“Modelización para la Predicción de Sensores de Gas para Monitorización de Actividad Doméstica”

—

Jorge Durán, Jaime Enríquez, Marcos de las Heras

Índice

Introducción

- 8 sensores de gas, un sensor de temperatura y un sensor de humedad.
- Dos estímulos: vino y plátano, y ausencia de estímulos.
- Clasificar respuestas a los estímulos mediante técnicas de aprendizaje automático.

Descripción y Análisis de Dataset.

- Dataset formado por dos ficheros:

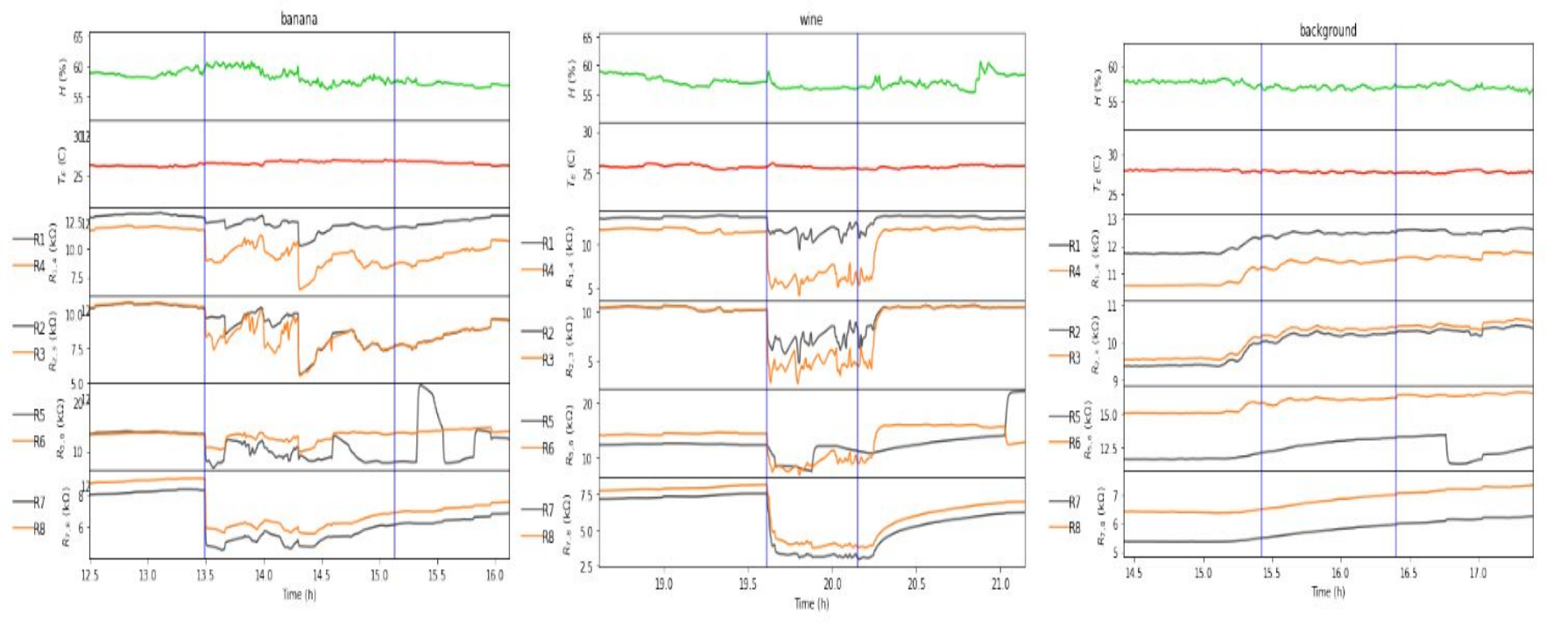
  - ❖ HS_Sensor_Metadatos.dat
    - ➢ id
    - ➢ date
    - ➢ class
    - ➢ t0
    - ➢ dt
  - ❖ HS_Sensor_Metadata.dat
    - ➢ id
    - ➢ time
    - ➢ R1-R8
    - ➢ Temp.
    - ➢ Humidity

- Atributos reales

Descripción y Análisis de Dataset.

- 100 inducciones: 36 vino, 33 plátano, 31 background.
- 928991 datos en HS_Sensor_Metadatos.dat.
- No hay datos para el experimento 95
- Los experimentos 14 y 76 no tienen datos posteriores a la inducción.

