MEMORIA DE LA PRÁCTICA BICIMAD

1. MOTIVO:

La finalidad de esta práctica es, empleando los datos facilitados por la empresa de alquiler público de bicicletas eléctricas BiciMAD de Madrid, clasificar cada estación de bicicletas en función de los usuarios que hacen uso de ellas. En nuestro caso, estudiaremos el riesgo de accidente de los usuarios etiquetando cada estación con una puntuación que vendrá determinada por la edad, el tipo, el horario, …

Esta puntuación estará comprendida entre los valores 0, correspondiente a una estación con poco riesgo de accidentes, y 1, correspondiente a una estación con bastante riesgo.

El programa utilizará los datos obtenidos para comparar el nivel de riesgo de cada estación.

Para clarificar el análisis de los datos, se presentará una tabla con la información ordenada y clasificada según se desee y un mapa en el que vendrán indicadas las estaciones más destacadas del estudio.

Nuestro estudio estará basado en el sentido común, ya que carecemos de un informe estadístico. Sin embargo, con nuestras puntuaciones podemos establecer unos buenos resultados para poder disminuir el riesgo de accidente al utilizar una bicicleta de esta empresa. Obviamente, sería necesario revisar estas puntuaciones en un futuro y optimizar el mecanismo del programa con un estudio estadístico sobre el tema.

Estos resultados podrían servir para aumentar la seguridad policial en las estaciones con más riesgo, construir más carriles bici (en especial, que comuniquen dichas estaciones), promover campañas de marketing enfocadas a la seguridad vial, etc.

2. ESTRATEGIA:

La función principal del programa consiste en analizar el comportamiento de cada usuario durante un día y otorgar una puntuación comprendida entre 0 y 15 en función de su edad, el tipo de usuario, numero de viajes realizados y los intervalos horarios de los mismos. Finalmente normalizaremos la puntuación entre 0 y 1.

Una vez asignada la puntuación a cada usuario y obtenida la lista de estaciones con las que ha tenido contacto, se atribuirá a cada estación la puntuación del usuario. Posteriormente, se irán añadiendo a cada estación las puntuaciones de los usuarios que han tenido contacto con las mismas, para obtener una media total y conseguir así el objetivo del estudio.

Un usuario con puntuación menor de 0.5 será clasificado como "usuario con poco riesgo" y un usuario con puntuación mayor de 0.5, como "usuario con bastante riesgo". Por tanto, esta puntuación quedará recogida en las estaciones con las que el usuario ha tenido contacto.

3. PUNTUACIONES

A cada cliente se le asigna una puntación entre 0 y 15 en función de:

Como se ha dicho en el anterior punto, a cada cliente primeramente se le atribuirá una puntuación entre 0 y 15. Esta puntuación es el resultado de la suma de tres niveles de puntuación, cada nivel con atribución desde 0 hasta 5 puntos:

1) Edad:

Aquí otorgamos los puntos al cliente según la edad. Diferenciamos los siguientes rangos:

- Menores de 18 años: 4 puntos (Consideramos que tienen menos experiencia a la hora de manejar una bici con estas edades y no conocen las normas de circulación al no tener el carnet de conducir)

- Entre 19 y 26 años: 5 puntos (Edades propicias para actuar sin la cordura y sensatez suficiente ante una situación complicada. A esto le añadimos un porcentaje de jóvenes que pueden coger una bici bajo los efectos de alguna sustancia estupefaciente)

- Entre 26 y 40 años: 3 puntos (Bicis utilizadas para el trabajo por lo que baja el riesgo. Pero se mantiene el riesgo de utilizarla después de alguna celebración, comidas, fiestas, etc.)

- Entre 41 y 65 años: 1 punto (Es menos probable que la gente utilice la bici a estas edades por mitad de la ciudad. En caso de cogerla sería de una forma muy prudente)

- Mayores de 66 años: 1 punto (Consideramos que solamente se coge alguna bici esporádicamente y con mucho cuidado)

2) Nivel de tipo de usuario (ocasional u anual):

En este nivel, consideramos que si un cliente tiene el abono anual es que tiene más experiencia con las bicicletas, y, por tanto, se reduce mucho el riesgo

- Cliente ocasional: 3,5 puntos
- Cliente anual: 1 punto

3) Nivel intervalo de horas:

Interpretamos que en el intervalo de mañanas, entre semana, hay bastante tráfico por lo que existe un riesgo moderado y según va pasando el día el riesgo va aumentando por el cansancio, ingerir alcohol, etc. Para el fin de semana, hacemos una interpretación parecida, aunque aumentando el riesgo ya que el uso de las bicicletas sería mayoritariamente recreativo y más propicio de accidente.

Si un cliente utiliza la bicicleta dos veces en un día, dependiendo del intervalo de llegada y de salida, hemos adjudicado las puntuaciones suponiendo si era la ida o vuelta del trabajo, si era la vuelta de alguna celebración, la ida hacia alguna discoteca o simplemente un viaje casual. Hemos creado una matriz simétrica, que aparece más abajo para indicar las puntuaciones.

