Reciclaje en la Ciudad de Buenos Aires

Denise Martin Caldareri, Emiliano Nicolas Di Bennardo y Leonardo Maestri
Universidad Tecnológica Nacional, Buenos Aires, Argentina

Abstracto
En este reporte se analizará la evolución del reciclaje por comuna dentro de la Ciudad Autónoma de Buenos Aires a lo largo de los años 2015 y 2016. También se generará un modelo predictivo de tipo clasificador para identificar las comunas de la Ciudad Autónoma de Buenos Aires a partir de sus pesajes semanales. El informe se desarrollará en siete etapas: introducción, descripción del dataset, análisis exploratorio de datos, material y métodos, resultados, discusión y conclusiones, referencias.

Palabras Clave
Reciclaje, puntos verde, Buenos Aires, medio ambiente, comunas

1 Introducción

En nuestro trabajo realizaremos un análisis sobre los puntos verdes que se encuentran en la Ciudad Autónoma de Buenos Aires, comprendido dentro del periodo 2015 y 2016, donde buscaremos conclusiones que nos ayuden en un futuro a incentivar a elevar la cantidad de kilogramos reciclados, cumpliendo con los objetivos planteados en la Ley 1.854 "Basura cero"[3] promulgada en enero 2006 respecto a la gestión de los residuos sólidos urbanos, orientada a la eliminación progresiva de los rellenos sanitarios, con medidas dirigidas a la reducción de generación de residuos, la recuperación y el reciclado.

2 Descripción del dataset

El dataset que utilizaremos para realizar este trabajo fue extraído de Data Buenos Aires, en la rama de medioambiente, donde se analiza el pesaje semanal de distintos materiales reciclables que fue recibido en los 32 puntos verdes de la Ciudad Autónoma de Buenos Aires entre 2015 y 2016.

La información que contiene este dataset se encuentra a cargo de los empleados que trabajan en cada punto verde, donde se les solicita a los mismos que completen a través de un programa la cantidad de vidrio, papel, cartón, metal, plástico, telgopor y tetrabrik para los 32 puntos verdes. Dentro de estos 32 puntos verdes, se encuentran dos puntos verdes especiales, que además de recolectar todo lo mencionado anteriormente, se encargan también del reciclado de electrodomésticos, aparatos de informática y aceite vegetal usado. Estos tres últimos puntos no serán tenidos en cuenta en nuestro análisis para poder darle al trabajo un análisis integral y equitativo de todos los puntos verdes.

El dataset elegido nos obliga a hacer una limpieza exhaustiva del mismo, este posee una proporción considerable de datos completados con valores nulos, cadenas vacías, cadenas con espacios y filas con ceros en cada uno de sus registros. Este complicación se debe a que el dataset es alimentado por distintas organizaciones en distintas comunas, las cuales no tienen un acuerdo de estandarización para completar los datos.

La limpieza del archivo nos permitió observar que los últimos tres meses de 2016 se encuentran sin datos, por lo que nuestro análisis en el trabajo estará comprendido entre enero 2015 y septiembre 2016. El sistema utilizado para la carga de documentación generó también en varios meses una quinta semana con pocos días para completar los 30 o 31 días del mes, esto llevó a que el archivo posea muchos datos vacíos en estas semanas que no fueron cargadas por la mayoría de las organizaciones de las comunas, por lo que procederemos también a eliminar estas quintas semanas vacías para poder optimizar el estudio.

Optamos por ejecutar un criterio de selección para los puntos verdes ya que algunos de estos tenían en su mayoría valores inutilizables, ya sea porque no estaban cargados o porque estaban presentes pero vacíos. El criterio elegido es el siguiente:

- Semana rellena con 0, "", " ", NaN, considerada vacía.
- Mes con 3 o más semanas vacías, considerado vacío.
- Mes no cargado en el dataframe, considerado vacío.
- Punto con 4 o más meses vacíos en 2015 y 2016, considerado vacío.

Finalmente se eliminan los puntos verdes considerados vacíos según el criterio, lo que nos dejó 28 puntos.

