Enunciado Proyecto # 2

Instrucciones

Utilizando los conceptos aprendidos de aprendizaje no supervisado en clase, se necesita realizar un análisis exploratorio del desempeño de los estudiantes de la carrera de Ingeniería en Informática y Sistemas en función de su historial de cursos, tanto aprobados como no aprobados, y de las variables que puedan ayudar a encontrar comportamientos a nivel global del desempeño académico de los alumnos

El objetivo de esto es poder encontrar información o patrones "ocultos" que no siempre son fácilmente visibles y que también por el gran volumen de estudiantes y la cantidad de secciones que existen es muy difícil identificar.

Específicamente se busca por medio de técnicas de aprendizaje no supervisado resolver planteamientos tales como:

Implementar Agrupamiento (Clustering) – 20pts

1. ¿Cómo podemos clasificar a los estudiantes en grupos en función de su rendimiento académico?
    - Aquí lo que se pretende encontrar es patrones de comportamientos en cuanto al desempeño que ha tenido el estudiante, en el cual se plantean (pero no se limita a) los siguientes ejemplos
        - Similitudes por promedio simple
        - Similitudes por promedio ponderado
        - Similitudes por promedio de cursos numéricos (matemáticas, física, química, ingeniería económica, investigación de operaciones, etc..)
        - Similitudes por promedio de cursos específicos a su área de ingeniería (para informática por ejemplo podrían ser Bases de Datos, Ingeinería de Software, Sistemas Operativos, etc...)
        - Similitudes por promedio de cursos CFI (Magis, ética, etc...)
        - Similitudes por número de cursos perdidos
        - Similitudes por número de cursos aprobados

    - Lo que se busca encontrar por medio de identificar dichos patrones numéricos es agrupar a estudiantes y clasificarlos en categorias, como por ejemplo:
        - Estudiantes que van bien en cursos de ingeniería y CFIs, con un promedio aceptable (por ejemplo de 70 a 75pts) pero mal en cursos numéricos
        - Estudiantes que van bien en cursos numéricos y CFI pero mantienen un promedio en cursos de ingeniería más bajo, con un promedio alto (de 80pts+)
        - Etc…

A ciencia cierta se desconoce que se puede llegar a encontrar y por esto es que busca aplicar una técnica de análisis exploratorio como "K-means" que pueda dar indicios de los diferentes grupos y categorías de estudiantes que pueden llegar a existir en función de todas las variables

Implementar Reglas de Asociación – 10pts

2. Se desea también encontrar según el registro histórico de ciclos cual es el comportamiento de los índices de aprobación y no aprobación de cursos en función del resto de cursos que llevan los alumnos para responder a preguntas tales como
    - ¿Cómo podemos identificar qué tanto es el índice de aprobación de ciertos cursos en función del ciclo? Lo que se busca es tratar de encontrar reglas de asociación, Por ejemplo:
        - *"Primer Ciclo -> Reprobar Matemática I" con una confianza del 60%,* esto implicaría que siempre en primer ciclo reprueban Matemática I el 60% de las veces y sería interesante compararlo con el Segundo ciclo
        - *"Primer Ciclo -> Cursos Numéricos" con una confianza del 40%,* esto implicaría que siempre en primer ciclo reprueban algún curso numérico el 40% de las veces lo cual sería una visualización más general muy interesante de encontrar
    - ¿Cómo podemos identificar si existe una regla que nos muestre si reprobar un curso (o más cursos) implica también reprobar otro? ¿de igual forma con aprobar? Lo que se busca es tratar de encontrar reglas de asociación, Por ejemplo:
        - *"Reprobar Lenguajes Formales y Arquitectura del computador I -> Reprobar Bases de Datos I" con una confianza del 35%,* esto implicaría que reprobar esos dos cursos influye en un 35% de las veces reprobar otro, dado que estos tres se imparten en un mismo ciclo
        - *"Aprobar Estrategias de Razonamiento -> Aprobar Matemática I" con una confianza del 75%,* esto implicaría que aprobar este curso influye en un 75% aprobar el otro

El objetivo de aplicar reglas de asociación es tratar de visualizar cada ciclo como si fuese una "transacción" en la cual hay diferentes cursos y diferentes resultados, de modo que puedan encontrarse posibles reglas y patrones que ayuden a la facultad a ver si existe causalidad entre el ciclo y el curso, o los cursos vs otros cursos, para lo cual debe de aplicar medidas de soporte y confianza según su criterio (tomando en cuenta lo recomendable) y no ignorar medidas como el "lift"

Ambas técnicas de exploración servirán enormemente para que la Facultad pueda encontrar casos de alumnos que están teniendo problemas en ciertas áreas del pensum de su carrera pero que en general tienen un promedio aceptable o bueno, con el objetivo de tomar medidas de apoyo y soporte para reforzar y ayudar a los alumnos que tengan algún problema.

