Universidad Nacional Autónoma de México
Facultad de Ciencias

Proyecto

Modelos No Paramétricos y de Regresión
Semestre 2023 - 2

Ruth Selene Fuentes García
Jorge Iván Reyes Hernández

Alumno: Olvera Trejo Alberto

---

1. Introducción

En este trabajo se exponen diferentes modelos con el objetivo de poder explicar el área quemada en un incendio forestal. La primera aproximación al problema es crear una regresión lineal múltiple y después hacer transformaciones a las variables o a los datos con el objetivo de poder mejorara el desempeño de la regresión. Al final del trabajo se tomo un punto de corte a partir de él se decidió si un incendio es peligroso o no, dependiendo del tamaño de área quemada. Dado que ahora la variable respuesta es dicotómica, entonces tratamos el problema como uno de clasificación, por lo que una aproximación es usar regresión logística. Finalmente, se hace una comparación entre regresión lineal múltiple y regresión logística.

2. Datos

Se tiene un conjunto de datos de 517 observaciones, donde cada una es un vector de 13 entradas. Las variables son las siguientes:

- X - coordenada x dentro del parque Montesinho (1-9)
- Y - coordenada y dentro del parque Montesinho (1-9)
- month - mes del año
- day - día de la semana
- FFMC - índice FFMC (Fine Fuel Moisture Code). Representa la humedad de los combustibles de hojarasca forestal bajo la sombra. Toma valores entre 18.7 y 96.20
- DMC - índice DMC (Duff Moisture Code). Representa la humedad del combustible de la materia orgánica descompuesta debajo de la hojarasca. Toma valores entre 1.1 y 291.3

- DC - índice DC (Moisture Code). Representa el secado en profundidad del suelo. Toma valores entre 7.9 y 860.6
- ISI - índice ISI (Intial Spread Index). Representa la velocidad prevista de propagación del incendio.
- temp - temperatura en grados Celsius. Toma valores entre 2.2 y 33.3
- RH - porcentaje de humedad relativa. Toma valores entre 15 y 100
- wind - velocidad del viento en km/h. Toma valores entre 0.4 y 9.4
- rain - lluvia en mm/m^2. Toma valores entre 0 y 6.4
- area - hectáreas quemadas del bosque. Toma valores entre 0 y 1090.84

Análisis Exploratorio de los Datos

Lo primero fue hacer un histograma de todas las variables, la cual es la figura 1