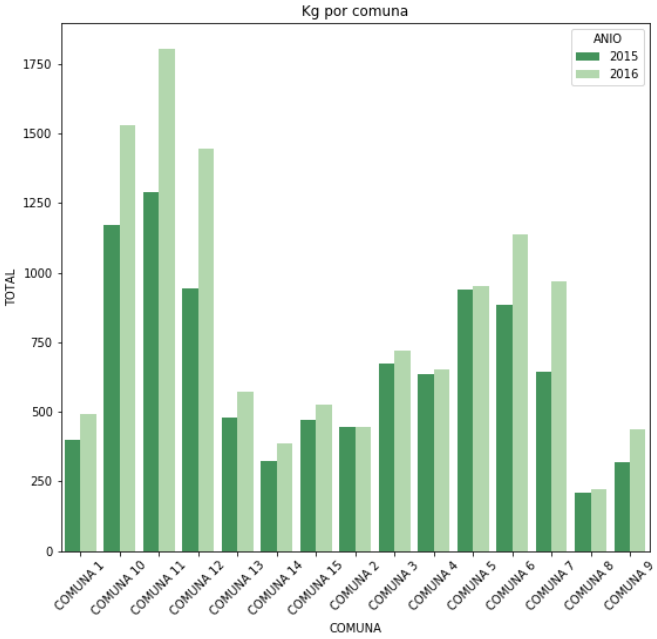
Los puntos verdes que pasaron la selección fueron tratados para completar sus semanas vacías mediante las medias de las semanas no nulas del mes al que pertenecen y de no ser posible mediante la media calculada entre el mes anterior y el siguiente.

3 Análisis exploratorio de datos

3.1 Pesajes totales 2015 y 2016

Comenzamos el EDA visualizando y comparando los kilogramos totales reciclados por comuna durante 2015 y

2016. En este caso, se recurre a un gráfico de barras, a través del cual se observa que las comunas con mayor volumen de reciclaje son: la número 11, 10 y 12. Por otro lado, las comunas con menores índices de reciclado son: la número 8, 9 y 14.

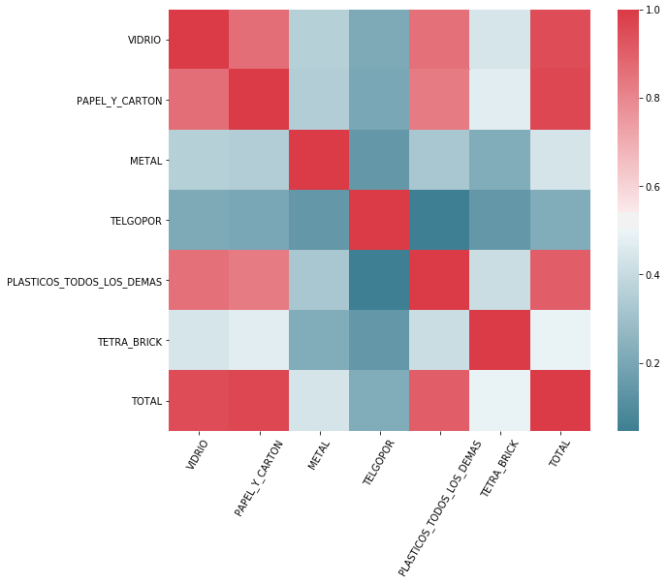


3.2 Test de Pearson entre las variables

El test de Pearson es un indicador que varía entre -1 y 1. Indica si dos variables supuestas independientes, se comportan linealmente una de la otra al analizalas juntas. Una correlación r > 0 implica correlación lineal positiva y r < 0 implica una correlación lineal negativa. Si r = 0 quiere decir que no existe correlación lineal, aunque puede existir otro tipo de correlación no lineal. El método sigue la fórmula:

$$r = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\left[\sum_i^n (x_i - \overline{x})^2 (y_i - \overline{y})^2\right]^{1/2}}$$

