**Preprocesamiento de Datasets**

**Dataset del Gimnasio**

* Lo primero fue leer el archivo, ajustar el nombre de las columnas y tipo de datos.
* Después rellenamos algunos valores nulos, que fueron pocos, en la columna de humedad
* Posteriormente pasamos a generar del dataset original, un dataset con resolución de 10 segundos, el cual tenía el promedio de las mediciones con su respectiva desviación estándar y kurtosis.
* Se realizó la estandarización de los datos, y una división de 80-20 para training testing.
* Utilizamos este dataset promediado para aplicar PCA, pero encontramos que para obtener el 95% de la variabilidad explicada eran necesarios utilizar 8 de los 12 features disponibles. En este enfoque algunos de los atributos se consideran fuertemente correlacionados por lo que no se considera la mejor opción para utilizar.
* Se decidió utilizar el dataset promediado a 10seg pero solamente utilizando las medias de las mediciones, descartando 8 de los 12 features. Dichas medias corresponden a temperatura, humedad, presión y altitud. Se tienen fuertes sospechas que la presión y la altitud no deben ser utilizados para la estimación de la ocupación.
* Posteriormente se procedió a generar un dataset balancedo partiendo del dataset de 10seg con PCA. Para el balanceo se utilizó oversampling usando el método de ADASYN.
* Finalmente se generó un nuevo dataset partiendo del original, pero con resolución de 1 minuto sin promedio de muestras. Dicho dataset también fue balanceado usando la técnica de ADASYN, más no se le aplico PCA.
* Se encontró que aunque se obtuvo el mejor resultado con el dataset de 10seg sin procesar, no hubo una reducción significativa en los datasets con PCA, y PCA con Ovesampling. Igualmente en el dataset de 1 muestra por minuto tampoco hubo perdida significativa. La diferencia en accuracy entre el dataset de 10seg pormediado y el dataset de 1 minuto es de 0.00617.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Testing** | | | | |
|  | Avg | PCA | Balanced | 1 minute |
| SVM | 1 | 1 | 0.99671 | 0.981481 |
| KNN | 1 | 1 | 0.99671 | 1 |
| DT | 1 | 0.990131 | 0.993421 | 1 |
| **Mean** | 1 | 0.99671033 | 0.99561367 | 0.993827 |

**Dataset Dressing Room**

* Se reutilizó el script del dataset del gimnasio: se leyó el archivo, se ajustaron nombres de columnas y tipos de datos, se corrigieron los pocos valores nulos.
* Se detectó que en 6k/11k instancias del dataset el campo de fecha estaba corrupto y no pudo ser recuperado.
* El dataset se promedio en conjuntos de 10 muestras sin respetar los cambios de día.
* Se generaron los mismos datasets con PCA, y PCA con Oversampling usando ADASYN.
* No se pudo generar el dataset con 1 muestra por minuto por el impedimento de la columna de fecha.
* Se obtuvieron resultados congruentes con el dataset del gimnasio, es decir, el mejor resultado se obtuvo con el dataset sin procesar, pero los resultados con los otros dos datasets no fueron significativamente menores.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Testing** | | | | |
|  | Avg | PCA | Balanced |  |
| SVM | 0.996721 | 0.996721 | 0.996721 |  |
| KNN | 0.996721 | 0.996721 | 0.993442 |  |
| DT | 0.990163 | 0.986885 | 0.983606 |  |
| **Mean** | 0.994535 | 0.99344233 | 0.99125633 |  |