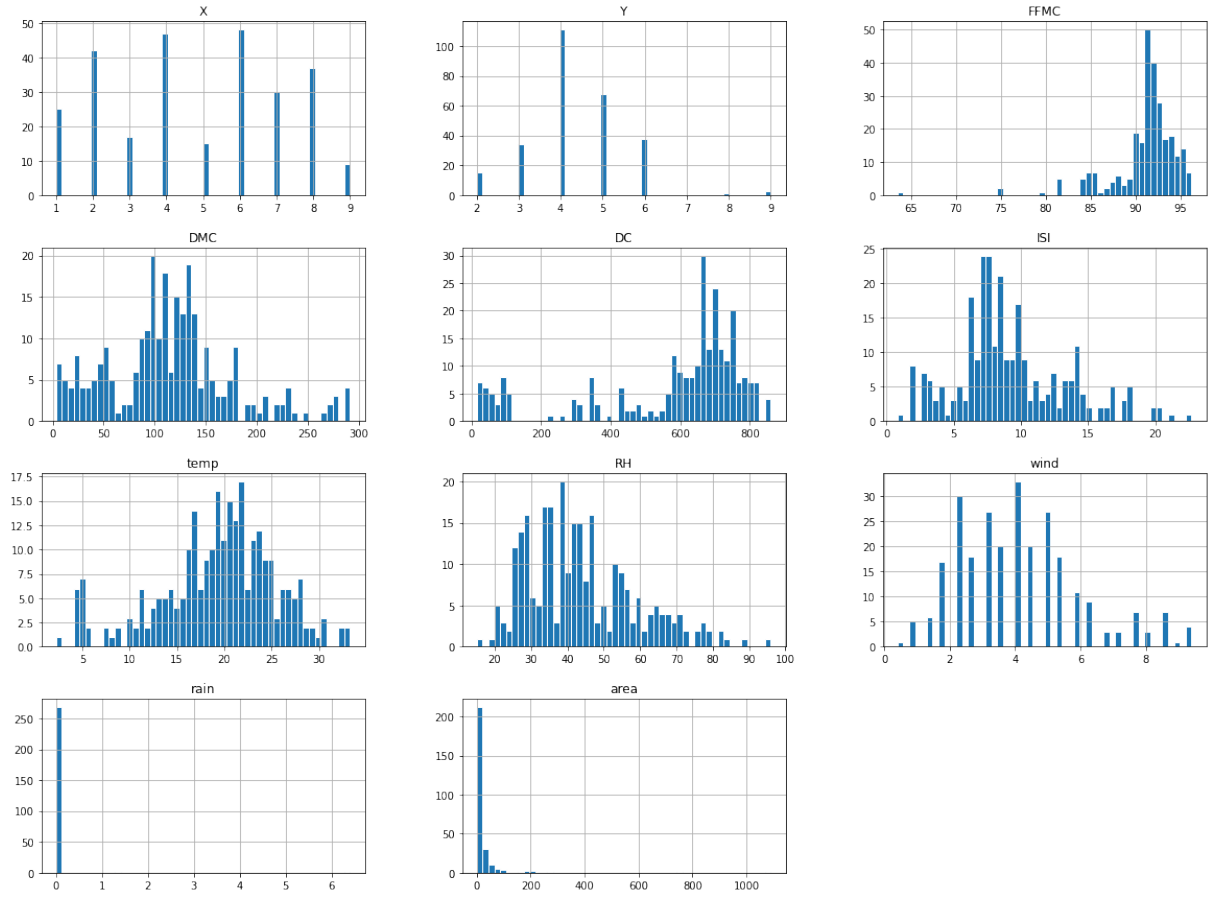

Figura 1: Histograma de todas las variables

El histograma de las variables FFMC e ISI muestra que tienen valores atípicos, por lo que se eliminaron dichas observaciones. Por otro lado, parece que la variable DC es bimodal, mientras que temp y RH tienen colas muy pesadas.

El siguiente paso fue hacer un boxplot (figura 2) y eliminar las observaciones que están a más de una distancia del rango intercuartil de la media.

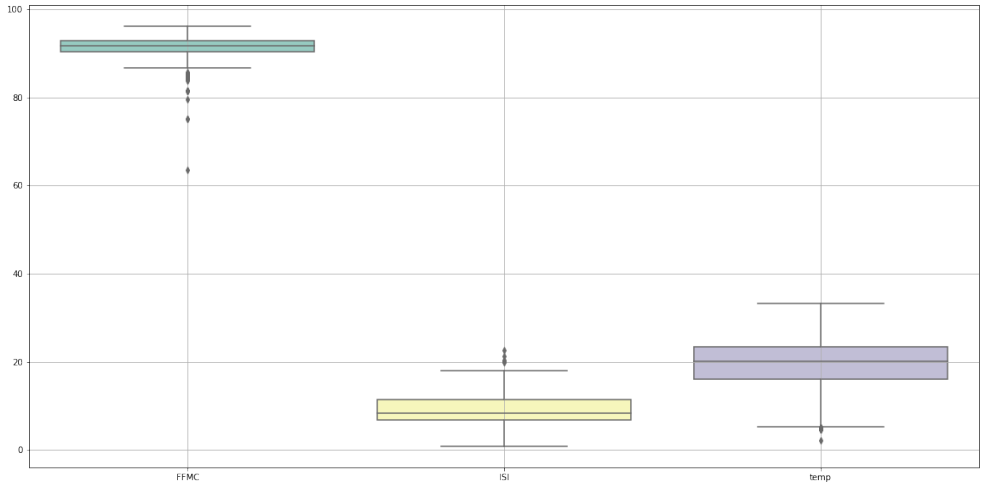

Figura 2: Boxplot de las variables FFMC, ISI y temp

Finalmente, se hizo otro boxplot para el área, el cual sirvió para detectar que también tiene valores atípicos, por lo que se eliminaron las observaciones cuya área quemada es más de 150 hectáreas (figura 3)

