Sistemas de Información para el Análisis de GVD
Proyecto - Curso 2024

Análisis de datos de EDUCACIÓN

1 Introducción

En este proyecto se propone implementar un sistema que permita analizar datos sobre algunos aspectos de la Educación en Uruguay, los cuales se encuentran disponibles en la Web. Se trabajará con algunos datasets extraídos del Catálogo Nacional de Datos Abiertos de Uruguay [1], el cual permite acceder a datos abiertos de organismos públicos, academia, organizaciones de sociedad civil y empresas privadas de Uruguay. Algunos de los datasets fueron levemente modificados para ser más interesantes a efectos de este proyecto.

Se analizarán datos de ofertas educativas de educación secundaria y educación primaria, incluyendo datos sobre los liceos y las escuelas existentes en cada ubicación, y las cantidades de docentes y estudiantes que hay en cada departamento de Uruguay. Por otro lado, se analizará el uso que los docentes y los estudiantes les dan a la plataforma informática CREA [2], perteneciente al Plan Ceibal, así como el uso de la biblioteca.

Teniendo en cuenta distintos perfiles de personas u organizaciones, que puedan estar interesados en analizar los datos de Educación, se desea construir una plataforma de análisis de datos que integre todas las fuentes de datos consideradas en este proyecto. Dicha plataforma debe proveer información confiable a través de una interfaz amigable y flexible orientada a la toma de decisiones.

La plataforma de análisis tendrá como objetivo fundamental el permitir realizar análisis multidimensional de la información, de tipo Business Intelligence (BI), y opcionalmente aplicar algoritmos de aprendizaje automático para deducir otra información de interés.

2 Objetivos del proyecto

El objetivo general de este proyecto es implementar una plataforma de datos sobre la cual se construya un DW, una interfaz multidimensional (cubos de datos) que se cargue a partir del DW y de datos curados no cargados en el DW. Por otro lado (opcionalmente), interesa que se transformen algunos datos fuentes para aplicar luego alguna técnica de aprendizaje automático.

Objetivo 1

El objetivo principal de este proyecto es realizar un análisis multidimensional sobre ciertos aspectos de los datos fuentes.

Para alcanzar este objetivo se deberá:

1- Realizar un **diseño conceptual multidimensional** completo que surja del análisis de los requerimientos, generando el **modelo conceptual**.

2- Diseñar e implementar un **modelo lógico** que dé soporte al modelo conceptual desarrollado en el punto 1, teniendo en cuenta las restricciones impuestas por las herramientas a utilizar.

3- Diseñar e implementar los **procesos de carga** del modelo lógico utilizando *Pentaho Data Integration* (también conocido como *Kettle*).

4- Implementar los cubos y el *front end* del sistema, correspondientes a los puntos 1 y 2.

Objetivo 2

El segundo objetivo, que solamente será para los estudiantes que tomen esta opción, es realizar una prueba de concepto en donde se prepararán ciertos datos de las fuentes para predecir información a partir de ellos, utilizando alguna técnica de aprendizaje automático.

3 Requerimientos

En esta sección se describen los requerimientos funcionales y no funcionales de la solución a desarrollar.

3.1 Requerimientos funcionales

A continuación, se presentan los requerimientos funcionales, que son requerimientos de análisis de la información, que se deben satisfacer en el proyecto.

Requerimiento 1:

Se desea analizar la información sobre las Ofertas de Educación Secundaria, es decir los distintos tipos de formaciones que un estudiante podría cursar en un determinado año o semestre. **Interesa ver los datos de cada oferta educativa, agrupados según distintos criterios, como el nivel, la opción, el liceo, la ubicación, y otros.** Según estas perspectivas se quieren estudiar indicadores de cantidades y porcentajes. A continuación, se describe en detalle cada uno de estos aspectos.

La información debe ser analizada según el liceo, la ubicación, el plan, el nivel, la modalidad y el turno.

La ubicación se refiere a la ubicación del liceo y se debe obtener a partir de la dirección del mismo. Se deben obtener las coordenadas geográficas (latitud, longitud), y agrupar según zona, localidad, departamento y zona-país. La zona se determina por el código postal de la ubicación, mientras que

la zona-país se refiere a “norte”, “sur”, “este” y “oeste”. También se quiere poder agrupar las ubicaciones según el área, la cual sería “rural” o “urbana”. Si fuera posible (utilizando alguna otra fuente de datos), también es deseable manejar un nivel de agregación más fino que el de zona, correspondiente a la noción de barrio.

