Búsqueda de la relación entre la producción de plástico a nivel mundial y la educación de las personas

MINERÍA DE DATOS & EL PARADIGMA BIG-DATA

Agustín Jofré Millet
Eduardo Alcober Clemente

ÍNDICE

1- Introducción

Nuestro proyecto surge como respuesta ante la evidente explosión del uso del plástico y su impacto en la vida humana. Intuitivamente, una población mejor preparada intelectualmente, sería más eficiente a la hora de tomar decisiones tanto individuales como en grupo, adaptándose mejor a los cambios sociales en particular y de sus propias vidas en general.

Resulta curioso encontrar cómo cada día salen personas mejor preparadas de las Universidades y centros de formación, y cómo todavía el Ser Humano no ha podido controlar la invasión de residuos, causada por él mismo, en su ecosistema, siendo incapaz de crear una forma de vivir y sostenerse que sea saludable tanto para él mismo, como para su entorno.

2- Objetivos (hipótesis, alcance, restricciones)

La minería de datos es un campo multidisciplinario, que se encuentra en la inserción estadística, aprendizaje automático, administración de base datos y visualización de datos.

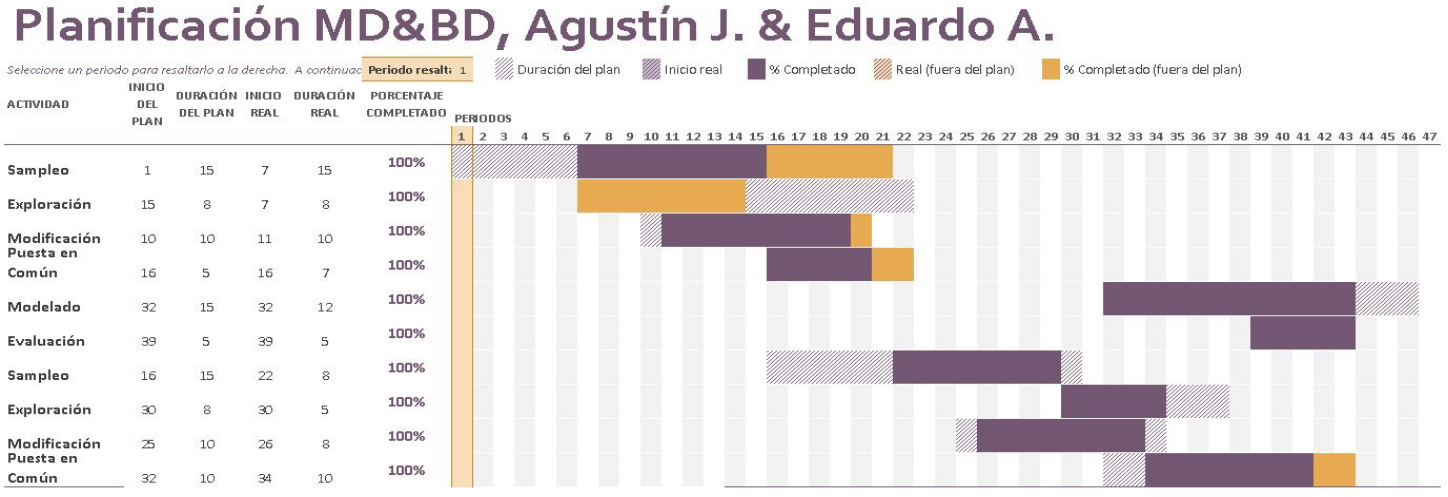
Nuestro objetivo es tratar de encontrar,si existe, una relación entre la educación y la contaminación mundial, basándonos en las cifras de producción de plástico mundial durante mitad del siglo pasado y comienzos del actual.

Vemos la necesidad de crear este estudio estadístico y predictivo, para estudiar la evolución que han vivido los sistemas educativos, respecto a la gestión de residuos. Los últimos informes del PNUMA[1] e ISWA[2], permiten afirmar que se producen entre 7000 y 10000 millones de toneladas de residuos en el mundo.