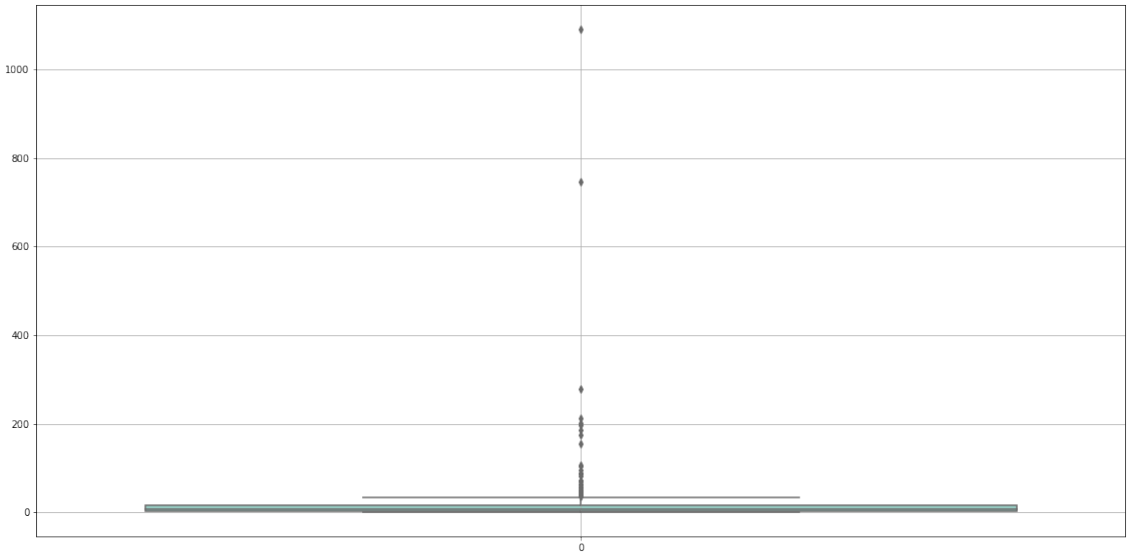

Figura 3: Boxplot de la variable área

3. Modelos

3.1. Modelos de Regresión Lineal Múltiple

Antes de crear el primer modelo hay que considerar que se cuentan con variables categóricas, por lo que hay que transformarlas a variables dummies. Una vez hecho lo anterior se creó el primer modelo, el cual incluyó todas las variables. El resultado obtenido fue el siguiente:

- R^2 0.148
- R^2 ajustada 0.052
- $Prob(F - statistic)$ 0.0561

El valor obtenido de la R^2 ajustada es muy bajo, prácticamente cero, por lo que el siguiente paso es hacer el análisis de residuales para ver si hay algún supuesto que no se cumpla y este causando errores en el modelo.