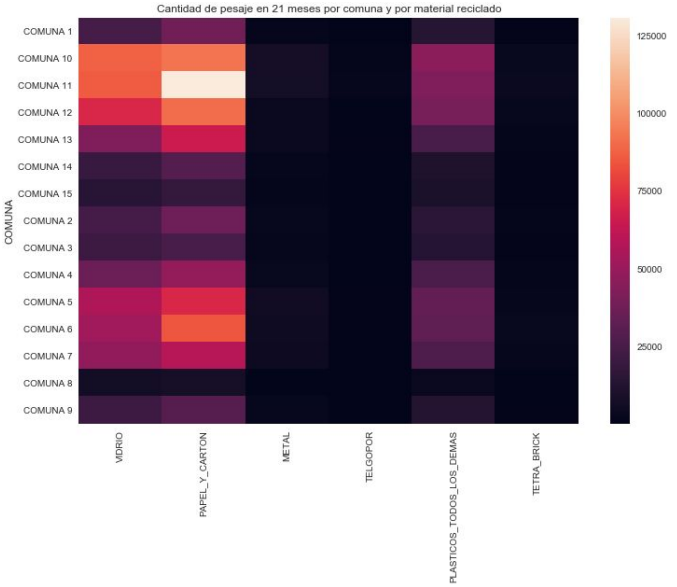
Visualizamos sus resultados mediante un gráfico Heatmap.



Analizando la matriz de correlatividad, se evidencia que los materiales reciclados que más influyen en el total son vidrios, plásticos, papel y cartón. Además, estos se encuentran fuertemente relacionados unos con otros. El resto de materiales tienen pocos aportes al total. En caso de telgopor y tetra brick se puede deber a que no todos los puntos verdes reciben estos materiales.

3.3 Análisis de materiales totales por comuna

En este punto del EDA visualizamos la cantidad total de pesaje de cada material a lo largo de 2015 y 2016. Buscamos saber cuales son los materiales más reciclados y en que comunas. Utilizamos nuevamente un Heatmap.

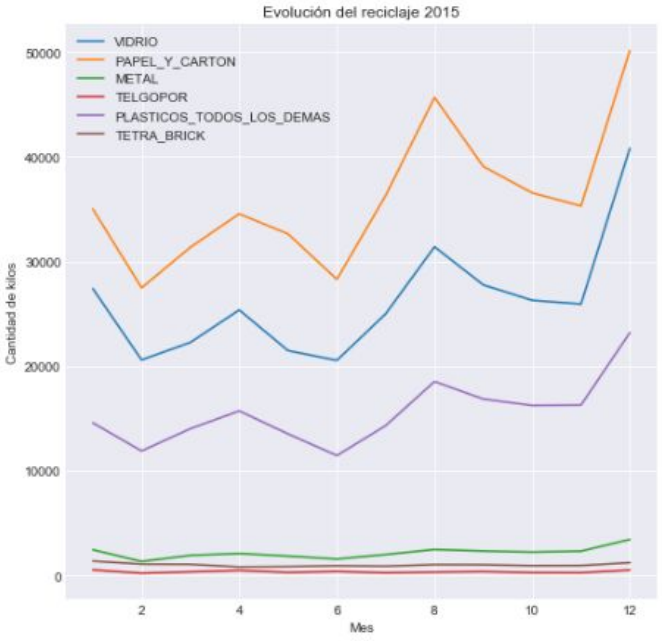


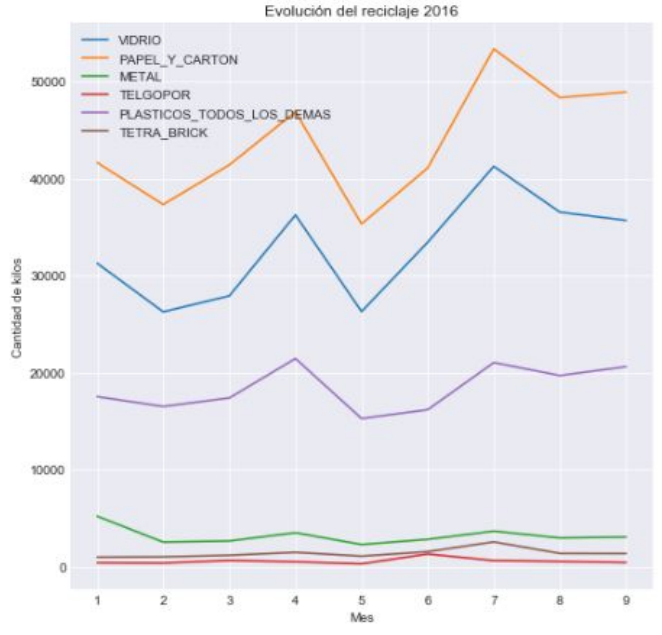
En este gráfico podemos ver resultados que se corresponden con el Test de Pearson hecho ya que vemos como vidrios, papeles y cartones son los más reciclados y por ende los que más aportan al total. También confirmamos otra vez como en el punto 3.1 que las comunas 10, 11 y 12 son las de mayores volúmenes y las 8, 9 y 14 las de menores

