM, PM2.5 ATM, PM10 ATM .





4. Implementació del treball: Algoritmes de Machine Learning (I)

Conjunts d'entrenament i test

Principi de Pareto, en molts casos, el 80% dels efectes són conseqüència del 20% de les causes.

- Entrenament: 80%
- Test: 20%

Xarxes Neuronals, s'han de crear subseqüències pels subconjunts d'entrenament i de test per poder predir les dades en la finestra de temps que es vulgui fer la predicció.

4. Implementació del treball: Algoritmes de Machine Learning (II)

Tècniques per evitar el sobreentrenament

Xarxes Neuronals

DropOut: La capa de drop-out té la funció molt especifica en les xarxes neuronals, que és, prevenir el sobreentrenament. Desactiva un número aleatori d'entrades de la capa.

EarlyStopping: Funcionalitat de la llibreria keras de python, evita l'execució de masses èpoques d'entrenament degut a que poden provocar un sobreajustament del conjunt de dades d'entrenament.

4. Implementació del treball: Algoritmes de Machine Learning (III)

Avaluació dels models

Xarxes Neuronals

Mètriques de regressió:

- Mean squared error (MSE)
- Root mean squared error (RMSE)
- Mean absolute error (MAE)

Isolation Forest

Es tindrà en compte els valors atípics detectats pel diagrama Boxplot (nombre teòric).

- Valors predits / Nombre teòric

4. Implementació del treball: Algoritmes de Machine Learning (IV)

Determinació d'anomalia

Xarxes Neuronals

Funció de pèrdua > Càlcul del threshold => Anomalia

```
#càlcul del threshold de test
def calculate_threshold(X_test, X_test_pred):
    distance = np.linalg.norm(X_test - X_test_pred, axis=1);
    """Sorting the scores/diffs and using a 0.80 as cutoff value to pick the threshold"""
    distance.sort();
    cut_off = int(0.80 * len(distance));
    threshold = distance[cut_off];
    return threshold
```

Isolation Forest

La mateixa funció de predicció del model, retorna 1 si és anomalia, -1 si no ho és.
Amb l'ajuda del paràmetre "contamination" es pot acabar d'ajustar el model.

4. Implementació del treball: Algoritmes de Machine Learning (V)

Models Utilitzats

Xarxes Neuronals Recurrents (RNN)

- **Long Short Term Memory (LSTM)**

  Extensió de les RNN, bàsicament amplien la seva memòria per aprendre d'experiències importants que han passat fa molt temps.

- **Gated Recurrent Unit (GRU)**

  Usen el mateix principi que les LSTM, però estan simplificades de manera que el seu rendiment és similar però són més eficients computacionalment.

Xarxes Neuronals Profundes (DNN)

Xarxa artificial amb múltiples capes entre les capes d'entrada i sortida. Normalment són xarxes de retroalimentació en que les dades flueixen des de la capa d'entrada a la cap de sortida sense retrocedir.

Isolation Forest

Algorisme d'aprenentatge no supervisat per identificar anomalies quan les dades no estan etiquetades.

4. Implementació del treball: Desplegament del models a producció

- Plantejament inicial
- Replantejament final

6. Resultats (I)

Resultats òptims de la xarxa neuronal recurrent GRU

Sequence	Activation	Optimizer	DropOut	Units	Epochs	BatchSize	Validation	RMSE	MSE	MAE	Seconds
1h	tanh	adam	0.118	55	36	9	0.2	0.582	0.338	0.154	58.817
3h	tanh	adamax	0.118	55	36	9	0.2	0.584	0.341	0.165	58.063
6h	tanh	adamax	0.118	55	36	9	0.2	0.588	0.345	0.174	101.704
12h	tanh	adamax	0.405	40	68	45	0.1	0.603	0.363	0.200	96.051
1d	tanh	adamax	0.118	55	36	9	0.2	0.657	0.432	0.2568	204.052
3d	tanh	adadelta	0.118	55	36	9	0.2	0.969	0.9385	0.641	1631.155
7d	sigmoid	adamax	0.405	40	68	45	0.1	1.023	1.046	0.689	1643.763

Resultats òptims de la xarxa neuronal recurrent LSTM

Sequence	Activation	Optimizer	DropOut	Units	Epochs	BatchSize	Validation	RMSE	MSE	MAE	Seconds
1h	tanh	adam	0.405	40	68	45	0.1	0.582	0.338	0.159	27.413
3h	tanh	adamax	0.118	55	36	9	0.2	0.586	0.343	0.169	105.727
6h	tanh	adamax	0.118	55	36	9	0.2	0.593	0.351	0.189	146.728
12h	tanh	adamax	0.405	40	68	45	0.1	0.659	0.434	0.259	72.936
1d	tanh	adamax	0.405	40	68	45	0.1	0.830	0.688	0.389	7 7.335
3d	tanh	adadelta	0.118	55	36	9	0.2	0.941	0.885	0.612	1523.291
7d	sigmoid	adamax	0.118	55	36	9	0.2	1.0230	1.046	0.689	3432.494

