**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 |  |

Dataset Residencial

Preparación

Primero el dataset se encontraba en un archivo de Excel, con hojas individuales para cada día de medición. Adicionalmente se contaba con una hoja que contiene de manera continua todos los datos respecto a ocupación y ventiladores tomados de las fotos en intervalos de cada 10 min. Igualmente, el formato de las horas se encuentra en UTC por lo que es necesario considerar esto.

Lo primero que se hizo fue generar un archivo CSV con todos los datos de el Excel. En total fueron 11 días de medición que se unieron. Los días no son totalmente consecutivos y se encuentran divididos en 4 periodos. El primero del día 15 de mayo, el segundo periodo del 24 al 29 de mayo, el tercero del 31 de mayo al 1 de junio, y por ultimo el cuarto del 4 al 5 de junio.

La medición durante cada día varía, la medición más larga fue de 24 horas con aproximadamente 70k registros, mientras que la más corta fue de corta fue de apenas 1 hora con 3.5k registros. En promedio las mediciones fueron de 8 horas con aproximadamente 25k registros. Sin embargo, hubo 3 días que la medición se hizo durante varios muestreos en lugar de un muestreo continuo. Esto debido a la disponibilidad del investigador para realizar la recolección de datos.

Después de eliminar datos nulos por errores de muestreo, obtuvimos un dataset final de 295823 instancias. Este dataset se guardó en un archivo CSV con volumen aproximado de 16 mb. Este archivo se llamo dataset-1seg.csv. Posteriormente, generamos 8 datasets más como listamos acontinuación:

* dataset-10sec.csv.- Dataset con resolución de 1 muestra cada 10 segundos.
* dataset-10sec-avg.csv.- Dataset promediado usando intervalos de 10 segundos.
* dataset-30sec.csv.- Dataset con resolución de 1 muestra cada 30 segundos.
* dataset-30sec-avg.csv.- Dataset promediado usando intervalos de 30 segundos.
* dataset-1min.csv.- Dataset con resolución de 1 muestra cada minuto.
* dataset-1min-avg.csv.- Dataset promediado usando intervalos de 1 minuto.
* dataset-5min.csv.- Dataset con resolución de 1 muestra cada 5 minutos.
* dataset-5min-avg.csv.- Dataset promediado usando intervalos de 5 minutos.