Finalmente, si un cliente utiliza la bicicleta tres o más veces en un día es complicado designar ante qué tipo de cliente estamos y el riesgo que conlleva, así que en estos casos sólo diferenciaremos la puntuación teniendo en cuenta el día de la semana de los mismos (los fines de semana atribuiremos más puntos).

Los intervalos horarios que utilizaremos para un mismo día serán:

a) Intervalo 1: De 7:00 a.m. a 09:59 a.m. (Horario de ir a trabajar y mucho tráfico).

b) Intervalo 2: De 10:00 a.m. a 12:59 p.m. (Horario de compras, paseos y menos tráfico).

c) Intervalo 3: De 13:00 p.m. a 15:59 p.m. (Horario de comidas).

d) Intervalo 4: De 16:00 p.m. a 20:59 p.m. (Horario de ocio).

e) Intervalo 5: De 21:00 p.m. a 22:59 p.m. (Horario de vuelta del trabajo y cenas, cansancio)

f) Intervalo 6: De 23:00 p.m. a 23:59 p.m. y de 00:00 a.m. a 06:59 a.m. del mismo día.

(Horario nocturno donde los usuarios pueden estar en discotecas y hay poco transporte púbico).

Estos datos sobre los intervalos en combinación con el día de la semana que se produce el viaje (haremos diferencias entre día de entre semana y fines de semana) nos dan las siguientes tablas de puntuaciones:

ENTRE SEMANA – 1 VIAJE						
INTERVALO	1	2	3	4	5	6
PUNTUACIÓN	3	2.5	1.5	3	3	2

FINES DE SEMANA – 1 VIAJE						
INTERVALO	1	2	3	4	5	6
PUNTUACIÓN	1	1	2	3	4	4

En las siguientes tablas, referentes a dos viajes, se muestra en la primera columna el intervalo de salida y en la primera fila el intervalo de llegada:

ENTRE SEMANA – 2 VIAJES						
INTERVALO	1	2	3	4	5	6
1	3	2.5	1.5	3	3	2
2	2.5	2	1.5	2.5	2.5	3
3	1.5	1.5	1	2	2.5	3
4	3	2.5	2	2	2.5	3
5	3	2.5	2.5	2.5	2.5	4
6	2	3	3	3	4	4

FINES DE SEMANA – 2 VIAJES						
INTERVALO	1	2	3	4	5	6
1	1	1	2	3	4	4
2	1	1	2.5	3.5	3.5	4
3	2	2.5	2	3.5	3.5	4
4	3	3.5	3.5	3.5	4	4.5
5	4	3.5	3.5	4	4	5
6	4	4	4	4.5	5	5

Ante la posibilidad de no poder encontrar en los archivos de datos información sobre edad y/o tipo de usuario, en esos casos la puntuación del mismo irá sobre 10 o sobre 5 puntos, y la media se hará en consecuencia teniendo en cuenta la falta de parámetros.

4. Procedimiento

A continuación expondremos todos los pasos que se han llevado a cabo para realizar el análisis y su implementación correspondiente.

1) En un primer momento cargamos los datos originales de la rdd y con un primer mapeo extraemos los datos que nos interesan que son:
    - El nombre del usuario (id)
    - La edad del usuario
    - El tipo de usuario
    - Origen y destino del viaje

- Hora en la que se realiza el viaje

Con unas funciones auxiliares llamadas edad (nos devuelve la puntuación correspondiente a la edad), tipo (nos devuelve la puntuación correspondiente al tipo de usuario), fecha_hora (nos devuelve una tupla, en la que el primer elemento nos indica si el viaje se realiza entre semana "semana" o en fin de semana "finde", el segundo es la estación de origen y el tercero es la estación de destino), viaje1 (devuelve la puntuación derivada al usuario que realiza un único viaje) y viajes2 (análogo a viaje1 pero con dos viajes) establecemos las puntuaciones a cada usuario según los datos anteriormente expuestos.

Finalmente transformamos en diccionarios los datos del primer mapeo de la forma: *{id: …, edad:…, tipo:…., viajes: [(fecha, origen, destino), (…., …., …..), ….]}.* Este ultimo elemento se corresponde con el resultado de ejecutar la función fecha_hora. Creamos una segunda rdd con estos diccionarios

2) En un segundo lugar, queremos obtener la puntuación por usuario utilizando las funciones auxiliares que hemos creado, para obtener una rdd de diccionarios con el nombre del usuario, su puntuación media y las estaciones por las que ha pasado.

Para realizar esta media, hay que tener en cuenta que muchas veces no aparecen los datos de la edad o el tipo (la información con respecto a los viajes siempre aparece). Por tanto creamos unos parámetros de edad y tipo que utilizaremos para realizar la media. Las puntuaciones se encuentran en un rango entre 0-5. Por lo tanto podemos llegar a obtener una puntuación máxima de 15 puntos. Entonces debemos dividir entre 5,10 o 15 en función de los parámetros que tengamos.

