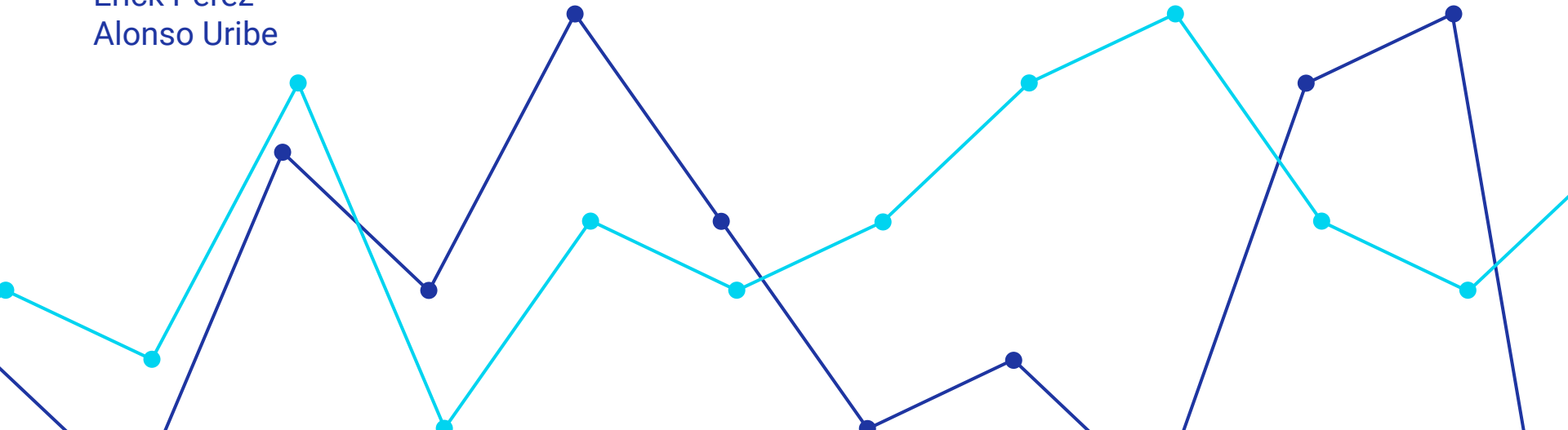
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

Análisis de Matrículas Educación Superior en Chile

Grupo 6
Javiera Alegría
Adriana Concha
Erick Pérez
Alonso Uribe



Contexto y motivación

Estudiar y trabajar con datos de **ingreso a la educación superior** puede ser muy útil para **predecir** las características de las **futuras matrículas**, como también , para **identificar problemas** actuales como futuros, lo cual permitiría tomar medidas de manera anticipada.

Esquema de Registro Matrícula de Educación Superior

Se cuenta con datos desde 2010 a 2019, con un archivo por año.

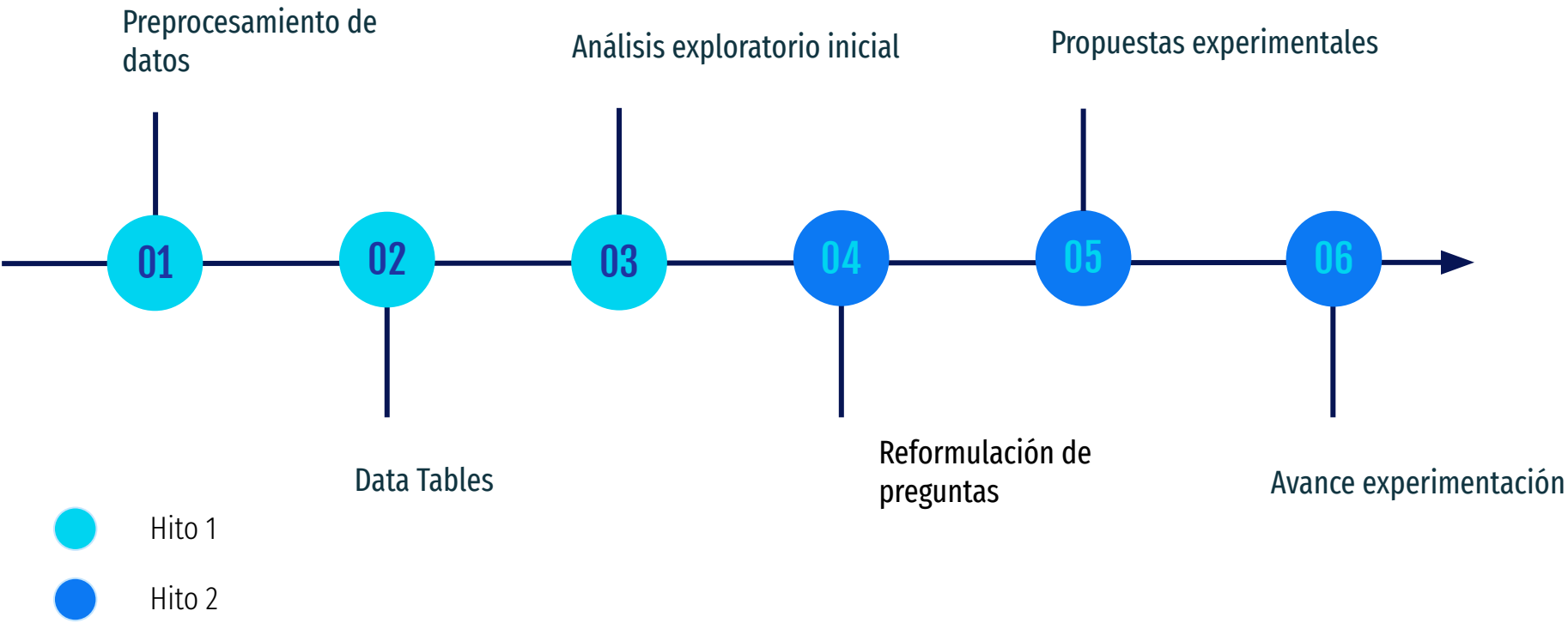
Fuente de la información

Servicio de Información de Educación Superior (SIES).