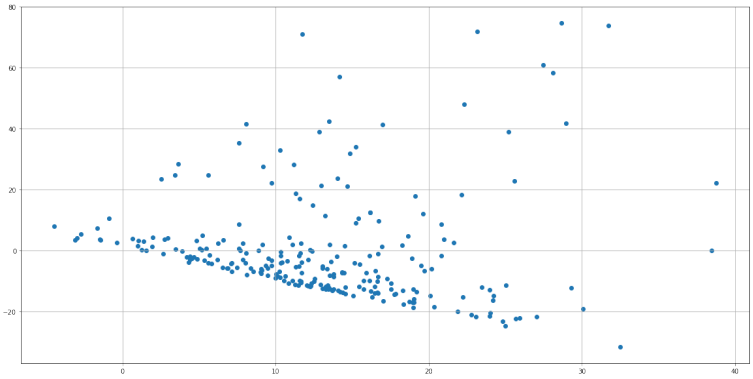

Figura 4: Gráfica de residuales contra el valor estimado

Primero se hizo la verificación del supuesto de la homocedasticidad tomando la prueba de Goldfeld-Quandt, la cual arrojó el resultado de que la varianza es constante, pero por otro lado al hacer la gráfica (figura 4) de los residuales contra el valor estimado, se observó que los puntos no son tan aleatorios como deberían, por lo que sospechamos que la varianza no es constante. Para poder resolver dicho problema se hicieron tres modelos extra los cuales consisten en transformar la variables respuesta con las funciones $y^2, \ln(y)$ y $\sqrt{y}$

Al hacer el gráfico de residuales contra estimaciones obtenemos la gráfica 5 A primera vista parece que

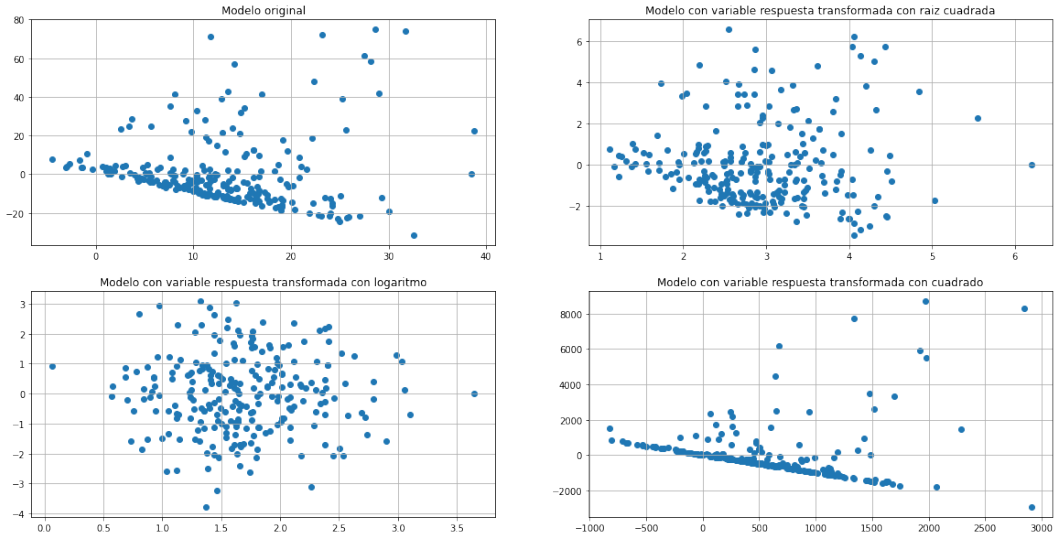

Figura 5: Residuales contra estimaciones de los 4 modelos

los modelos del ln y de la raíz cuadrada tienen varianza constante, pero para verificar hay que hacer la prueba de Goldfeld-Quandt, la cual arrojó los siguientes resultados:

- Modelo original: La varianza es constante
- Modelo $\sqrt{y}$: La varianza es constante
- Modelo $\ln(y)$: La varianza es constante
- Modelo y^2: La varianza no es constante

Se consideran los modelos que pasaron ambas pruebas, los cuales son los de raíz cuadrada y ln.