volúmenes, estas se ven más oscuras siguiendo la escala en nuestro Heatmap.

3.4 Análisis de estacionalidad

Se construyen dos series de tiempo, una por cada año, a fin de analizar la evolución de los kilogramos totales reciclados por material y ver si se puede identificar una tendencia general y/o estacionalidad, es decir valores mayores en ciertos meses del año.





A partir de los siguientes gráficos se puede determinar que el reciclaje de los distintos materiales presenta una tendencia creciente desde 2015 a 2016, con puntos máximos en los meses de abril y diciembre así como también durante el período comprendido entre julio y agosto.

La diferencia entre los meses de noviembre y diciembre se podría explicar por la incidencia que tienen las fiesta de fin de año y el inicio del verano. Durante diciembre las personas suelen dar y recibir regalos empaquetados en papel, cartón y plásticos además de aumentar su consumo de bebidas alcohólicas embotelladas en vidrio. Estos fenómenos implicaría un aumento en la utilización de envases de vidrio y envoltorios.

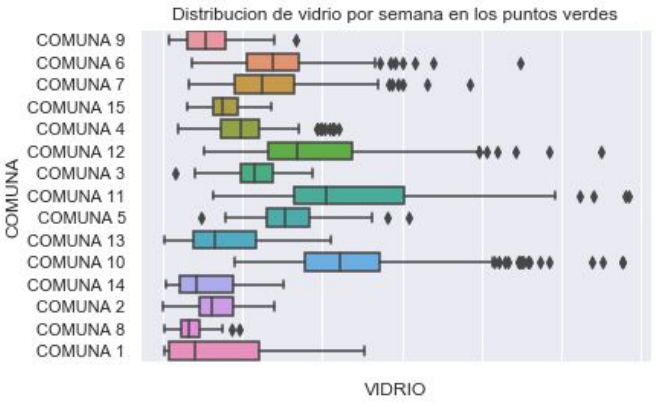
Durante los meses de enero, febrero y marzo, hay una disminución en el volumen total de materiales reciclados con un repunte en el mes de abril. Esto se puede deber a que los habitantes de la ciudad suelen migrar hacia otros puntos del país con el objetivo de disfrutar de sus vacaciones.

Durante los meses de mayo y junio hay una disminución de los kilogramos de material reciclado con un notable incremento posterior, en los meses de julio y agosto.

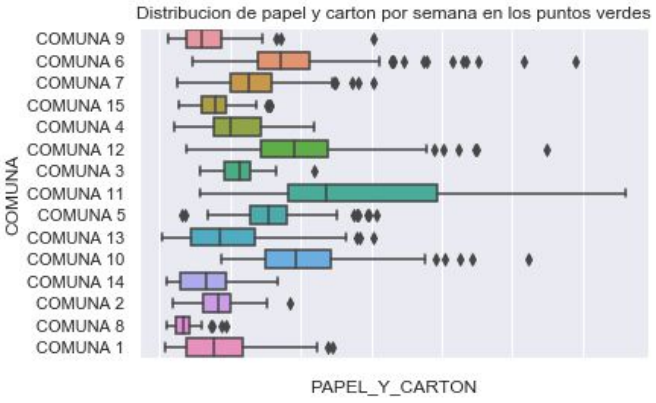
3.5 Análisis de cada material

Utilizando gráficos de tipo Boxplot buscamos ver la cantidad de outliers (anomalías) en los pesajes de cada semanas en cada comuna.