3- Metodología

Hemos escogido la metodología SEMMA, debido a su flexibilidad de aplicación y a su diferencia con el resto de metodologías estudiadas ya que ésta no tiene en cuenta de la misma forma el *business intelligence*, que no aplica en nuestro caso, y nos permite, etapa tras etapa, ir acondicionando las conclusiones de la anterior, que servirán de entrada a la siguiente, sin ser víctimas de nuestras propias ideas preconcebidas ni prejuicios al respecto: para exponer nuestras conclusiones queremos basarnos únicamente en los datos que hemos conseguido recopilar, tanto a nivel Nacional, como Mundial, para ofrecer determinada perspectiva al respecto y ofrecer una puesta en situación al lector, lo más completa posible, que ilustre nuestros avances.

3.1- Planificación



[1] Programa de las Naciones Unidas para el Ambiente
[2] International Solid Waste Association

Al principio fue difícil encontrar Datasets que permitieran el análisis de nuestra hipótesis, por eso trabajamos en paralelo para aprovechar mejor el tiempo dándonos feedback durante el proceso de búsqueda de datos. Tras la primera puesta en común, donde compartimos nuestros avances individuales y marcamos nuevas direcciones, fijamos los datasets que íbamos a utilizar y continuamos trabajando en paralelo para observar desde los dos puntos de vista, qué resultados tenemos y su relación con nuestra hipótesis. La fase de Modelado y Evaluación la ilustramos al final, ya que llegar a ella nos costó dos iteraciones de las tres fases primeras, terminando de elaborar los resultados del experimento en equipo, dentro de los plazos previstos. Aunque hay etapas que han comenzado antes de lo que teníamos programado, y otras que se han extendido más, hemos cumplido bastante bien con los objetivos que nos hemos ido fijando. Tampoco hemos dejado que el calendario mermarse nuestro avance, dejando que fueran los resultados de nuestros análisis proyecto los que marcaran el ritmo.

4- Implementación

El camino que hemos seguido en la investigación, según lo marca la metodología tradicional SEMMA, consiste de los siguientes puntos de forma secuencial:

- Sample
    - Durante esta etapa hemos buscado toda la información posible al respecto. No ha sido fácil debido a la diversidad de fuentes consultadas y formatos fuente. Las principales fuentes que hemos utilizado por ofrecer información relevante al respecto de nuestro estudio han sido:
        - https://ourworldindata.org/
            - Our World in Data is focussing on the powerful changes that reshape our world
        - http://www.esrl.noaa.gov/gmd/ccgg/trends/index.html
            - Trends in Atmospheric Carbon Dioxide, Mauna Loa, Hawaii
        - https://www.madrid.org
            - Tanto en el apartado de *estadísticas* como en otros subapartados de la página, hemos encontrado información actual importante que procedemos a presentar.
- Explore
    - Exploración de los datos.
- Modify
    - Modificación de los mismos y traducción a estructuras de datos manejables por R.
- Model
    - Elaboración del modelo de regresión
- Assess
    - Evaluación del modelo, comprobación de los errores y de la precisión de los modelos elaborados.

Hemos partido de dos dataFrames , por un lado tenemos el número medio de años que pasa una persona escolarizada por país y año y por otro la producción de plástico global, en millones de toneladas generadas por año (fig.2).

Analizando la tabla de población escolarizada por países y año , descartamos la columna de países y creamos una nueva tabla (fig 1) ordenada por años y calculamos de media de tiempo escolarizado. usando la función aggregate.data.frame(..).