Si nos falta el parámetro del tipo, solo contamos las puntuaciones de la edad y los viajes y dividimos entre 10. Análogo si faltan los dos parámetros que se dividiría entre 5 solamente.

Los diccionarios serían de la forma: *{id:…., puntos: ….., estaciones: […,…….,….]}*

Creamos una tercera rdd con estos diccionarios.

3) Finalmente queremos otorgarle la puntuación de los usuarios a cada estación y establecer una media con la cuál poder determinar que estaciones tienen un riesgo menor o mayor de accidentes, que es el objetivo de esta práctica.

   - La idea es para cada estación sumarle la puntuación de las n personas que la han usado y dividirla entre n para obtener una media, sabiendo que si la puntuación es menor que 0,5 se corresponde con una estación con poco riesgo y si es mayor que 0,5 será una estación con mayor riesgo.

   - Para ello es necesario aislar cada estación haciendo una rdd mas grande en la cual podamos separar estaciones de usuarios.

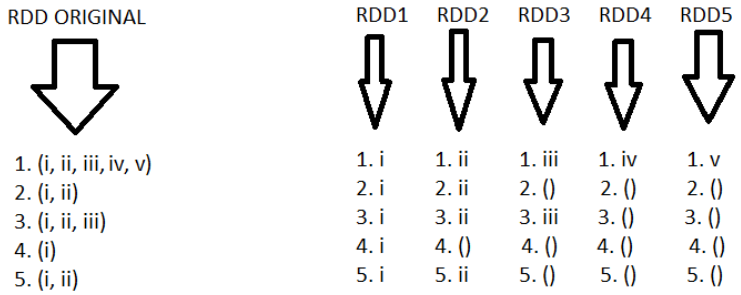
   - Con un mapeo separamos las estaciones asignándoles la puntuación del usuario que la haya utilizado, obteniendo n listas correspondiente a los n usuarios. Las listas serán de la forma *: {(estacion13, {puntos: …., numero_personas: 1}), …….}* En cada lista aparecen las estaciones por las que ha pasado un usuario. El numero de personas siempre es 1 ya que solo lo utiliza un único usuario. Finalmente sumaremos este numero de personas para cada estación para saber el numero de usos total.

   - La idea es transformar esta rdd de listas en una rdd en la que cada línea sea la información correspondiente a una estación, es decir pasar una lista de la forma *[(estacion13, {puntos: ..., numero_personas: 1}),('estacion47', {'puntos': ..., 'numero_personas': 1}), ...]* en líneas de la siguiente forma :

     (estacion13, {puntos: ..., numero_personas: 1})
     ('estacion47', {'puntos': ..., 'numero_personas': 1})
     …..
     …..

     Para ello primero calculamos el numero de elementos que tiene la lista más larga de la rdd original de listas mediante una reducción. Este número será k.

Crearemos una lista de k rdds donde cada línea de estas rdds estará formada por la componente k-ésima de la rdd original. Por supuesto habrá líneas en las que componente k-ésima no exista, en estos casos añadimos una tupla vacía para después filtrarla y quitarlas. Añadimos una imagen explicativa.



Con un mapper extraemos el elemento k-ésimo de cada línea, unimos las rdds y filtramos para quitar las tuplas vacías llegando a la rdd que pretendíamos.
Finalmente unimos estas rdds filtrando las tuplas vacías llegando a la rdd que pretendíamos. Esta será nuestra quinta rdd.

- Realizamos un groupByKey para juntar todas las puntuaciones correspondientes a la misma estación mediante un mapeo. Nos quedaría una rdd con las líneas de la siguiente forma: (estación19, [{puntos: ….., numero_personas: 1}, {puntos:…., numero_personas: 1}, …….])

  Esta será nuestra sexta rdd.

- Para terminar transformamos la rdd anterior en una lista de diccionarios donde aparece la estación, la media de las puntuaciones (sumando las puntuaciones y dividiendo entre el numero total de usos), el numero total de usos, las estaciones con menos riesgo (cuya media sea menor que 0,5) y las estaciones con mayor riesgo (cuya media sea mayor que 0,5)

Para esto realizamos un mapeo extrayendo la información de cada estación y juntando los elementos de las puntuaciones. Esta será la rdd final que luego exportaremos en un archivo nuevo.

5. IMPRESIÓN

Finalmente creamos otra función para poder visualizar los resultados en forma de tabla y de gráfica. Podemos ordenar la tabla en función de las estaciones, la media de éstas, sus usos, las estaciones con menor riesgo y las estaciones con más riesgo. Además, puedes clasificarlas en un orden creciente o decreciente, para comprobar las que más y menos riesgo tienen.

6. COMO EJECUTAR LA PRÁCTICA

He habilitado un espacio Gmail con la práctica cargada en el drive junto a varios archivos de movimientos de BiciMad. Entrar en la siguiente cuenta:

Practicagrupo007@gmail.com

Contraseña: GRUPO007

Todos los datos han sido obtenidos de https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1)