Predictor de incendios con detectores de humo

Prieto Julian

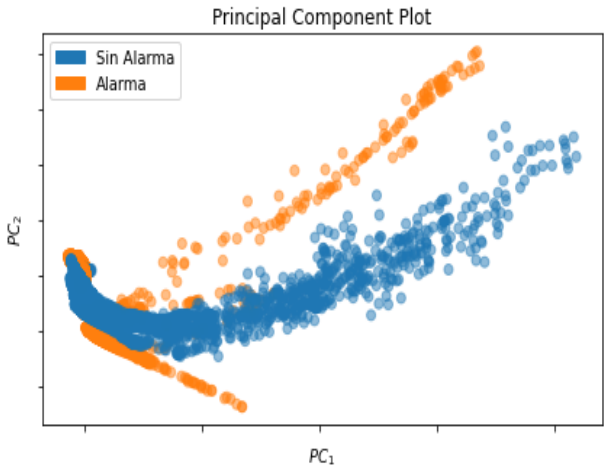
6/11/2022

Introducción

En este informe se detalla la creación de un modelo predictor que pueda detectar incendios, con la utilización de un set de datos con más de 60.000 lecturas de ambientes y fuentes de humo. En cada lectura, se registra la temperatura y la humedad ambiente, los compuestos orgánicos volátiles (medido en partes por mil millones), concentración de CO2, cantidades de hidrógeno molecular bruto y de gas etanol, presión del aire (hPa), tamaño de las partículas y su concentración según su tamaño, el momento exacto en el que se realizó la lectura, y un valor que representa el conteo de los registros.
Todos estos registros fueron tomados en diferentes entornos así como en interiores y exteriores normales, zonas con fuego de leña o con chimenea de gas, parrillas exteriores de leña, carbón y gas, y exteriores con una alta humedad.

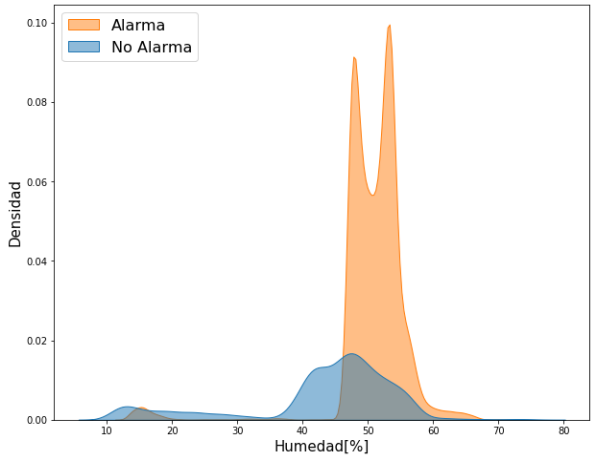
Análisis Exploratorio

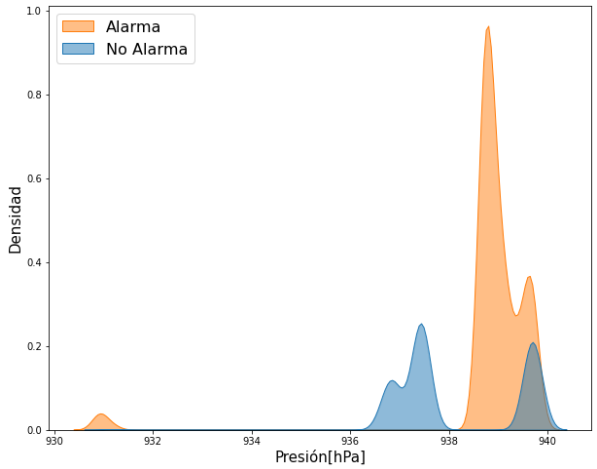
Primeramente, se eliminaron las columnas de UTC (tiempo de la prueba) y CNT (conteo de registros) ya que estas no tienen importancia al momento de clasificar los datos.
Luego, con la intención de poder graficar datos de múltiples dimensiones (+2), se realiza un análisis por componentes principales, el cual intenta mostrar con la mayor fidelidad posible los datos del conjunto calculando componentes principales y ordenándolas por mayor varianza, para así poder ver las variables más influyentes en la clasificación. Este método permite graficar de la mejor manera posible los datos para así poder visualizar la clasificación de estos.
En este caso, el análisis por componentes principales y su respectivo gráfico, logra representar un 65% de la información del set de datos, y como se ve abajo, en este se consigue observar una cierta tendencia seguida por los datos según si el resultado es positivo (Alarma) o si es negativo (Sin alarma).