6. Resultats (II)

Resultats òptims de la xarxa neuronal recurrent GRU amb dues capes

Sequence	Activation	Optimizer	DropOut	Units	Epochs	BatchSize	Validation	RMSE	MSE	MAE	Seconds	
1h	tanh	adam	0.118	0.243	45	43	30	0.2	0.574	0.329	0.157	43.191
3h	tanh	adamax	0.405	0.331	43	56	11	0.1	0.579	0.335	0.178	157.909
6h	tanh	adamax	0.118	0.243	45	43	30	0.2	0.583	0.340	0.190	145.750
12h	tanh	adamax	0.118	0.243	45	43	30	0.2	0.609	0.371	0.215	274.628
1d	tanh	adadelta	0.405	0.331	43	56	11	0.1	0.841	0.707	0.505	1747.804
3d	tanh	adadelta	0.405	0.331	43	56	11	0.1	0.923	0.852	0.562	4529.597
7d	sigmoid	adam	0.405	0.331	43	56	11	0.1	1.019	1.039	0.695	6280.207

Resultats òptims de la xarxa neuronal recurrent LSTM amb dues capes

Sequence	Activation	Optimizer	DropOut	Units	Epochs	BatchSize	Validation	RMSE	MSE	MAE	Seconds	
1h	tanh	adam	0.118	0.243	45	43	30	0.2	0.573	0.329	0.155	40.125
3h	tanh	adamax	0.118	0.243	45	43	30	0.2	0.580	0.337	0.180	80.390
6h	tanh	adamax	0.118	0.243	45	43	30	0.2	0.591	0.349	0.207	127.388
12h	tanh	adamax	0.118	0.243	45	43	30	0.2	0.787	0.619	0.316	246.600
1d	tanh	adadelta	0.405	0.331	43	56	11	0.1	0.857	0.734	0.491	1,547.029
3d	tanh	adadelta	0.405	0.331	43	56	11	0.1	0.942	0.887	0.553	4,229.342
7d	tanh	adadelta	0.118	0.243	45	43	30	0.2	1.011	1.023	0.682	2,678.760

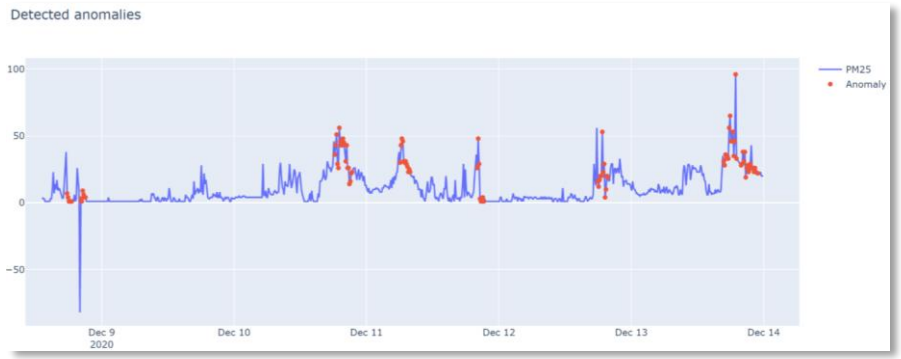
6. Resultats (III)

Resultats òptims de la xarxa neuronal profundes

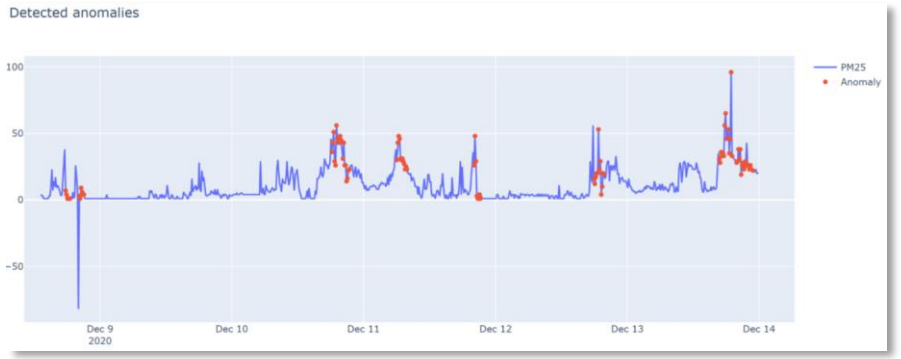
Sequence	Activation	Optimizer	DropOut	Units	Epochs	BatchSize	Validation	RMSE	MSE	MAE	Seconds
1h	sigmoid	adamax	0.779	22	85	23	0.2	0.023	0.001	0.014	18.778
3h	sigmoid	adamax	0.779	22	85	23	0.2	0.032	0.001	0.020	22.037
6h	tanh	adamax	0.779	22	85	23	0.2	0.033	0.001	0.020	27.322
12h	sigmoid	adamax	0.779	22	85	23	0.2	0.041	0.002	0.025	29.853
1d	sigmoid	adam	0.779	22	85	23	0.2	0.043	0.002	0.026	39.955
3d	sigmoid	adamax	0.796	97	98	43	0.1	0.043	0.002	0.025	173.990
7d	tanh	adam	0.796	97	98	43	0.1	0.046	0.002	0.026	313.318

