IIC2433 - Minería de Datos - 2° Semestre 2020
Profesor: Vicente Domínguez
Sección: 1

Informe Entrega Parcial

Grupo 14

1. Introducción

Una de las causas comunes de estrés y malestar psicológico de los estudiantes universitarios es la carga académica, que presenta múltiples factores asociados al bienestar emocional de los alumnos. Desde la dificultad misma del ramo hasta la demanda en tiempo que significa la asignatura pueden ser algunos de estos factores. Por ello, tanto para la universidad como para los estudiantes es necesario desarrollar herramientas que ayuden a un mejor enfrentamiento de la carga académica. Dentro de los distintos aspectos que se pueden abordar para resolver este problema existe la opción de ayudar a los estudiantes a planificar mejor sus mallas curriculares, de tal forma de evitar semestres sobrecargados que puedan culminar en un desequilibrio académico. Desde una perspectiva de la minería de datos esta solución es similar a diseñar un sistema recomendador de ramos, orientado a evitar el riesgo académico de los estudiantes.

Tomando los contenidos del curso se puede entonces hacer una aproximación a un sistema recomendador que, basándose en la experiencia de alumnos previos (base de datos) permita consultar qué combinaciones de ramos según cada alumno tendrán más probabilidad de causar un desequilibrio académico.

Por lo tanto, el objetivo final o problemática que se quiere resolver con el procesamiento de datos y las predicciones correspondientes es evitar la toma de ramos que en un mismo semestre puedan causar un riesgo académico para el estudiante, con el trasfondo de disminuir el malestar emocional de los alumnos.

En el presente documento se expondrán todos los aspectos necesarios para el desarrollo de la herramienta planteada.

2. Descripción de los datos

Dataset

Para la realización del proyecto haremos uso de tres *datasets* que nos proporcionó la Dirección de Pregrado. Estos contienen la información de todos los ramos cursados hasta el presente año por los alumnos de las admisiones **2015, 2016 y 2017**. Cada fila del *dataset* representa un curso que fue tomado por un alumno en un semestre y año en específico, junto con su nota final. Cabe destacar que los datos están anonimizados, y es por esto que se cuenta con un código de identificación especial para cada alumno.

Las columnas del *dataset* son las siguientes:

Columna	Explicación
COD ALUMNO	Llave única que representa a un alumno.
AÑO ADMISIÓN	Año de ingreso del alumno a la universidad.
PROGRAMA CÓDIGO	Código del programa inscrito, este puede hacer referencia a Ingeniería Civil Plan Común o al Diplomado inscrito si es que tiene.
PROGRAMA	Nombre del programa inscrito (plan común o diplomado)
MAJOR CÓDIGO SELECCIONADO	Código del major inscrito por el alumno. Como solo se están analizando alumnos de los años 2015 a 2017, todos poseen majors inscritos, por lo que esta columna no podrá tomar valores NaN.
	Nombre del major inscrito por el alumno.
MINOR SELECCIONADO	Nombre del minor inscrito.
AÑO	Año en el que se cursó el ramo.
SEMESTRE	Semestre en el que se cursó el ramo. (1 o 2)
SIGLA	Sigla del ramo.
SECCIÓN	Sección cursada.
NOMBRE CURSO	Nombre del ramo.
CRÉDITOS CURSO	Créditos del curso.
NOTA FINAL	Nota final del ramo en una escala de 1 a 7.
NOTA FINAL ALFA	Nota final del ramo en una escala alfabética. Un ramo puede tener una nota en escala numérica o alfabética, en cualquier caso, una de las columnas quedará con valores NaN.

A continuación se muestra un ejemplo de fila dentro del *dataset*:

	COD ALUMNO	AÑO ADMISIÓN	PROGRAMA CÓDIGO	PROGRAMA	MAJOR CÓDIGO SELECCIONADO	MAJOR SELECCIONADO	MINOR CÓDIGO SELECCIONADO	MINOR SELECCIONADO	AÑO	SEMESTRE	SIGLA	SECCIÓN	NOMBRE CURSO	CRÉDITOS CURSO	NOTA FINAL	NOTA FINAL ALFA
0	330775	2015	40013	INGENIERÍA CIVIL	M133	MAJOR EN COMPUTACIÓN E INGENIERÍA DE SOFTWARE ...	N161	MINOR DE PROFUNDIDAD EN DATA SCIENCE Y ANALYTICS	2015	1	ING1004	8.0	DESAFÍOS DE LA INGENIERÍA	10.0	5.6	NaN