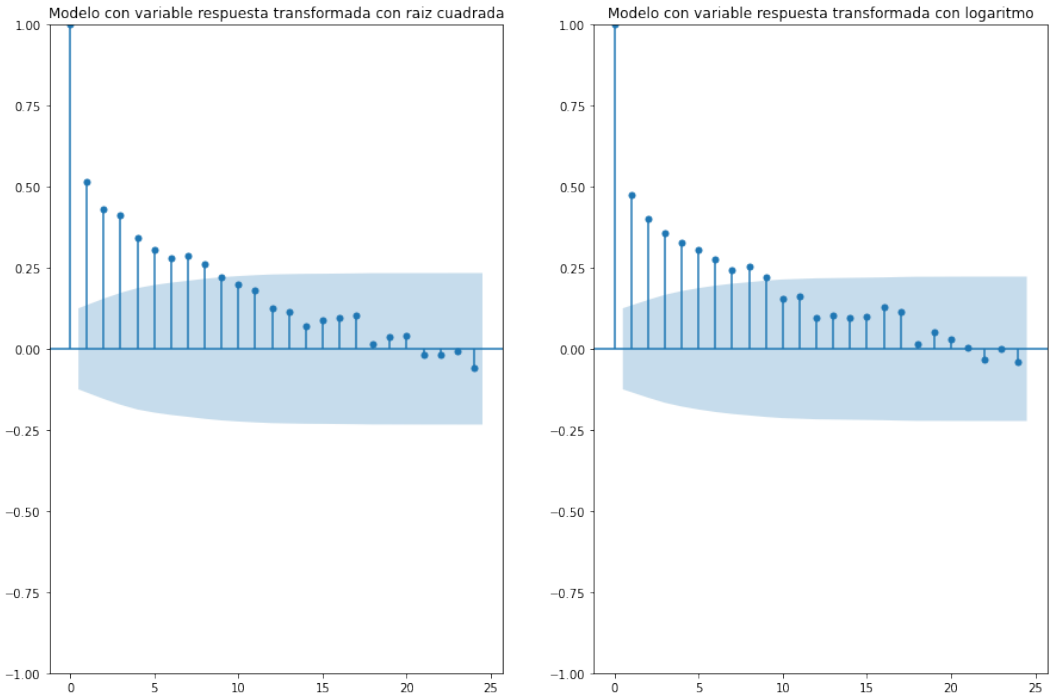

Figura 6: Autocorrelograma

El siguiente paso es verificar que no haya autocorrelación. El autocorrelograma de cada modelo está mostrado en la figura 6. El resultado obtenido es que ambos modelos definitivamente tienen problemas con la autocorrelación.

Por último, resta el supuesto de la normalidad de los residuales el cual se verifica haciendo el qqplot de ambos modelos (figura 7). El resultado obtenido a vista es que el modelo con la variables respuesta $\ln(y)$ es el único que cumple la normalidad de los residuales. Para tener más seguridad de dicha afirmación se realizó la prueba Shapiro, la cual arrojó como reslutado:

- Modelo $\sqrt{y}$: No cumple el supuesto
- Modelo $\ln(y)$: Sí cumple el supuesto

Por lo tanto, el mejor modelo hasta el momento es el que cuenta con la variable respuesta $\ln(y)$, a pesar de que no cumple el supuesto de la no autocorrelación. Hay que recordar que al momento de transformar las variables categóricas (tiempo) a dummies se obtuvieron $7 + 12 = 19$ variables extra, por lo que hay que ver el resultado de la R^2 ajustada, pues dicho valor penaliza el número de regresores empleados.