Se puede observar los casos donde no sonó la alarma marcados por puntos azules, y luego los casos donde sí sonó marcado con color naranja.
Siguiendo con este análisis, es posible detectar qué variables son las más influyentes en cada componente principal. Observando los datos del análisis, se puede ver que en la primera componente principal, las variables más influyentes para la clasificación son tanto el tamaño de las partículas como también su concentración en el ambiente. Y con respecto a la segunda componente principal, las dos variables más importantes son la humedad del ambiente y la presión del aire.

Teniendo esto en cuenta, es posible graficar las densidades de dos de estas variables segun su clasificacion:





Esto nos permite visualizar de manera sencilla de qué forma se distribuyen los valores de ambas variables en casos donde el detector de humo activó su alarma y en los casos donde no lo hizo[1].

Conclusiones del análisis exploratorio

Dado que, más allá de representar gran parte de la información, el gráfico obtenido gracias al análisis por componentes principales no brinda visualmente mucha información por sí solo, lo más relevante es el hecho de poder reconocer las variables más influyentes a la hora de clasificar si hubo o no un incendio. Las cuales, según los datos, son las mediciones de las partículas de materia en el aire como también las mediciones de humedad del ambiente y la presión del aire.

Clasificación Supervisada

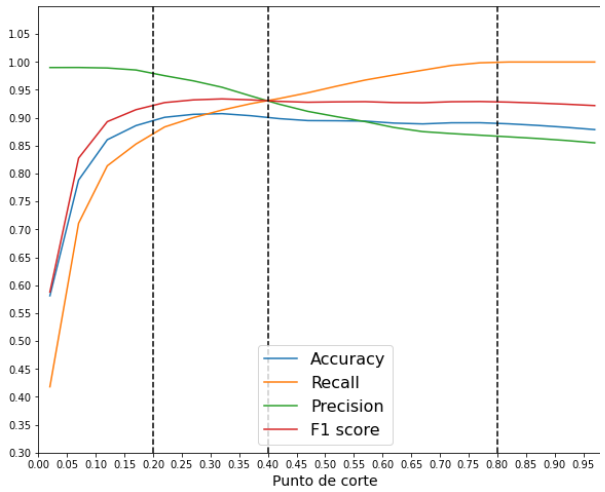
Aclaración:

Para este informe se realizó la tarea de crear un total de 5 modelos predictores distintos, cada uno con un método diferente, entre ellos: Bayes Ingenuo, Discriminante Lineal/Cuadrático y Regresión Logística. Ambos fueron entrenados y probados con los mismos datos en las mismas condiciones y, con la intención de descartar algunos y quedarse con uno solo, se comparó sus rendimientos utilizando el valor dado por el área bajo la curva ROC[2] (Consultar Apéndices Técnicos). Este, arrojó que los modelos con mejores rendimientos son los de Regresión Logística con una puntuación de 0.96, y el modelo de Discriminante Cuadrático con una puntuación de 0.97. Por este motivo, a lo largo de este análisis se habla y se realiza un análisis más extenso y con más métricas sobre ambos modelos.

Regresión Logística

El modelo predictor se diseñó y se entrenó para que por sí solo, pueda clasificar datos. Esto lo hace calculando dos probabilidades: una de que el resultado sea negativo, y otra de que el resultado sea

positivo. De entrada, el modelo utiliza un punto de corte[3] de 0.5, lo que logra unos resultados bastante decentes si hablamos de aciertos, falsos negativos y falsos positivos. Pero, con la intención de poder presentar varios escenarios con diferentes puntos de corte, se realizó un gráfico donde se muestran, con sus respectivos colores, distintas líneas que representan los valores de 4 diferentes métricas: ***Accuracy*** (porcentaje de aciertos), ***Recall*** (habilidad de clasificar correctamente todos los registros positivos), ***Precision*** (habilidad de no generar falsos positivos) y el ***F1-score*** (media armónica entre el ***Recall*** y el ***Precision***), según varios puntos de corte.



En este se marcan 3 escenarios principales que difieren entre sí en su ***Recall*** (Linea naranja) y en su ***Precision*** (Linea verde), mientras que el ***Accuracy*** (Línea azul) y el ***F1-score*** (Línea roja) se mantienen estables en todo momento (A partir del primer escenario).
El primer escenario, con un punto de corte de 0.2, se logra una poca cantidad de falsos positivos, pero sacrificando en gran medida los falsos negativos, ya que estos aumentan de manera considerable. Por otro lado, con un punto de corte de 0.8, se consigue todo lo contrario. Ya que en este caso, la cantidad de falsos negativos decae casi a cero, pero la cantidad de falsos positivos aumenta. Por último, analizando con un punto de corte de 0.4, se consigue equilibrar la cantidad de falsos negativos y falsos positivos, por lo que los valores del ***Recall*** y de ***Precision*** son similares[4].