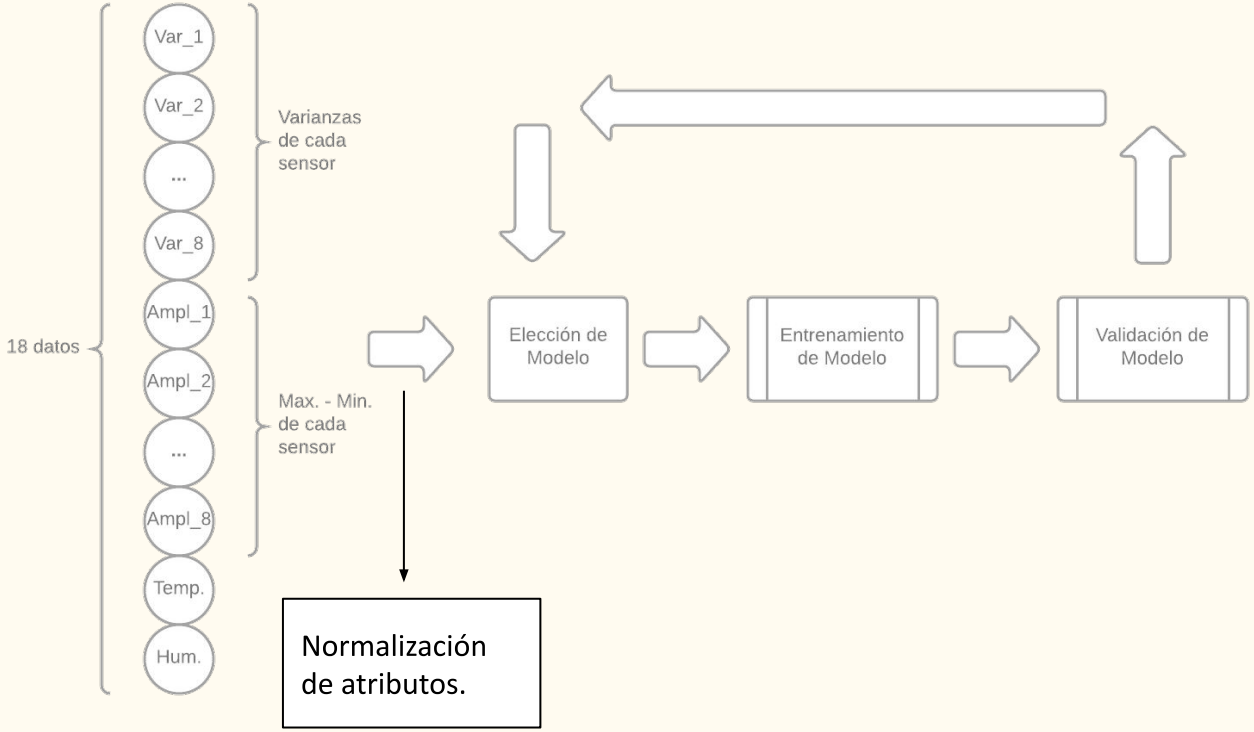
Descripción y Análisis de Dataset.



Elección de Atributos

- Variación parece ser clave.
  - Varianza por sensor.
  - Rango Max-Min.
  - Rango Intercuartil.
  - ...
- Media y Mediana no son útiles.
  - Muy similares para cada sensor.
- Temperatura y Humedad pueden ayudar.
  - Medias.
  - Varianzas.

Elección de Atributos



Elección de Modelos

- Módulo *LazyPredict*:
  - Entrenamiento de muchos modelos.
  - Poco código.
  - Se extraen conclusiones sobre modelos muy rápidamente.
- La mayoría son modelos son de Sci-kit Learn.
- Otros tienen sus propias librerías:
  - XDGBoosting: método de bagging con árboles de decisión + boosting de gradiente.
  - LightGBM: método de bagging con árboles de decisión muy similar a XDGBoosting
  - ...
- Métodos de Bagging obtienen los mejores resultados.

Model	Accuracy	F1 Score
LogisticRegression	0.79	0.77
ExtraTreesClassifier	0.77	0.76
LinearSVC	0.76	0.74
CalibratedClassifierCV	0.76	0.73
XGBClassifier	0.76	0.75
RandomForestClassifier	0.76	0.75
LinearDiscriminantAnalysis	0.75	0.73
RidgeClassifierCV	0.74	0.71
RidgeClassifier	0.74	0.72
NuSVC	0.74	0.71
SGDClassifier	0.73	0.70
BaggingClassifier	0.73	0.72
NearestCentroid	0.73	0.70
SVC	0.72	0.70
LGBMClassifier	0.70	0.69
AdaBoostClassifier	0.70	0.67
DecisionTreeClassifier	0.69	0.69
KNeighborsClassifier	0.68	0.65
Perceptron	0.68	0.64
ExtraTreeClassifier	0.67	0.65
GaussianNB	0.66	0.58
BernoulliNB	0.66	0.59
PassiveAggressiveClassifier	0.64	0.61
QuadraticDiscriminantAnalysis	0.60	0.55
LabelSpreading	0.60	0.60
LabelPropagation	0.60	0.59
DummyClassifier	0.34	0.34

Elección de Modelos

- Se escogen cinco de los modelos entrenados con *LazyPredict*.
  - Regresión Logística.
  - Random Forest.
  - K Nearest Neighbors.
  - Support Vector Machine Classifier.
  - Extreme Gradient Boosting.
- También se utiliza un modelo de Redes Neuronales.
  - Malos resultados.
- Entrenamiento y validación cruzada para cada uno.

