Instituto Tecnológico de Costa Rica

Escuela de Ingeniería en Computación

Entregable pdf del proyecto:

Big data

Profesor:

Manuel Zumbado Corrales

Estudiantes:

Joshua Solís Fuentes – Carné:2023064637

Explicación de la estructura del programa:

En el repositorio de GitHub que se puede acceder con el siguiente link https://github.com/Josh-SolisF/Proyecto-BigData se encuentran todos los archivos necesarios para la ejecución del proyecto.

Este sería el manual de ejecución que sigue el flujo de como correr el programa:

Manual de Ejecución

Debe ejecutarse en este orden:

1. bash build_image.sh (crear imagen con Docker).

2. bash run-postgres.sh (levanta la imagen de la Base de Datos).

   * Credenciales: usuario= `postgres`, contraseña=`pword`, puerto=`5433`

2. bash run_image.sh. Se levanta un contenedor con la imagen del paso anterior (`bd-tarea3`), el contenedor se llamará `bd-tarea3-container`. Si no esta disponible bash en esta etapa, puede ejecutar el comando: `docker run -p 8888:8888 -it --rm --name bd-tarea3-container bd-tarea3 /bin/bash`

3. Procure asegurarse de estar dentro del shell de Bash .

4. bash connect-psgres.sh, esto se encarga de conectarlo a la base

5. bash jupyter-server.sh. Ejecutamos el servidor de Jupyter.

6. Copiar el enlace al final de la salida generada por el paso anterior. No trate de acceder únicamente contectando al puerto 8888 en `localhost`, porque le pedirá este token o clave de autenticación.

Cosas que considerar sobre la ejecución del cuaderno de Jupyter:

- El cuaderno ya tiene las salidas de la última ejecución, la cual fue probada varias veces previo a realizar la entrega del trabajo. Si de igual forma se desea ejecutar, tenga los siguientes datos en cuenta:
- Durante todo el cuaderno se arroja unos warnings que se deben a la configuración de spark, cada vez que se dispara un hilo que utiliza una cantidad "alta" de memoria. Esto mejor no cambiarlo para evitar causar un error fatal.
- Tanto el entrenamiento de RF, como el uso de Cross-Validation, así como la generación de los cuartiles y otras cosas más son muy lentos si se ejecutan en arquitecturas distintas a la de la plataforma del contenedor (amd64) por lo que se debe tener paciencia.

Contenido del programa:

Todo el programa se encuentra en el jupyter notebook, se utilizó de bibliotecas externas para lograr correr las pruebas unitarias, se hizo de esta manera porque es mucho más sencillo de seguir los bloques de código y prosa que ofrece esta herramienta. Aunque muchas tareas duran más porque se esta perdiendo un poco de capacidad de procesamiento lo hace más fácil de seguir.

Propuesta:

Fuentes de datos analizadas:

Asequibilidad de los alimentos

https://data.chhs.ca.gov/dataset/food-affordability-2006-2010/resource/916e2a2e-383b-4af5-9f5b-310500961cb5

Datos o features:

Ind_id, ind_definition, reportyear, race_eth_code, race_eth_name, geotype, geotypevalue, geoname, country_name, country_flips, region_name, region_code, cost_yr, median_income, affordability_ratio, LL95_affordability_ratio, UL95affordability_ratio, se_food_afford, rse_food_afford, food_food_decile, CA_RR_affordability, ave_fam_size, version.

Desempleo del estado de california del 2004-2013

https://data.chhs.ca.gov/dataset/unemployment-2004-2013/resource/2ecd7fda-2317-4fa1-8a6e-1834cfa39cc0

Ind_id, ind_definition, reportyear, race_eth_code, race_eth_name, geotype, geotypevalue, geoname, country_name, country_flips, region_name, region_code,, Labor_Force, Unenmployment_rate, II_95ci, ul_95ci, se, rse, place_decile, ca_rr, version

En este caso ambos de estos datasets son del condado de california por lo que su unión es bastante natural, uno trata sobre la asequibilidad de los alimentos y el otro sobre el desempleo, ambos de estos temas se relacionan. Si hay una mayor tasa de desempleo es posible que la accesibilidad de los alimentos se vea influenciada, ambos ocurren en el mismo lapso del tiempo sin embargo en el caso del segundo dataset lo especifica más a que año, en el segundo solamente se muestra el rango entre el 2006-2010.

La “clave” de unión sería el código de región que comparten en común y hacer el análisis a partir de la zona

La variable a predecir sería “Alta seguridad Alimentaria” (Alta/Baja)

Si un condado tiene alta inseguridad alimentario si el porcentaje del ingreso gastado en alimento supera un umbral en este caso propondría 30%

Programa:

El programa en su totalidad se encuentra en el jupyter_notebook, incluyendo la carga de datos, el preprocesado y las pruebas unitarias. Por lo que el cuaderno contiene una mejor estructura para entender y documentar el código se recomienda correrlo para tener un mejor entendimiento del como funciona el programa