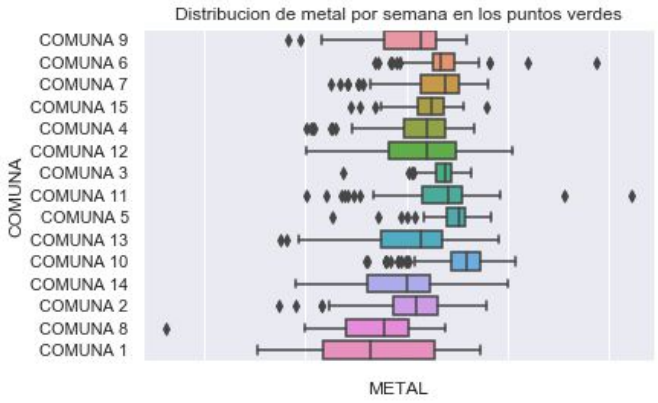
3.6.1 Vidrio



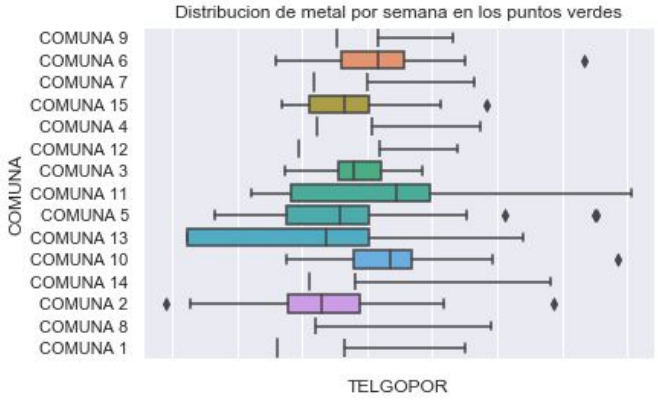
3.6.2 Papel y carton



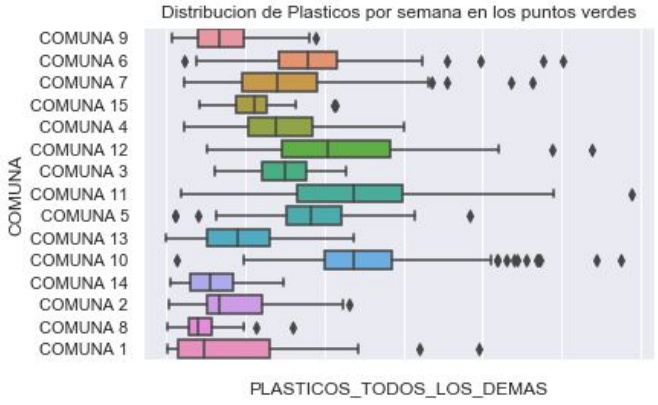
3.6.3 Metal



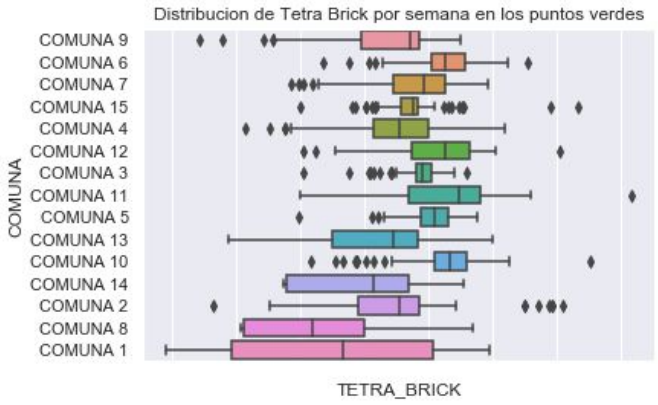
3.6.4 Telgopor



3.6.5 Plasticos



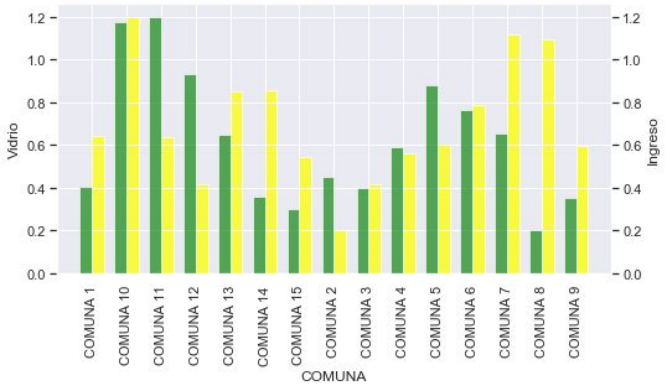
3.6.6 Tetrabrick

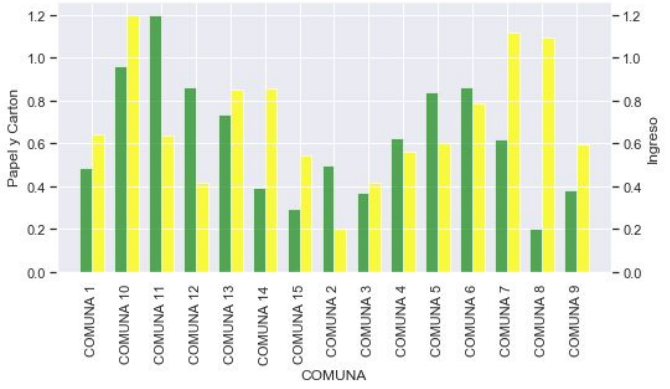


Como podemos observar, la cantidad de outliers se concentra en la mayoría de los materiales en las comunas 11, 10, 12 y 6. Esto tiene sentido ya que son las que mayores Kgs producen y tienen mayores posibilidades de tener semanas fuera del percentil 97. Es importante tener en cuenta esta observación ya que al filtrar estos outliers debemos hacerlo comuna por comuna, de lo contrario eliminaremos en su mayoría datos de las comunas que más producen y dejariamos anomalías en las distribuciones de comunas que producen menos.

3.7 Relación de ingresos económicos por comuna con pesajes

En este punto investigaremos si hay relación entre el ingreso medio mensual per cápita de cada comuna y los residuos que fueron separados en las mismas. Realizamos gráficos de barras donde comparamos el ingreso medio con los pesajes por comuna de vidrio, papel y cartón, que son los materiales más importantes.





Analizando los gráficos vemos que existe una correlación significativa pero no perfecta entre estos 2 factores en la mayoría de las comunas que tiende a que haya más reciclados mientras mayor es el ingreso. Esta se cumple menos en las comunas 8, 9, 11 y 14.

