Universidad Tecnológica Nacional

Facultad Regional Córdoba

Ingeniería en Sistema de información

Ciencia de Datos

TPI

“Primera Entrega”

Grupo 3

Integrantes:

- **82217 Anzulovich, Valentina - valenanzu.utn@gmail.com**
- **80359 Berrotaran, María Luz - luzberrotaran@gmail.com**
- **78678 Salas, Bruno Matías - brunosalas1@hotmail.com**
- **82508 Sonzini, Enrique - enriquesonzini@gmail.com**
- **81841 Yarbouh, Yamili - yamiliyarbouh@gmail.com**

Docentes:

- **Pablo Alberto Sacco**
- **Franco Mana**
- **Marisa del Carmen Callejas**

Curso: 5K2

Fecha de presentación: 18/09/2024

Índice

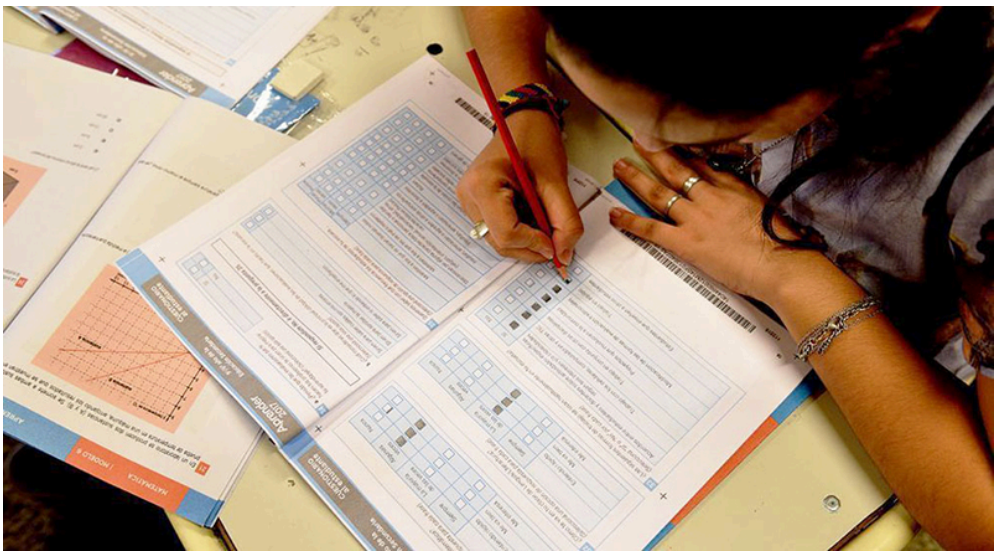
Introducción

Para este trabajo nos hemos puesto como objetivo presentar el dataset seleccionado. En nuestro caso estamos hablando de los datos provenientes de las pruebas Aprender de los años 2019 y 2022, especificando por qué y cómo llegamos a elegirlos. Además vamos a detallar todo el proceso que fue necesario para unir las tablas correspondientes y unificar sus contenidos.

Desarrollo

Para empezar fue fundamental elegir el tema y objetivo del trabajo. Se nos ocurrieron varias ideas, como la relación entre el uso de pantallas y los problemas de vista o sueño, luego decidimos enfocarnos en algo sobre lo que tuviésemos conocimientos como lo es el estudio.

Buscando datasets, nos encontramos con las pruebas Aprender donde encontramos una gran variedad de datos y tuvimos que determinar cuál podría ser el valor a predecir. En este caso nos decidimos por dos valores, que fueron las notas obtenidas por los estudiantes de secundario en matemática y en lengua en el examen.



El dataset de las pruebas Aprender tiene la respuesta de los alumnos a preguntas de distintos caracteres, como pueden ser factores socio-económicos, que actividades hace el chico fuera de la escuela, con quienes vive, qué responsabilidades tiene, entre otras. Con esto podríamos tratar de predecir el rendimiento que tendría un determinado alumno en las materias de las que tenemos el resultado. Las materias se evalúan con una base de 72 ítems distribuidos en diferentes áreas de conocimiento. Para el puntaje utiliza la metodología TRI estableciendo la media en 500 y una desviación estándar de 100. Sobre esto establece 4 categorías: Por debajo del básico, Básico, Satisfactorio, Avanzado.

Una vez definido el objetivo había que determinar cuáles iban a ser los datos a trabajar y para eso era necesario definir cuáles de las tablas disponibles iban a ser nuestro conjunto

de análisis. Optamos por elegir las encuestas que apuntaban al nivel secundario y surgió la posibilidad de trabajar con 4 años diferentes.

Teniendo entre nuestras opciones los años 2016, 2017, 1019 y 2022 se abrió un abanico de posibles combinaciones para que fueran nuestro set de datos. De las cuales, tras un análisis exhaustivo terminamos eligiendo la de los años 2019 y 2022. Esto fue así porque además de ser datos más actuales y suficientes para hacer el trabajo, presentaban muchas coincidencias en las columnas que compartían.

Debíamos definir cuál sería la metodología con la que abordariamos el trabajo, definiendo el plan de proyecto. Utilizando SCRUM definimos que íbamos a hacer un sprint por entrega del TP. Después nos planteamos un posible horario para realizar daylis en caso de ser necesario, terminamos definiendo que las íbamos a hacer antes de las clases porque eran los momentos en que nos podíamos reunir todos, por lo tanto haremos 2 “daylis” por semana. Además creamos un board en Miro para la organización de las tareas a realizar y quién será el encargado de cada una.

Para poder llevar a cabo la comparación de las tablas fue necesario para comenzar poder identificar los datos con los que contábamos, ya que las referencias con las que contaban no estaban disponibles de forma directa. Fue necesario hacer una traducción de columnas de cada tabla a un código más representativo.

Una vez hecho esto, ya teníamos las preguntas que hacían referencia a lo mismo con nombres iguales, pero fue ahí cuando notamos que no eran realmente iguales. Es decir los valores de codificación no significaban los mismo en la tabla de 2019 que en la de 2022. Nos dimos cuenta que iba a ser necesaria una transformación de los valores en los casos que no hubiera coincidencia.

Planteamos los siguientes objetivos para analizar en el dataset:

- Predecir el puntaje de lengua y matemática del alumno.
- Analizar si el hecho de que el alumno trabaje afecta a su rendimiento.
- Análisis de rendimiento por región.
- Análisis de rendimiento de acuerdo a con cuántas personas vive.

Conclusión

Actualmente contamos con un dataset que tiene por objetivo predecir las notas que puede obtener un alumno en matemática o lengua en el examen Aprender. Todavía falta para que este set de datos sea de utilidad, y va a ser necesario hacer algunos ajustes respecto a los valores correspondientes a los distintos años. Pero tenemos la esperanza de que una vez tratados los datos vamos a obtener predicciones valiosas a partir de ellos.