El plan hace referencia al plan de estudios al que corresponde la oferta educativa. Se quiere poder ver el nombre del plan, pero además verlos agrupados según la franja de edad para la que está propuesto (extra-edad o no), según si es una reforma/adecuación o si es un plan nuevo, y según la época en que fue aprobado (antes del 2000 o después del 2000).

El nivel que se quiere analizar es el grado, que debe poder agruparse por ciclo (CB o Bachillerato).

La opción que se quiere analizar solo tiene sentido en el caso del bachillerato diversificado, para el caso de ofertas que no son de bachillerato diversificado se puede manejar un valor especial.

La modalidad se refiere a si la oferta es anual o semestral.

El turno se refiere al turno en el que se dicta la oferta. Los turnos se quieren poder ver agrupados en horarios y los horarios en franjas-horarias. La franja horaria podría ser “matutina”, “vespertina” o “nocturna”, o se podría proponer algún otro criterio que resulte adecuado para agrupar a los horarios.

De la información según todas las perspectivas descriptas, interesa analizar la cantidad de liceos, la cantidad de ofertas educativas, el porcentaje de liceos, el porcentaje de estudiantes por departamento y el porcentaje de docentes por departamento. Estas dos últimas medidas se quieren promediar al subir al nivel zona-país.

Requerimiento 2:

Se desea analizar la información sobre las Escuelas que imparten educación inicial y primaria. Interesa ver los datos de cada escuela, agrupados según distintos criterios, como la ubicación, la categoría, los niveles, y otros. Según estas perspectivas se quieren estudiar distintos indicadores de cantidades. A continuación, se describe en detalle cada uno de estos aspectos.

La información debe ser analizada según la escuela, la ubicación, la categoría, el nivel más bajo, el nivel más alto, y el turno.

La escuela está identificada por su nombre. Se quiere poder ver a las escuelas agrupadas según su tipo, el cual se refiere a si es una escuela especial, común o jardín de infantes.

La ubicación se refiere a la ubicación de la escuela y se debe obtener a partir de la dirección de esta. Al igual que en el caso de los liceos, se deben obtener las coordenadas geográficas (latitud, longitud), y poderse agrupar según zona, localidad, departamento y zona-país. La zona se determina por el código postal de la ubicación, mientras que la zona-país se refiere a “norte”, “sur”, “este” y “oeste”. También se quiere poder agrupar las ubicaciones según el área, la cual sería “rural” o “urbana”. Si fuera posible (utilizando alguna otra fuente de datos), también es deseable manejar un nivel de agregación más fino que el de zona, correspondiente a la noción de barrio.

La categoría se refiere a una categorización que se hace de las escuelas por parte de la administración de estas. Estas categorías básicas se quieren analizar según 2 agrupaciones: si son escuelas especializadas en alguna discapacidad o no, y si son escuelas de práctica o no.

El nivel más bajo es el nivel más bajo que tiene la escuela, es decir el nivel desde el cual parte la formación de los alumnos. Estos niveles se tienen que poder agrupar en ciclos, los cuales son "inicial" y "primaria".

El nivel más alto refiere al nivel más alto que alcanzan los niveles que tiene la escuela, y se trata en forma análoga a la anterior.

El turno refiere a los turnos que ofrece la escuela. Los turnos se deben poder agrupar por tipo de turno ("corto" o "largo"), y por franja horaria ("matutino", "vespertino" o "ambos").

De la información según todas las perspectivas descriptas, interesa analizar los siguientes indicadores: la cantidad de aulas, la cantidad de grupos por nivel, la cantidad de escuelas y la cantidad de niveles. Interesa agregar estas medidas como sumas y como promedios.

Requerimiento 3:

Se quiere analizar los datos de Centros Educativos en las distintas ubicaciones.

Interesa analizar la cantidad de centros según la ubicación, considerando todos los niveles de agregación posible, según el subsistema (si es Primaria o Secundaria), y según los turnos generales. Los turnos generales en este caso se refieren a si es matutino, vespertino, nocturno, o extendido.