4 Materiales y métodos

Debido a que nuestros datos son pesajes en kilogramos de distintos materiales reciclados, y que los mismos se encuentran ingresado semana por semana de forma muy variada, utilizaremos un modelo supervisado que prediga como salida la comuna a la que pertenece una lista semanal de pesajes que ingresa como entrada.

Nuestra hipótesis es que las comunas tienen combinaciones distintas de distribuciones para cada pesaje, lo que les otorga una cualidad única para, en caso de detectarla, poder predecir de que comuna es una lista semanal de pesajes.

En principio, se realiza un filtro de outliers con el propósito de eliminar anomalías en los resultados. Este filtro lo realizamos por cada comuna separada ya que como vimos en 3.6 estas manejan escalas distintas de pesajes, por lo cual sería incorrecto filtrar todas juntas porque eliminaría mayormente registros de las que más producen en promedio. Obtuvimos los percentiles 97 de cada pesaje en cada comuna y filtro mi dataset por cada uno.

Buscamos aplicar la técnica Cross validation que se realiza con las muestras de entrenamiento para evitar un sobreajuste de los modelos. Consiste en dividir nuestro training set en K folds (porciones) e iterar K veces. En cada iteración, una porción se utiliza como validación y el resto como train donde se entrena un modelo con train evaluará el resultado de clasificación con validación. Luego se realizará un promedio de los resultados de exactitud de todas las iteraciones.