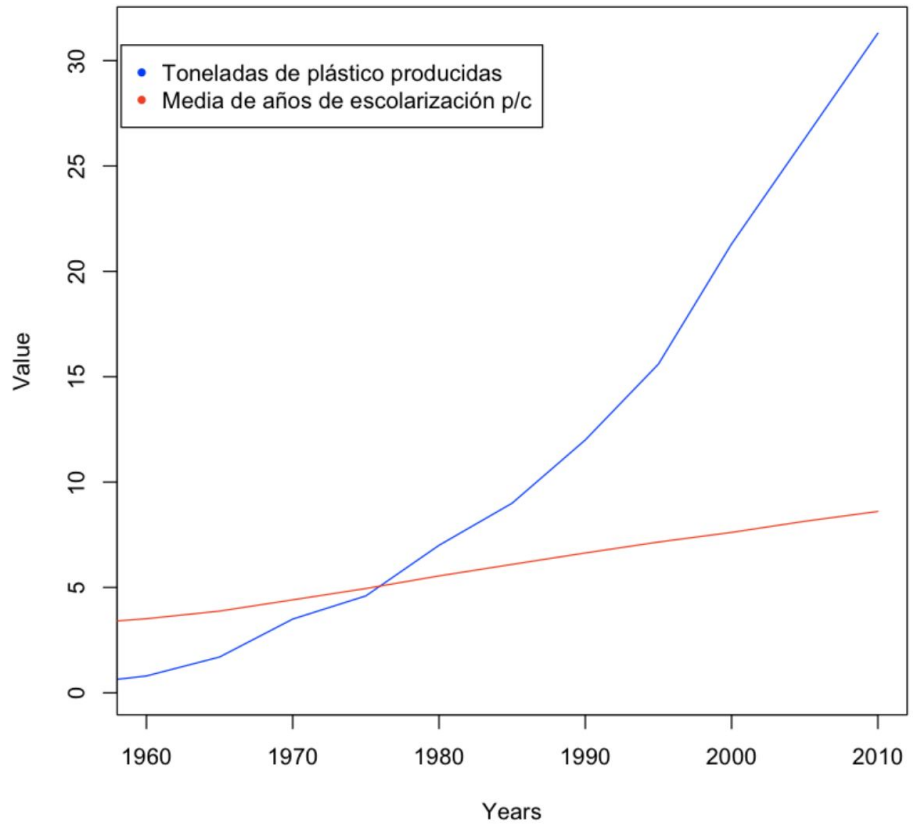
Tras analizar ambas tablas, procedemos a unirlas de forma ordenada por año. Se pierden datos, la tabla de escolarización va desde el año 1870 al 2010, de 5 en 5 años y la de producción de plástico desde 1950 a 2015, por lo que nos quedamos con datos desde 1950 a 2010 en ambas tablas.

Year	mean escolarizados
1870	0.5463964
1875	0.5912613
1880	0.6511712
1885	0.7189189
1890	0.7927027
1895	0.8721622
1900	1.0216216
1905	1.1727928
1910	1.3368468
1915	1.5254054

fig. 1.

Year	Prod de plastico global
1950	2000000
1951	2000000
1952	2000000
1953	3000000
1954	3000000
1955	4000000
1956	5000000
1957	5000000
1958	6000000
1959	7000000

fig. 2.



Para la unión de la tablas usamos la función merge, que combina dos data frame por columnas , ordenados por año y calculando la media de los valores, el resultado se muestra el la figura 3.

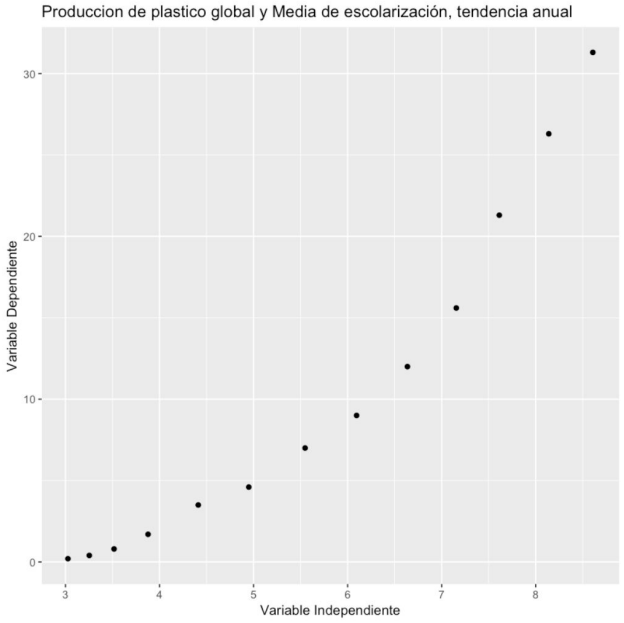
Year	Prod de plastico global	mean escolarizados
1950	0.2	3.026036
1955	0.4	3.253243
1960	0.8	3.516486
1965	1.7	3.878198
1970	3.5	4.411892
1975	4.6	4.949279
1980	7.0	5.547838
1985	9.0	6.095315
1990	12.0	6.636126
1995	15.6	7.155495
2000	21.3	7.612883
2005	26.3	8.139279
2010	31.3	8.607477

Figura 3.