Figura 1: Ejemplo fila *dataset*

Limpieza de Datos

Para el proyecto no necesitaremos ciertas columnas por diversas razones, en el siguiente apartado se muestra cada columna que no será utilizada con su respectiva justificación:

- NOTA FINAL ALFA: Estas notas no aplican a los cursos que vamos a analizar.
- PROGRAMA: Todos los datos por analizar se encuentran en el programa Ingeniería Civil, por lo que no aporta información.
- PROGRAMA CÓDIGO: Solo se considerará el código 40013.
- SECCIÓN: No afecta en nuestro análisis.
- MAJOR SELECCIONADO: Se utilizará el código del major, por lo que esta columna ya no es necesaria. Sin embargo, no es descartada completamente, pues puede ser útil en términos de análisis de datos y futura visualización.
- MINOR SELECCIONADO: Se utilizará el código del minor, por lo que esta columna ya no es necesaria
- **Valores** NaN: Se eliminarán todas las filas con valores NaN, pues son aquellas en donde las notas de los ramos son alfabéticas, por lo tanto no aporta información al análisis.

Definiciones y consideraciones

A continuación se explicarán las definiciones y consideraciones relevantes para el desarrollo de la predicción y análisis de resultados.

- **Alumnos similares académicamente:** Los atributos a considerar para comparar alumnos serán los siguientes:

  1. `CRÉDITOS INSCRITOS`: Es relevante saber en que etapa de la carrera se encuentra el alumno, por lo que se compararán únicamente alumnos que se encuentren en un rango de créditos similar, con una variación máxima de 100 créditos. Cabe destacar, que no es necesario que los estudiantes actualmente se encuentren en la misma etapa de la carrera.
  2. `PPA`: Promedio ponderado acumulado semejante. Este se debe calcular considerando los créditos tomados. Es decir, dos personas se comportan de manera parecida si es que tienen un PPA similar con X cantidad de créditos inscritos.
  3. `MAJOR`: Que se encuentren inscritos en majors iguales o compatibles (i.e. que compartan un número de cursos mayor o igual a cierto umbral).
  4. `MAX PROMEDIO`: Que su promedio más alto no supere un umbral de diferencia entre ambos. Siempre que estos promedios sean los máximos hasta el mismo semestre para ambos alumnos.
  5. `MAX SIGLA`: Que el ramo en el que tuvieron un mejor rendimiento sea similar, es decir, que pertenezcan al mismo departamento, esto lo podemos identificar mediante las tres primeras letras del curso. (MAT, IIC, FIS, entre otros). Siempre que estos ramos sean los máximos hasta el mismo semestre para ambos alumnos.
  6. `MIN PROMEDIO`: Que su promedio más bajo no supere un umbral de diferencia entre ambos. Siempre que estos promedios sean los máximos hasta el mismo semestre para ambos alumnos.
  7. `MIN SIGLA`: Que el ramo en el que tuvieron un peor rendimiento sea similar. Siempre que estos promedios sean los máximos hasta el mismo semestre para ambos alumnos.

- **Riesgo académico:** Se definirá el riesgo académico cuando:

  $$\texttt{PPA} - \texttt{PROMEDIO FINAL}_i >= \texttt{UMBRAL}$$

  Con `UMBRAL` un número decimal positivo.

  En otras palabras, si es que la variación entre el PPA del estudiante y el promedio final en el curso i es lo suficientemente alta para disminuir su promedio, se considerará que hay un riesgo académico.

  Cabe destacar que si el $\texttt{PROMEDIO FINAL}_i$ es mayor que el `PPA`, entonces la diferencia será negativa, por lo que nunca superará el `UMBRAL`.

- **Riesgo de reprobación:** Se definirá el riesgo de reprobación cuando:

  $$\texttt{PROMEDIO FINAL}_i \sim 4{,}0$$

  Es decir, cuando el promedio en el curso i es similar o menor a 4.0.

3. Temática central del proyecto

El proyecto se centrará en predecir el riesgo académico y riesgo de reprobación de un estudiante en base a los ramos que desea tomar en un semestre. Considerando su PPA, créditos inscritos hasta el momento,

sigla y promedio del mejor y peor curso que ha tenido hasta el momento, major inscrito y créditos inscritos en el semestre que se busca predecir. Cabe destacar que estos parámetros pueden cambiar dependiendo del rendimiento de los modelos. De esta manera, se puede ayudar a prevenir casos de reprobación de cursos, o bien, advertir que la carga académica de cierta combinación de ramos puede traer riesgos académicos en un semestre. Esto puede alivianar el estrés futuro que puedan sufrir ciertos estudiantes por inscribir una combinación de cursos que puedan perjudicar a su rendimiento académico y salud mental.