6. Resultats (IV)

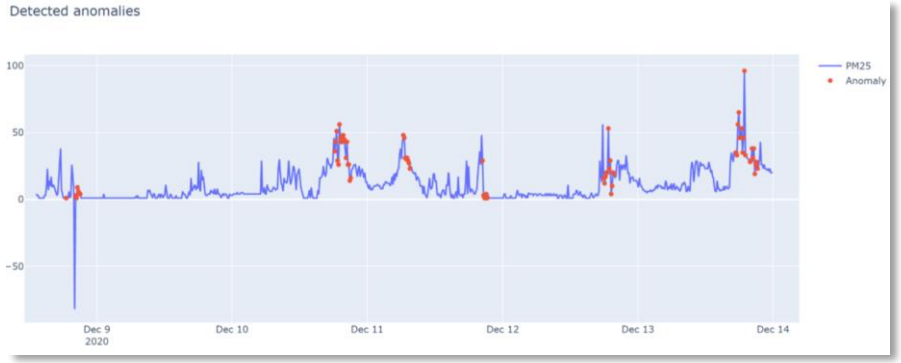
Visualització d'anomalies RNN-GRU – 1h



Visualització d'anomalies RNN-LSTM – 1h

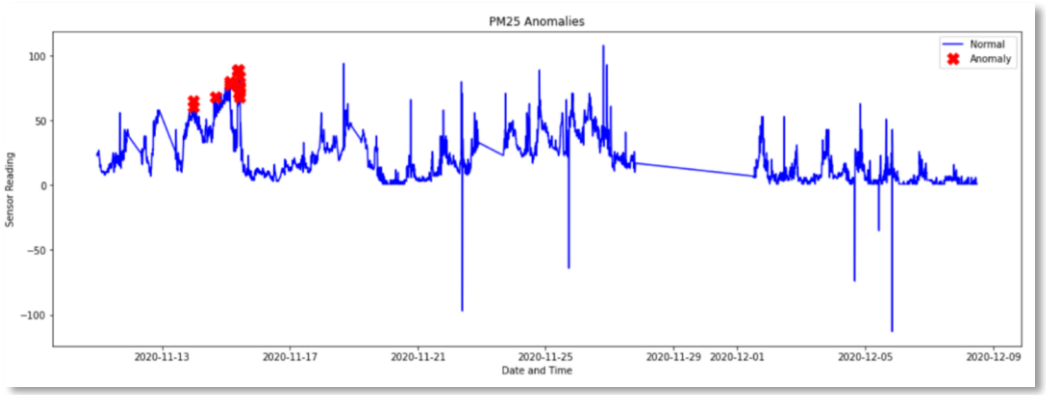


Visualització d'anomalies DNN – 1h

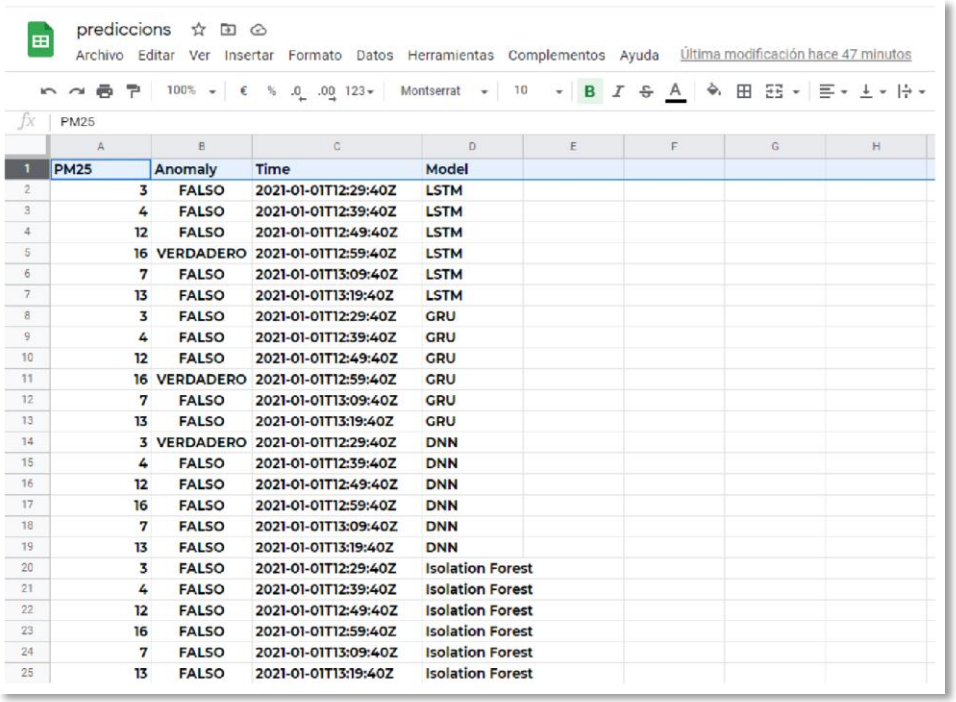


6. Resultats (V)

Visualització d'anomalies Isolation Forest



Detecció d'anomalies en temps real



prediccions

Archivo Editar Ver Insertar Formato Datos Herramientas Complementos Ayuda Última modificación hace 47 minutos

PM25

	A	B	C	D
1	PM25	Anomaly	Time	Model
2	3	FALSO	2021-01-01T12:29:40Z	LSTM
3	4	FALSO	2021-01-01T12:39:40Z	LSTM
4	12	FALSO	2021-01-01T12:49:40Z	LSTM
5	16	VERDADERO	2021-01-01T12:59:40Z	LSTM
6	7	FALSO	2021-01-01T13:09:40Z	LSTM
7	13	FALSO	2021-01-01T13:19:40Z	LSTM
8	3	FALSO	2021-01-01T12:29:40Z	GRU
9	4	FALSO	2021-01-01T12:39:40Z	GRU
10	12	FALSO	2021-01-01T12:49:40Z	GRU
11	16	VERDADERO	2021-01-01T12:59:40Z	GRU
12	7	FALSO	2021-01-01T13:09:40Z	GRU
13	13	FALSO	2021-01-01T13:19:40Z	GRU
14	3	VERDADERO	2021-01-01T12:29:40Z	DNN
15	4	FALSO	2021-01-01T12:39:40Z	DNN
16	12	FALSO	2021-01-01T12:49:40Z	DNN
17	16	FALSO	2021-01-01T12:59:40Z	DNN
18	7	FALSO	2021-01-01T13:09:40Z	DNN
19	13	FALSO	2021-01-01T13:19:40Z	DNN
20	3	FALSO	2021-01-01T12:29:40Z	Isolation Forest
21	4	FALSO	2021-01-01T12:39:40Z	Isolation Forest
22	12	FALSO	2021-01-01T12:49:40Z	Isolation Forest