Coeficientes

Los modelos de Regresión Logística permiten conocer los llamados coeficientes de regresión logistica[5]. Debido a que el modelo mostró que la variable TVOC influye de gran manera al calcular las probabilidades, podemos deducir que esta variable tiene cierta importancia en la tarea de detección de incendios.

Discriminante Cuadrático

El caso de este modelo es totalmente diferente al anterior. Al momento de graficar las métricas según diferentes puntos de corte, se puede visualizar como estas se mantienen estables con valores entre 0.85 y 1.0 para todos los casos. Hablando en forma general, este modelo genera una cantidad mínima de falsos negativos, mientras que la cantidad de falsos positivos se ve elevada.

Conclusiones de la Clasificación Supervisada

Si lo que se busca es un modelo que genere la menor cantidad de falsos negativos al momento de predecir, la opción más recomendable sería el modelo de Discriminante Cuadrático. Mientras que

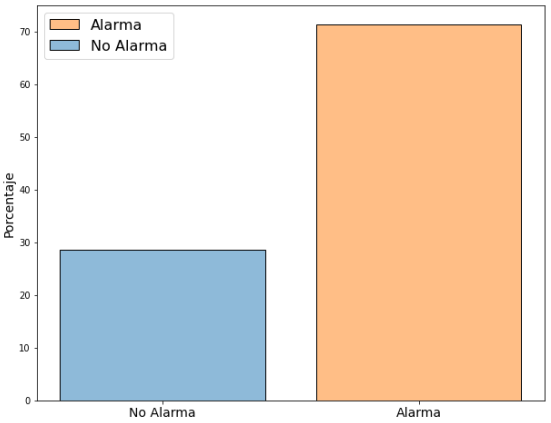
si se desea tener más opciones con respecto a los falsos negativos y falsos positivos, el modelo más flexible es el de Regresión Logística, el cual brinda 3 escenarios que satisfacen distintas necesidades en cuanto a sus métricas de clasificación.

Futuros Trabajos

En futuros trabajos se podría implementar la utilización del método de validación cruzada al momento de evaluar los modelos debido a que es una forma muy utilizada y arroja buenos resultados para la comparación de rendimientos.
Además, para expandir el análisis exploratorio, se podría incluir la matriz de correlación la cual muestra de manera sencilla de entender la importancia de cada variable con respecto a otra, es decir, la correlación entre dos variables. Esto puede permitir entender aún más el set de datos y cómo se relacionan sus variables entre sí.

Apéndices Técnicos

[1]Cabe aclarar que la distribución de los datos clasificados como positivos y como negativos no es pareja, y como se ve en el siguiente gráfico,



está claro que los casos positivos superan en cantidad a los negativos, por lo que es normal que en los gráficos de densidad mostrados en la sección Análisis Exploratorio muestren una gran diferencia en densidades según sus clasificaciones, ya que el 70% de los registros son casos donde se activó la alarma.

[2]La curva ROC(Receiver Operating Characteristic) es una representación gráfica del rendimiento del modelo clasificador que muestra la distribución de las fracciones de verdaderos positivos y falsos positivos. El valor que se toma en cuenta es el de AUC (Area Under the Curve), dónde cuánto más cercano a 1, mejor rendimiento. Gráficamente, la línea que recorre los bordes izquierdo y superior del gráfico representa un ‘test perfecto’, mientras que una línea diagonal con pendiente positiva representa un modelo inutil o con una clasificación totalmente aleatoria.

[3]El llamado punto de corte es el valor que es utilizado al momento de decidir la clasificación según ciertas probabilidades. Este puede ser modificado según las necesidades y los propósitos del modelo creado.

[4] A continuación, se muestra una tabla de métricas para los 3 escenarios mencionados con el modelo de Regresión Logística:

Punto de corte	0.2	0.4	0.8
Accuracy	0.9	0.9	0.89
Recall	0.87	0.93	1.0
Precision	0.98	0.93	0.87
F1-Score	0.92	0.93	0.93

[5]Los coeficientes de regresión logística pueden utilizarse para estimar la razón de probabilidad de cada variable independiente del modelo.

Referencias

Los links consultados para realizar este informe fueron los siguientes:

- 1.9. Naive Bayes — scikit-learn 1.1.3 documentation
- sklearn.metrics.plot_roc_curve — scikit-learn 1.1.3 documentation
- API Reference — scikit-learn 1.1.3 documentation
- 6 Métodos de clasificación | Estadística y Machine Learning con R (bookdown.org)
- seaborn: statistical data visualization — seaborn 0.12.1 documentation (pydata.org)
- Regresión Logística - Documentación de IBM