Elección de Modelos

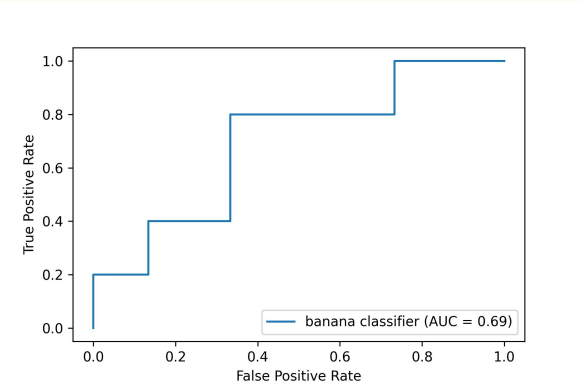
- Se refinan los parámetros de los dos modelos que mejores resultados han obtenido: Regresión Logística y Extreme Gradient Boosting.
- Se utiliza *GridSearchCV* de *Sci-kit Learn*.
- Para cada los mejores parámetros que se obtienen son:

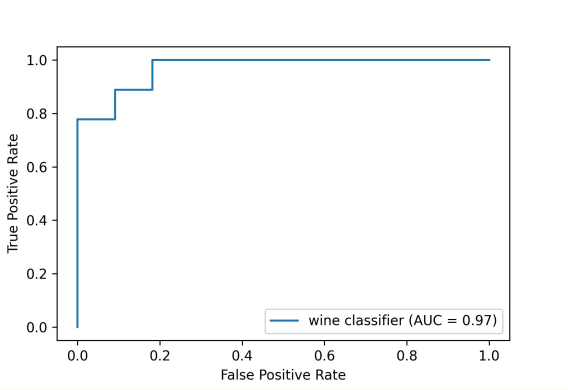
```
{'C': 0.1, 'penalty': 'l2'}
```

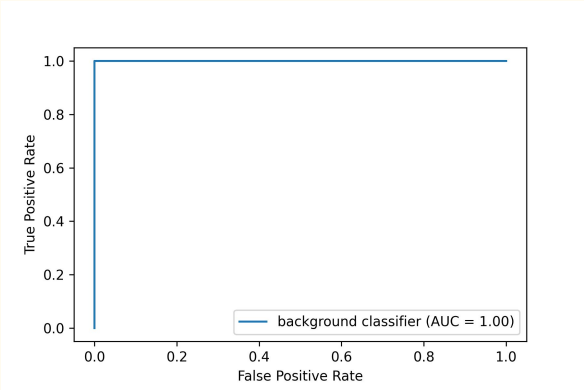
```
{'eta': 0.04,
 'gamma': 0.3,
 'max_depth': 5,
 'n_estimators': 30,
 'reg_lambda': 1}
```

Elección de Modelos

- Modelo de clasificación One VS Rest
- Curvas ROC por cada tipo de clase







Discusión de Resultados: Atributos

- Varianza de sensores de gas
- Máximos y mínimos de sensores
- Medias de Temperatura y Humedad
- Predicción general y banana: Durante la inducción
- Predicción wine: Durante y después de la inducción
- Predicción background: antes, durante y después de la inducción

Discusión de Resultados

- Valores aceptables para clasificador general
- Valores desaconsejables para clasificador banana
- Valores muy recomendables para clasificadores wine y background.

Referencias

[1] Ramon Huerta, Thiago Mosqueiro, Jordi Fonollosa, Nikolai Rulkov, Irene Rodriguez-Lujan. Online Decorrelation of Humidity and Temperature in Chemical Sensors for Continuous Monitoring. Chemometrics and Intelligent Laboratory Systems 2016.

[2] Pandas. McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

[3] Numpy. Oliphant, T. E. (2006). A guide to NumPy (Vol. 1). Trelgol Publishing USA.

[4] Sci-kit Learn. Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

[5] Matplotlib. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95.

[6] LazyPredict. https://github.com/shankarpandala/lazypredict

[7] Keras. Chollet, F., & others. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras

[8] XGBoost. Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, 2016

[9] LightGBM. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3149-3157.