23	16	FALSO	2021-01-01T12:59:40Z	Isolation Forest
24	7	FALSO	2021-01-01T13:09:40Z	Isolation Forest
25	13	FALSO	2021-01-01T13:19:40Z	Isolation Forest

7. Conclusions

- Resultats similars entre els dos tipus de Xarxes Neuronals Recurrents (LSTM i GRU)
- Bons resultats de les Xarxes Neuronals Profundes, segons les mètrica RMSE.
- Xarxes Neuronals Profundes més ràpides que les Xarxes Neuronals Recurrents.
- Les Xarxes Neuronals Profundes en la detecció en temps real, detecten més franges d'anomalies que les Xarxes Neuronals Recurrents.
- L'Isolation Forest en les proves realitzades han estat bones, però alhora de realitzar la predicció en temps real, no han detectat cap anomalia.
- Cal destacar la rapidesa de les prediccions en l'Isolation Forest envers de les Xarxes Neuronals.

8. Línies de futur

- Seguir l'evolució del dispositiu particle Argon.
- Realitzar el mateix estudi amb un data set de dades més gran, i a ser amb dades separades, per exemple interior i exterior.
- Realitzar un estudi multi-variant per veure el comportament de les dades, afegint nous sensors, per exemple el de temperatura, humitat.
- Millorar els models proposats i realitzar un estudi més profund per aconseguir una predicció més en temps real, sense necessitar 6 registres.

Gracias
Thank you
Gràcies

Antoni Llussà i Sala
Treball Final de Màster
Màster Universitari en Ciència de Dades