Dividimos nuestros datos en train y test, siendo este último grupo un 30% del total. El clasificador aprenderá la regla de decisión utilizando el train set (samples + labels). Luego clasificará las muestras de test (sin mirar las labels de test) y se medirá la exactitud de clasificación en testeo. Esto lo realizamos para

Se realiza una selección de features mediante la técnica Variance Threshold que consiste en remover las features que presentan una varianza menor a un umbral determinado. En nuestro caso elegimos un umbral de 0.2.

Evaluaremos con tres tipos de clasificadores para ver cuál se ajusta mejor a nuestro trabajo: K-nearest neighbors (KNN), Support Vector Machines (SVM) y Logistic Regression (LG).

4.1 K-nearest neighbors[2]

En el método KNN se clasifica cada nuevo set de pesajes con la comuna que corresponda, según tenga K vecinos más cercanos de una comuna o otra. El cálculo de la distancia que utilizamos es el default, que es la Euclidiana que se corresponde a la fórmula[5]:

$$dist\,(A,B) = \sqrt{\frac{\sum_{i=1}^{m}(x_i - y_i)^2}{m}}$$

Es importante destacar que la función de distancia utilizada podría afectar los resultados obtenidos.

Donde p y q son 2 puntos ubicados en un hiperespacio de n dimensiones (n pesajes distintos en nuestro caso). Se selecciona la etiqueta (comuna) qué más frecuente aparece entre las 15 diferentes.

Realizamos un ciclo de 50 iteraciones para probar con distintos valores de K (hiper parámetro, vecinos). El resultado fue que 5 vecinos era la cantidad óptima.

4.2 Support Vector Machines[1]

Este modelo clasificador lineal busca el hiperplano separador que maximiza el margen entre clases (comunas). Cuando las clases no son separables linealmente se acude a un soft-margin, penalizador que permite muestras mal clasificadas. Cada una de estas es penalizada por un costo C que el usuario selecciona (un ej. de hiper-parámetro). El margen separador queda definido por “s” muestras que son nuestros support vectors.

Nosotros además decidimos utilizar un kernel Gaussiano. Tiene la fórmula, según M. Bishop (2006):

$$K_{gaussian}(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

Este mapea nuestros datos a una dimensión desconocida (de Hilbert) donde son linealmente separables. Allí en ese nuevo espacio donde son mapeadas las muestras se aplican los productos internos (o similaridad).

Para los hiperparámetros de Costo y Gamma junto con los K folds de cross validation decidimos hacer un Grid Search para que se prueben todas las combinaciones de hiperparámetros y elegir la mejor.

4.3 Logistic Regression

Por último elegimos otro clasificador lineal. Es una regresión lineal precedida de una función de activación sigmoide lo que genera que la salida sea binaria y no continua como una regresión normal. Puede entenderse como una red neuronal de una sola capa y una sola neurona.

A cada muestra clasificada, le asigna una probabilidad de pertenecer a cada clase existente en el problema según la siguiente fórmula extraída de Christopher M. Bishop:

$$p(y_i|\mathrm{x}) = \sigma(w^T\mathrm{x})$$

Donde sigma representa la función sigmoide extraída de Christopher M. Bishop:

$$\sigma(\mathrm{x}) = \frac{1}{1+\exp(-\mathrm{x})}$$

El vector w traspuesto es el vector de pesos que ajusta el modelo en cada iteración y x es el vector de entrada (pesajes de la semana en este caso). Si la probabilidad es mayor a cierto threshold (0.5) entonces pertenece a una clase y viceversa.