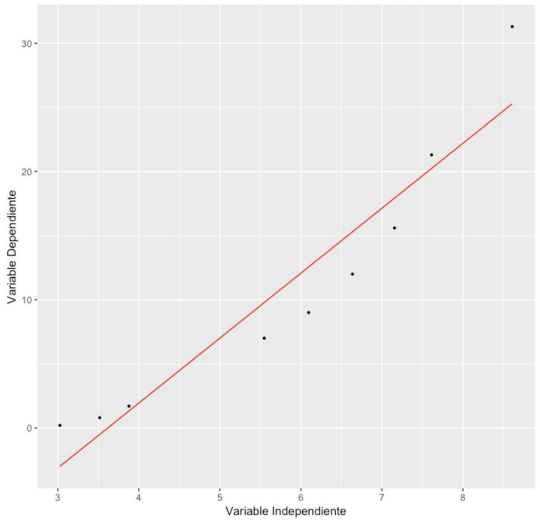
Regresión

Una vez tenemos los datos preparados, vamos a proceder a aplicar el modelo de regresión lineal , multivariable (regresión polinomial). Lo primero que hacemos es renombrar las columnas , la ‘x’ es la media de escolarizados, nuestra variable independiente y la ‘y’, producción de plástico global , nuestra variable dependiente.

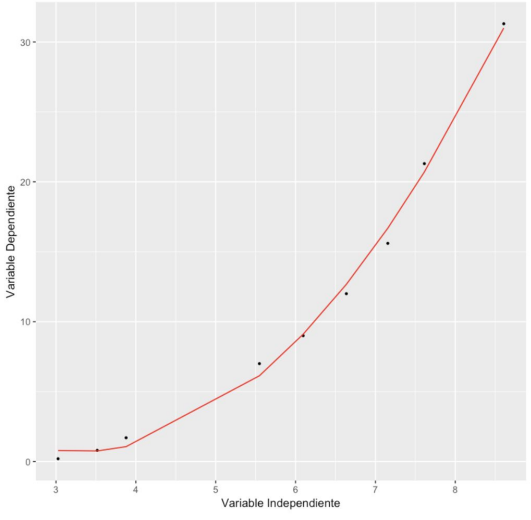
Obtenemos la siguiente gráfica con el total de datos.



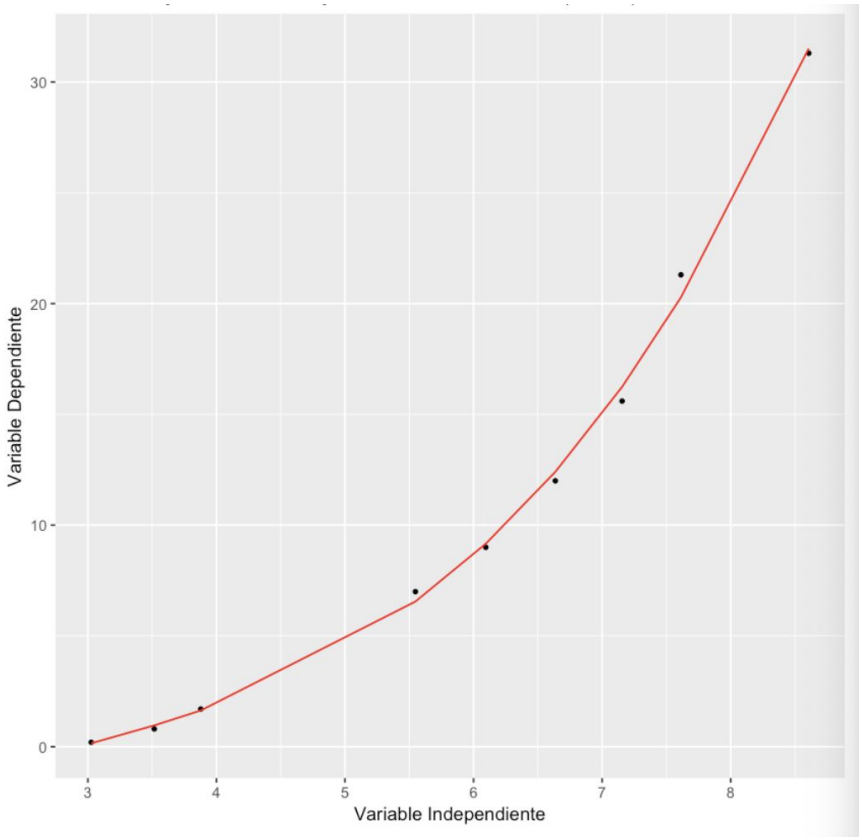
- Nuestra idea es predecir el impacto del desarrollo humano sobre la producción de plásticos.
- Dividimos, como es común, en proporción 70-30% los datos de entrenamiento y prueba respectivamente, utilizando la función split de la librería caTools. y procedemos a entrenar el modelo para tres modelos diferentes regresión lineal y multivariable . Las gráficas obtenidas son las siguientes.