Cantidad de registros y variables

12.070.384 registros en total (aprox. 1,2 millones por año), y 49 columnas.

Roadmap


Preprocesamiento de datos
Análisis exploratorio inicial
Propuestas experimentales
01
02
03
04
05
06
Data Tables
Reformulación de preguntas
Avance experimentación
Hito 1
Hito 2

¿Se puede predecir las características de las matrículas para años futuros considerando los datos actuales?

Definir las características de las matrículas y encontrar estructuras en una colección de instancias.

Investigación

- Particularmente nos quedamos con características macroscópicas que den cuenta de su naturaleza, a través de los años, ej: su precio promedio(regional, nacional, etc.). Se propone encontrar estructuras de clustering que den cuenta de ello.

Preprocesamiento

- Nueva selección de columnas relevantes. Por instancia se crea una concatenación de vectores One-hot por columna, por lo tanto, el módulo del vector, correspondiente a una instancia, es sqrt(número de columnas).

Análisis

- Transformar las instancias a vectores haciendo un embedding que represente un espacio rico en información.
- Generar clustering jerárquicos, visualizarlos y validarlos con Coeficiente de Silhouette, Cohesión, Separación e Inercia.
- Reducir dimensionalidad con métodos como PCA, validar con Silhoutte y visualizar la estructura.

¿Podemos predecir quien se cambiará de carrera?

Se busca entrenar un clasificador capaz de identificar a una persona que se cambiará de carrera, para ello se planea:

Preprocesamiento

- Eliminar 35 columnas que no aportan información útil.
- Aplicar restricciones sobre los datos, para de dejar solo los datos relevantes.
- Usar OneHot para encontrar correlaciones y eliminar columnas redundantes.

Análisis

- Utilizar la técnica Holdout con 30% de datos de entrenamiento, dada la cantidad de datos, estratificando las clases; “sí se cambiará” y “no se cambiará”.
- Probar con diferentes clasificadores (KNN, Naive Bayes, SVM, Árbol de Decisión) y técnicas de balance de clases (oversampling, subsampling, normal), para mantener la configuración que presente mejor desempeño.
- Analizar desempeños a través de la matriz de confusión, en especial de la métrica recall.

¿Cuánto valor le agrega la acreditación a una carrera?

Preprocesamiento

- Se seleccionan 14 columnas de las 49 del dataset original, con información relevante, tanto categóricas como numéricas.
- Se modifican 2 columnas para facilitar el manejo de datos.
- Se crean 3 nuevas variables con información relevante.
- Se elimina el 1% de los datos por valores incompletos.

Visualizaciones

- Se generan histogramas de todas las columnas.
- Se obtienen métricas de los histogramas.
- Se generan gráficos de la evolución temporal de las métricas anteriores.
- Visualización de clusters utilizando PCA, t-sne, UMAP.

Análisis

- Se cuantifica el valor agregado de la acreditación en una carrera, dado los resultados anteriores.

¿Existe una relación entre costo de carrera y otras variables?

Experimentación

- Utilización de método descriptivo Clustering con el objetivo de:
  - Relacionar variables valor del arancel, duración de la carrera y años de acreditación.
  - Relacionar variables modalidad y el tipo de jornada con el costo de carrera, realizando una codificación numérica de forma previa.
- Pruebas de técnicas K-Means y Clustering jerárquicos.

Validación

- Validar resultados utilizando métrica SSE, Matriz de Similitud y coeficiente de Silhouette.

Visualización

- Visualización de resultados a través de scatterplot del clustering y dendrogramas para identificar outliers y realizar post-procesamiento de datos.

Implementación

En este hito se implementó la segunda pregunta; ¿Podemos predecir quien se cambiará de carrera?, siguiendo la metodología planteada anteriormente.

Tras probar los clasificadores KNN, Naive Bayes, SVM y Árbol de Decisión con los datos de un solo año, se observó que KNN y SVM presentaban malas métricas y requerían muchos recursos. Así, solo se observan Naive Bayes y Árbol de Decisión con todos los datos.

El clasificador final utiliza Árbol de Decisión y Oversampling, al ser la combinación con mejor desempeño en las métricas relacionadas al recall de la clase verdadera, es decir, “Sí, se cambia de carrera”.

Con los resultados obtenidos podemos predecir satisfactoriamente quien se cambiará de carrera, y generar políticas que ayuden o guíen a la sociedad estudiantil de forma informada.

Conclusiones y desafíos futuros

- Se reformularon las preguntas de investigación a partir del feedback del Hito 1.
- Resultados preliminares utilizando técnicas de minería de datos. Particularmente, se encontraron resultados satisfactorios para la pregunta 2, para predecir quienes se cambiarán de carrera.
- Se espera terminar de implementar las propuestas metodológicas para todas las preguntas durante el Hito 3.