Datos

Se adjuntan los siguientes archivos con los datos necesarios para realizar el análisis:

1. ListadoPromedios.xlsx – Muestra de estudiantes de Informática y Sistemas con las siguientes columnas:
    a. Año – año de inicio de carrera
    b. ID – Identificador del estudiante
    c. Sede – sede a la que pertenece
    d. Facultad – facultad a la que pertenece
    e. carrera – carrera a la que pertenece
    f. prom_simp_x_ciclo – promedio simple del primer ciclo 2019
    g. prom_pond_x_ciclo – promedio ponderado del primer ciclo 2019
    h. cursos_x_ciclo – cursos asignados durante el primer ciclo 2019
    i. Promedio_Simple_Acumulado – promedio simple acumulado
    j. Promedio_Ponderado_Acumulado – promedio ponderado acumulado
    k. cursos_acumulados – cantidad de cursos (aprobados) acumulados
2. Pensum11001.xlsx, Pensum13001.xlsx y Pensum18001.xlsx
    a. Nombre_Sede - Nombre de sede
    b. Nombre_Facultad - Nombre de facultad
    c. Nombre_Carrera - Nombre de carrera
    d. Nombre_Titulo - Nombre de titulo
    e. Nombre_Curso - Nombre de curso
    f. Cred_Teo – cantidad de créditos teóricos
    g. Cred_Pra – cantidad de créditos prácticos
    h. Nombre_Periodo – Ciclo en que se imparte el curso
    i. No_Pensum – código del pensum
    j. No_Curso – código del curso
3. Carpeta notas – Posee un archivo .CSV por cada estudiante del archivo "ListadoPromedios", el nombre del archivo es el ID del estudiante
    a. No_Sede – ID de sede
    b. Sede – Nombre de sede
    c. No_Facultad – ID de facultad
    d. Facultad – Nombre de facultad
    e. No_Carrera – ID de carrera
    f. Carrera – Nombre de carrera
    g. No_curso – ID de curso
    h. Nombre_Curso – Nombre de curso
    i. Nota – Nota de curso
    j. Nota_Calculo_Prom – Nota de curso que sirve para calcular promedio
    k. Creditos – créditos del curso
    l. No_ciclo – ID de ciclo
    m. Nombre_Ciclo Eje – Nombre de facultad
    n. No_pensum – ID de pensum
    o. No_jornada – ID de jornada

p. No_seccion_fac – Número de sección
q. No_tip_examen – ID del tipo de examen
r. Tipo_Examen – Nombre del tipo de examen

Entregables

Los entregables esperados son:

- Scripts de R (archivo .r) realizados para cada análisis
- Datasets de R (archivo .rdata)
- Documento con Resultados obtenidos de cada análisis explicando con alto nivel de detalle
    - Como se proceso la información
    - Cada una de las herramientas utilizadas (Kmeans, a priori, etc..)
    - Cada script realizado en R
    - Cada gráfico o plot realizado
    - Resultados obtenidos
        - Clasificaciones encontradas (kmeans)
        - Reglas encontradas (a priori)
- Presentación en Power Point con el resumen de los resultados obtenidos en el documento previo
- Repositorio público en Azure Notebooks con cada script y archivo
- Opcional: Reportes de Tableau usando los datasets de R (+2pts extras)

Fecha de entrega

Viernes 29 de Noviembre del 2019 a las 23:55 horas (no habrá prorroga), para dicha fecha deberá de proveer lo siguiente:

- Subir al portal URL por medio del link llamado “Entregables Proyecto 1” un archivo comprimido con:
    - Archivos de R (.R y .Rdata)
    - Plots (en imágenes o pdfs)
    - Documentación de resultados
    - Presentación de Power Point
- Publicación de scripts y archivos en repositorio de Azure Notebooks (recuerde que se tomará en cuenta la fecha de modificación del repositorio)

El proyecto puede realizarse en grupos de un máximo de 3 integrantes.