Gráfica regresión lineal simple



Gráfica regresión lineal multivariable (de 2º grado)



Gráfica regresión lineal multivariable (de 3º grado)

- A primera vista la nube de puntos que forman los samples del conjunto de entrenamiento es una parábola. No obstante vamos a tener en cuenta la medida del error a la hora de aplicar regresión lineal para compararlo con los otros modelos construidos

	Regresión Lineal	***Regresión polinomial de segundo grado***	***Regresión polinomial de tercer grado***
coeficiente de determinación R-squared	**0.8952**	**0.9952**	**0.9989**
P-value	**9.308e-05**	**7.067e-08**	**4.704e-07**
Error residual	**3.507**	**0.7732**	**0.62**

5- Resultado y evaluación.

A primera vista con solo ver los resultados obtenidos, descartamos aplicar el modelo de regresión lineal. Por lo que a partir de ahora solo nos centraremos en los modelos multivariable.

El coeficiente R cuadrado alcanza su valor máximo al añadirle la característica x^3 a la regresión lineal de 2º Grado y, como era de esperar, éste es el mejor ajuste que hemos podido obtener. Podemos decir que ***el modelo estima con alta precisión***. Hemos usado esta medida para comparar, aunque se puede observar en los *summary's* de las tres regresiones cómo el *error standard residual,* así como el *R-squared ajustado* siguen la misma progresión de acercamiento al ajuste óptimo en x^3.

En cuanto al P-value, este está determinado por el valor de *F-statistic* y es la probabilidad de que nuestros resultados hayan ocurrido por casualidad. Como se puede apreciar en la tabla la probabilidad es bastante baja.

El valor residual se define como la diferencia entre el verdadero valor de la variable dependiente y el valor predicho por el modelo. Cuanto más pequeño sea el valor de los residuos mejor será el ajuste del modelo a los datos y más acertadas serán nuestras predicciones.

Se ha realizado una validación simple , pintando la matriz o tabla de confusión de ambos modelos .Como se puede ver en las tablas de abajo el modelo de regresión de 2º grado ha predicho dos valores mal, no en cambio de polinomio de 3º grado .

y_poly_predict	0.2	0.8	1.7	7	9	12	15.6	21.3	31.3
0.758411673891553	0	1	0	0	0	0	0	0	0
0.79067670835423	1	0	0	0	0	0	0	0	0
1.06620204504408	0	0	1	0	0	0	0	0	0
6.13548586043936	0	0	0	1	0	0	0	0	0
9.10348127132576	0	0	0	0	1	0	0	0	0
12.6684768351055	0	0	0	0	0	1	0	0	0
16.6844667937822	0	0	0	0	0	0	1	0	0
20.7017833061681	0	0	0	0	0	0	0	1	0
30.9910155058893	0	0	0	0	0	0	0	0	1

matriz de confusión con polinomio de 2º grado

y_poly_predict	0.2	0.8	1.7	7	9	12	15.6	21.3	31.3
0.153348935204709	1	0	0	0	0	0	0	0	0
0.960993249852532	0	1	0	0	0	0	0	0	0
1.63433652632677	0	0	1	0	0	0	0	0	0
6.55240894223326	0	0	0	1	0	0	0	0	0
9.16233623823936	0	0	0	0	1	0	0	0	0
12.4052909951512	0	0	0	0	0	1	0	0	0
16.2404769639676	0	0	0	0	0	0	1	0	0
20.2785047360548	0	0	0	0	0	0	0	1	0
31.5123034129698	0	0	0	0	0	0	0	0	1

matriz de confusión con polinomio de 3º grado

Por último, aunque vemos que el modelo se comporta correctamente, vamos a proceder a comprobar los resultados con el porcentaje de datos que nos dejamos para validación.