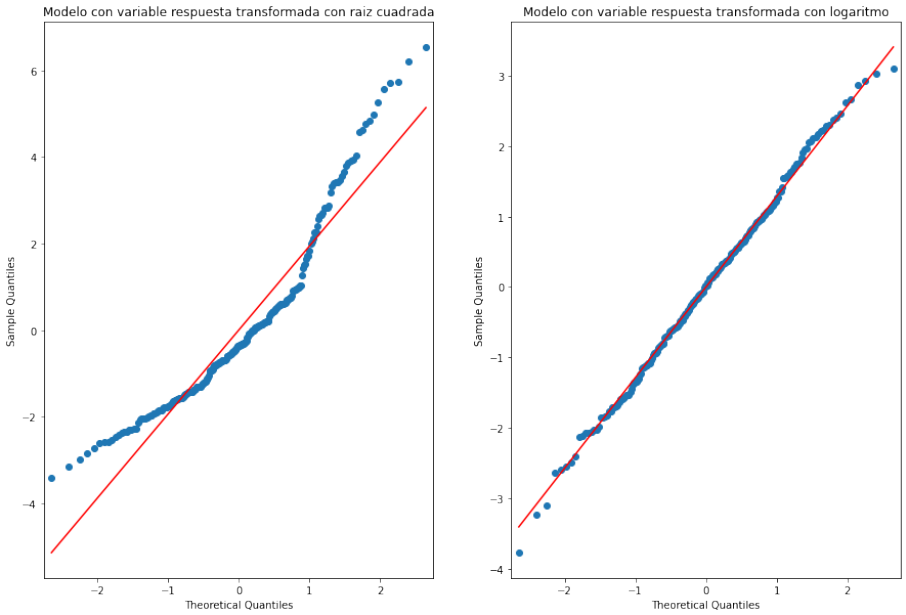

Figura 7: QQplot de ambos modelos

Los resultado de la regresión múltiple son:

1. R^2 0.147
2. R^2 ajustada 0.05
3. $Proba(F-statistic)$ 0.0622

Desafortunadamente éste modelo no es mejor que el anterior. El siguiente paso es hacer backward stepwise selection para eliminar variables e intentar incrementar el valor de R^2 ajustada. Las variables que no pasaron la prueba t (prueba que nos indica si dicha variable es importante en el modelo) fueron: $DCM, DC, mont\_oct, day\_sat$, por lo que se quitaron del modelo. El resultado obtenido fue el siguiente:

1. R^2 0.104
2. R^2 ajustada 0.019
3. $Proba(F-statistic)$ 0.226

El valor de la R^2 ajustada no aumenta, por lo que se procede con otro enfoque.

Recordemos que ningun modelo cumplió el supuesto de la varianza cosntante y una hipótesis es que se estan considerando variables temporales y por ello hay una correlación. Para poder eliminar dicho problema se eliminó la variable del día de la semana y se modificó la variable del mes asignandole un 1 a aquellos meses donde hace calor y un 0 en donde hace frío. Los resultado de dicho modelo son:

1. R^2 0.053

2. R^2 ajustada 0.013

3. $Proba(F - statistic)$ 0.217

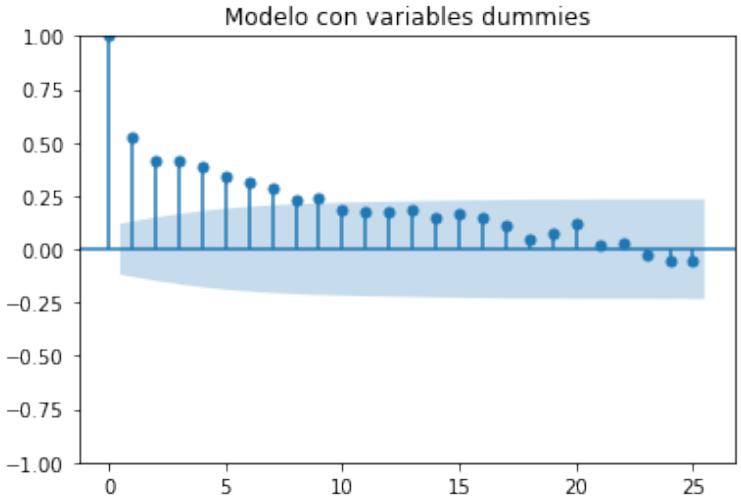

Figura 8: Autocorrelograma modificando el tiempo

Sigue sin aumentar el valor de la R^2 ajustada, por lo que dicha transformación no cumple el objetivo. Por último, el autocorrelograma del modelo (figura 8) muestra que el problema no se solucionó.

Llegando a este punto, y viendo los resultados obtenidos con los modelos creados, podemos afirmar que el área quemada del bosque no puede ser modelada en término de un modelo de regresión lineal múltiple debido a que los resultados son pésimos.