Cabe aclarar, que el estudiante que quiere consultar, deberá ingresar a la interfaz únicamente los nombres de 3 ramos en los que quiera predecir su rendimiento académico al tomarlos juntos, esto porque si se ingresa una cantidad mayor se reduciría mucho el dataset de predicción, lo que haría que finalmente no se pueda asegurar un resultado. Lo anterior también se justifica en base a que es común que se tomen 3 ramos de alta dificultad y 2 ramos de mediana o baja dificultad para equilibrar la carga académica.

Para abordar el problema se decidió utilizar un proceso de filtrado de datos y posteriormente el modelo de predicción KNN , en donde se transformará cada estudiante en un vector del espacio, para luego encontrar a sus vecinos más cercanos y de esta manera comparar su rendimiento en los ramos que se buscan.

Ejemplo de dataset construido previo al proceso:

```
('FIS1523',
'MAT1630',
'MAT1640',
'IIC2233',
'TTF027',
5.8,
6.0,
4.5,
5.4,
7.0,
ppa= 5.6,
mejor promedio hasta este semestre = 6.6,
peor promedio hasta este semestre = 4.1,
major inscrito = 'MAJOR COMPUTACIÓN',
sigla mejor promedio = 'MAT1610',
sigla peor promedio = 'LET003',
creditos inscritos hasta el semestre = 100,
creditos inscritos en el semestre = 50)
```

Este procedimiento constará de 2 etapas:

i) Filtrar a los individuos que no sean iguales en ciertas columnas específicas:

Como ejemplo, si un alumno quiere predecir cómo será su rendimiento acádemico si toma los ramos IIC2233, MAT1630 y MAT1640, entonces se filtrarán en el dataset todos los alumnos que hayan incorporado estos ramos en su semetre, dejando afuera los que no tienen esos ramos en el semestre. Además, cabe la posibilidad de que filtremos por otros atributos que consideremos relevantes, como por ejemplo por la cantidad de créditos que tomó el alumno en el semestre, esto puede variar dependiendo de cómo sea el rendimiento de nuestro modelo.

Una vez que se aplique el proceso de filtrado para el ejemplo mostrado quedará el siguiente vector:

```
('FIS1523',
'MAT1630',
'MAT1640',
5.8,
6.0,
```

```
4.5,
ppa= 5.6,
mejor promedio hasta este semestre = 6.6,
peor promedio hasta este semestre = 4.1,
major inscrito = 'MAJOR COMPUTACIÓN',
sigla mejor promedio = 'MAT1610',
sigla peor promedio = 'LET003'
```

Los atributos eliminados fueron: créditos inscritos, créditos inscritos hasta el semestre, dos ramos y sus respectivos promedios.

ii) Con los individuos que pasan el filtro se realiza la búsqueda de vecinos más cercanos por KKN:

En el caso del ejemplo, el vector utilizado para la búsqueda será la que llegó del filtrado.

Cabe mencionar que los parámetros que se están ocupando para el filtrado y para el KNN podrán cambiar según el avance del proyecto. Dependerá de las relaciones que se vayan encontrando.

En esta misma línea, se generarán distintos modelos en base a los atributos previamente descritos, es decir, se generarán modelos con distintas combinaciones de atributos y ponderaciones, para luego evaluar su rendimiento, comparar los resultados y encontrar el modelo que mejor se adapte a los requerimientos del proyecto. Junto con esto, también se evaluarán distintas métricas de distancia para encontrar aquella que de mejores resultados al momento de predecir.

4. Trabajo pendiente

El trabajo que nos queda por delante es normalizar los datos y separarlos en set de entrenamiento, validación y test, luego crear los diferentes modelos de KNN y entrenarlos en base a los distintos parámetros mencionados anteriormente. Además, debemos evaluar el rendimiento de cada uno de ellos y encontrar el mejor predictor.

En conjunto a lo anterior, se generarán visualizaciones para apoyar el entendimiento de las conclusiones a las que lleguemos.

Finalmente, se planea crear un sitio que permita introducir ciertos cursos, y recibir una predicción sobre el resultado