Agregamos nuevas características al modelo, en este caso 2, de segundo y tercer grado, con vistas a aumentar su precisión, calculamos el valor de predicción, utilizando los valores entrenados.

Como resultado vemos que se predicen de forma acertada.

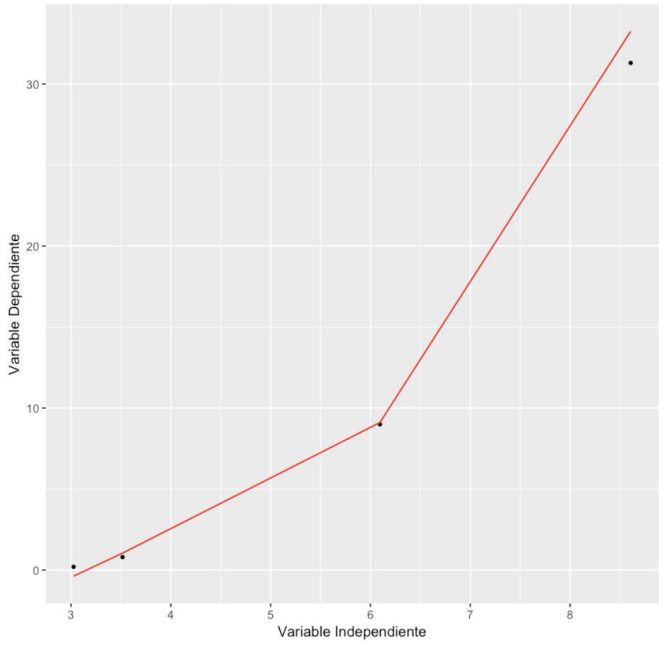


Gráfico y predicción con datos de validación

y_poly_test_predict	0.4	3.5	4.6	26.3
0.51659715806332	1	0	0	0
2.82223870350523	0	1	0	0
4.34460506488376	0	0	1	0
25.7783588835262	0	0	0	1