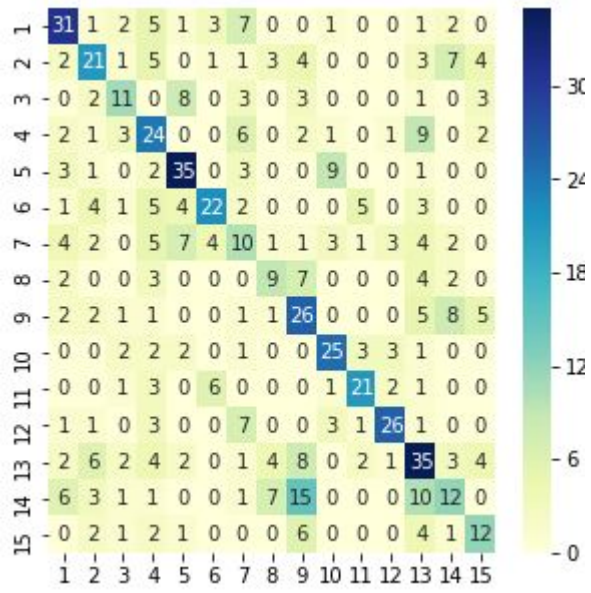
Este método también tiene un hiper parámetro C que representa el costo que penaliza las muestras mal clasificadas. También tiene otro penalizador al vector W bajo la norma L1, en nuestro caso,. De esta manera evita que el modelo quiera sobre-ajustarse a los datos de training. En otras palabras, evita que existan posiciones de W muy altas y

otras casi nulas, o viceversa, que solo algunas posiciones de W se activen y el resto no.

5 Resultados

El mejor de los 3 analizados fue el Super Vector Machine, con un accuracy de 50,91% para predecir entre 15 comunas distintas.Este se utilizó con un Costo = 3 y un valor Gamma = 1.

A continuación mostramos su matriz de confusión que permite ver en qué predicciones fallo más nuestro modelo.



Como podemos ver las comunas 14 y 19 son las que más se confundieron entre sí tratando de predecirlas, lo que podría indicar que sus distribuciones de pesajes son muy parecidas. Esto es inesperado ya que son muy geográficamente distantes entre sí y tienen ingresos significativamente distintos.

6 Discusión y conclusiones

Como nuestro modelo logró una significativa accuracy de 50% para clasificar entre 15 clases distintas, podemos suponer que tienen si distribuciones de features distintas. Creemos que nuestro modelo clasificador podría ser mucho más exacto si tuviéramos más datos y si no hubiéramos tenido que llenar varios valores con medias en nuestro dataset para registros faltantes o nulos.

Al haber visto la gran diferencia entre el tetrabrik, con el vidrio, plástico y papel, creemos que se debe fomentar el reciclaje de este tipo, dando conocimiento a la población de cómo está formado un tetrabrik y porque es importante separarlo. Este es un tipo de material que la población de CABA no sabe bien cómo reciclar y es uno de los más presentes en las cocinas argentinas.

Lo más importante de nuestro trabajo es que tenemos la certeza de que de tener más datos sobre materiales reciclables podríamos hacer un análisis más exacto y profundo para generar información para tomar decisiones. Desperdiciamos mucho tiempo en limpiar los datos por falta de estandarización en cómo estos se completan.

Juntando los errores más comunes en el llenado de los datos proponemos consejos para su estandarización en el futuro:

1) El llenado de los datos debe hacerse mediante una interfaz que asegure que los datos ingresados por pesaje sean solo números.

2) Los números ingresados deben utilizar coma (,) para los valores decimales y no deben ingresarse con punto (.) para separar millares.

3) En caso de no ingresarse una semana, esta debe completarse automáticamente con valores nulos (NaN) en cada pesaje.

7 Referencias

1. Andrew Ng, CS229 Lecture notes**.** Recuperado de: http://cs229.stanford.edu/notes/cs229-notes3.pdf
2. k-nearest neighbors algorithm: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
3. Ley 1.854 “Basura cero”: https://www.buenosaires.gob.ar/areas/leg_tecnica/sin/normapop09.php?id=81508&qu=c&cp&rl=1&rf&im&mot_toda&mot_frase&mot_alguna
4. Christopher M. Bishop: Pattern Recognition and Machine Learning, 2006. Cambridge, U.K. Springer
5. Li-Yu Hu, Min-Wei Huang, Shih-Wen Ke and Chih-Fong Tsai, 2016, The distance function effect on k-nearest neighbor classification for medical datasets. Recuperado de: https://springerplus.springeropen.com/articles/10.1186/s40064-016-2941-7#citeas