Una última aproximación al modelo es la siguiente: tomar un umbral y a partir de ahí decidir si un incendio es grave o no. Por ejemplo, si tomamos como umbral las 5 hectareas, entonces todos los incendios por abajo de dicho valor serán considerados como no graves (0) y el resto serán considerados como graves (1)

3.2. Modelos de Regresión Logística

Debido a que la variable respuesta ahora toma valores $0, 1$, entonces es un problema de clasificación. Una aproximación es tomar el modelo de regresión logística (notar que el exponente de la formula es el modelo de regresión lineal múltiple)

$$\frac{1}{1 - \exp^{-\left(\beta_0 + \sum_{i=1}^{n} \beta_i X_i\right)}}$$

Se crearan diferentes modelos, cada uno con un diferente umbral y después se procederá a graficar la precisión de cada uno de ellos. Antes de proceder a mostrar los resultados es importante mencionar que un valor muy alto de umbral es muy mala idea ya que si tomamos un umbral de 1000 y únicamente hay 2 incendios que quemaron más de 1000 hectáreas, entonces el modelo nos va a mandar todas las observaciones a la categoría 0 indicando que ningún incendio es grave. Dicho modelo es pésimo debido

a que realmente no está clasificando, únicamente manda todo a una categoría. Sucede lo análogo si consideramos un umbral demasiado chico.

El vector de umbrales considerado es el siguiente $[2.5, 5, 7.5, 10, 12.5, 15, 17.5, 20, 22.5]$

La gráfica de precisión para cada modelo se muestra en la figura 9

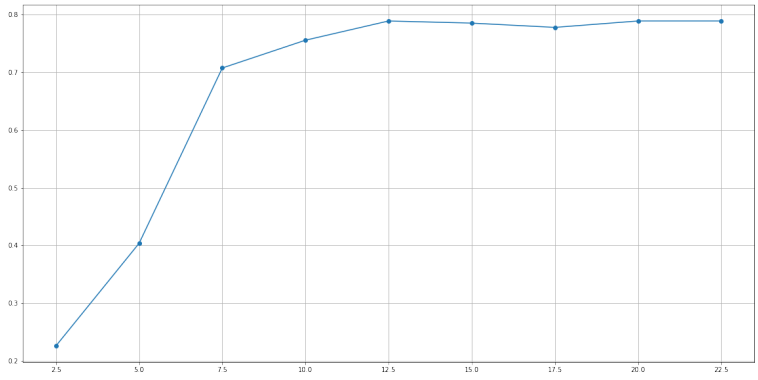

Figura 9: Precisión para cada modelo

Tomaremos el modelo con umbral 7.5 debido a que el siguiente modelo (el de 10) no aumenta mucho la precisión, entonces no vale la pena aumentar tanto el umbral para ganar tan poca precisión.

La matriz de confusión es la siguiente:

$$\begin{pmatrix} 179 & 34 \\ 45 & 12 \end{pmatrix}$$

Por último se hicieron dos modelos, uno únicamente para los meses de calor y otro únicamente para los meses de frío. El método que se siguió fue el mismo el cual es tomar el vector de umbrales y tomar aquel valor de manera que sea el mejor y no sea tan grande.

Para los meses de calor se tomó un umbral de 7.5 y se obtuvo una precisión de 0.75, mientras que para los meses de frio se tomó un umbral del 15 y se obtuvo una precisión de 0.75

4. Conclusiones

Al trabajar con la regresión lineal múltiple hay que tener cuidado con las variables temporales debido a que pueden influir en el supuesto de la correlación de los errores. Una manera de corregirlo es transformar dichas variables a unas dicotómicas. Usando lo anterior y usando la regresión logística, se obtuvieron los mejores modelos, los cuales consten en hacer un modelo para para los meses en donde hacer calor y otro para cuando hace frio, obteniendo una precisión del 0.75.

5. Referencias

P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data.