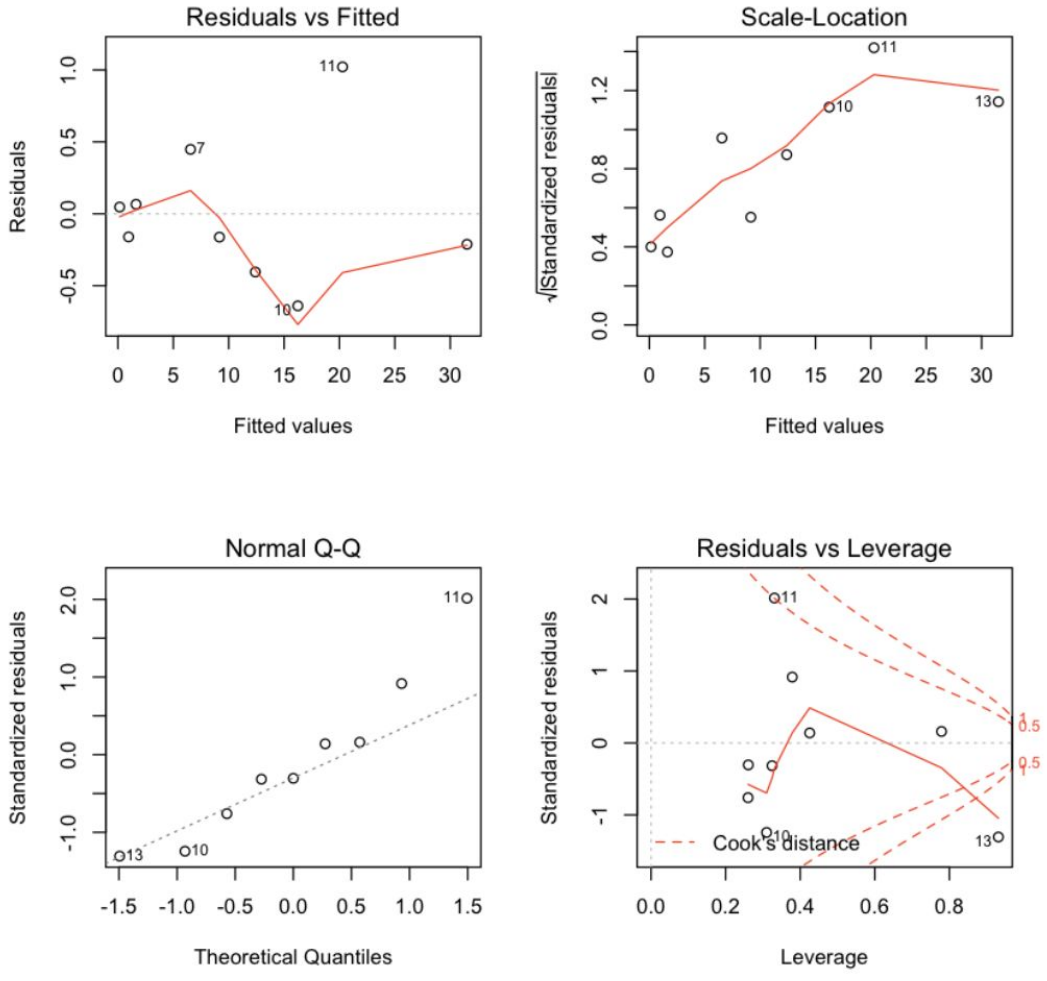
Matriz confusión datos de validación

6- Conclusiones .

Como se ha visto en los apartados anterior la matriz de confusión del modelo de 3º grado es perfecta. Por eso nuestra conclusión respecto a que modelo es mejor, es el de regresión polinomial de grado 3.

y_poly_predict	0.2	0.8	1.7	7	9	12	15.6	21.3	31.3
0.153348935204709	1	0	0	0	0	0	0	0	0
0.960993249852532	0	1	0	0	0	0	0	0	0
1.63433652632677	0	0	1	0	0	0	0	0	0
6.55240894223326	0	0	0	1	0	0	0	0	0
9.16233623823936	0	0	0	0	1	0	0	0	0
12.4052909951512	0	0	0	0	0	1	0	0	0
16.2404769639676	0	0	0	0	0	0	1	0	0
20.2785047360548	0	0	0	0	0	0	0	1	0
31.5123034129698	0	0	0	0	0	0	0	0	1

Con este modelo hemos obtenido un coeficiente de determinación R-squared de 0.9989, valores residuales bajos de 0.62, y un p-value muy bajo.



La otra conclusión obtenida y esta es respecto a los resultados del modelo elegido es que predice de forma acertada. La idea que se obtiene directamente al predecir un nuevo dato mayor a lo del eje de la variable independiente es que , la producción de basura aumenta considerablemente.

Se podría decir que a mayor escolarización más plástico se produce. Mirando los datos más a fondo, llegamos a la conclusión de que este aumento en la escolarización se debe a un aumento poblacional . Desde 1950 la población mundial ha seguido creciendo ,

incluso con una baja natalidad en algunos países, por que a esto se ha unido el aumento de la esperanza de vida.

Con ello lo que queremos decir es que a mayor población, mayor número de escolarizados, a la vez que con los años ha ido paulatinamente aumentando el número de años que un adulto pasa “estudiando”; cuando estas dos variables crecen a la vez, los recursos materiales que se utilizan para estudiar se disparan. Estos materiales básicos para el estudio, van desde lápices bolígrafos, libros, gomas de borrar, barra de pegamento, reglas, calculadoras, papel, tijeras, etc. Todos estos materiales, algunos más que otros son plásticos o vienen envueltos en plástico, y de aquí que la producción de plástico mundial aumente, aunque una vez encontrada la relación entre las variables, esto queda fuera del ámbito de nuestro estudio.

7- Bibliografía

- https://ourworldindata.org/ Our World in Data is focussing on the powerful changes that reshape our world
- http://www.esrl.noaa.gov/gmd/ccgg/trends/index.html Trends in Atmospheric Carbon Dioxide, Mauna Loa, Hawaii
- https://www.madrid.org
- https://medium.com/datos-y-ciencia/introducción-a-los-modelos-de-regresión-en-r-6ef5a4c47a8f
- http://www.ccg.unam.mx/~vinuesa/R4biosciences/docs/Tema9_regresion.html