Interesará ver la cantidad promedio de centros que hay en cada ubicación, discriminando por turnos y subsistemas. Por ejemplo, se quiere analizar en cada departamento, cuántas escuelas hay en promedio por localidad, que tengan horario extendido.

Atención: Este requerimiento deberá resolverse con un cubo implementado en el *modelo de datos documental*. Esto se explica en detalle en la **Sección 5**.

Requerimiento 4:

Este requerimiento tiene **2 opciones** de las cuales es obligatorio realizar **solamente una de ellas**.

Opción 1:

Se desea analizar los datos de actividades de los estudiantes y de los docentes.

Interesa analizar la cantidad de ingresos a CREA, la cantidad de comentarios, la cantidad de ingresos a biblioteca y la cantidad de préstamos pedidos a biblioteca. Además, interesa ver el indicador de cantidad de actividades total del estudiante o docente.

Se quiere estudiar estos indicadores según el año, el sexo, el rol (docente o estudiante), el subsistema, la ubicación y el nivel.

La ubicación se debe manejar con los niveles de agrupación que permitan los datos fuentes con los que se dispone.

Opción 2:

Utilizando como mínimo los mismos datos fuentes requeridos para la Opción 1, entrenar un modelo de aprendizaje automático (aplicando una técnica de su elección), que permita predecir alguna de las variables analizadas.

3.2 Requerimientos no funcionales

La solución deberá desarrollarse utilizando versiones estables de las herramientas que se presentan a continuación. Las versiones que se sugieren pueden ser sustituidas por versiones superiores si los estudiantes así lo desean.

3.2.1 Pentaho Business Intelligence (Pentaho BI). Los productos a utilizar de la versión *Community* son:

- *Pentaho BI Analytic* [3]: la plataforma *Pentaho* se basa en una aplicación web J2EE que permite publicar y gestionar soluciones, en un servidor que las implementa. Cada solución puede verse como una aplicación web que utiliza los diferentes servicios provistos por el servidor *Pentaho* (por ejemplo: motor OLAP, motor de workflow, etc.), que presenta la información al usuario mediante diferentes componentes (por ejemplo: reportes y gráficas dinámicas, vistas de análisis OLAP sobre cubos, tableros, diales con indicadores, etc.). Se sugiere utilizar la versión 9.4 de *Pentaho BI*.
- *Pentaho Analysis Services* [4]: Herramienta también llamada *Mondrian*, es un servidor OLAP del tipo ROLAP. Los modelos multidimensionales que son interpretados por *Mondrian* consisten en archivos XML denominados *Schemas*, en ellos se definen los cubos, dimensiones, niveles, jerarquías, etc. Además, es posible realizar consultas sobre los cubos OLAP, para esto se definen expresiones multidimensionales (MDX, *MultiDimensional eXpressions*) [5]. MDX es un lenguaje de consulta para bases de datos multidimensionales.
- *Pentaho Schema Workbench* [6]: Es posible crear *Schemas* de *Mondrian* utilizando la herramienta gráfica para el diseño de *Schemas* denominada *Schema Workbench*. Esta herramienta permite definir esquemas en *Mondrian* y luego publicarlos.
- *Pentaho Data Integration, Kettle* [7]: Esta es la herramienta de ETL del proyecto *Pentaho*. Se sugiere utilizar la versión 9.4.
- Documentación y tutoriales. En particular se sugiere documentación de *Pentaho BI Analytic* [3], de *Mondrian* [4] y de *Kettle* [7]. La plataforma utiliza RDBMs para almacenar la

información del sistema (usuarios, roles, etc.) y para almacenar datos. Consulte la documentación de *Pentaho* [3] para saber cuáles son los RDBMs soportados y cómo se configura la conexión [8]. Para este proyecto, como RDBMs, se sugiere utilizar *PostgreSQL 16.2* [9]. Además, se sugiere *PgAdmin4* [10] para gestionar de bases de datos *PostgreSQL*. Por otro lado, no hay restricciones respecto al sistema operativo sobre el cual debe correr el prototipo, queda a elección del estudiante.

- Componente *Visualizer* [11]: Este es un plug-in para *Pentaho*, de uso libre, que permite hacer consultas de tipo OLAP y crear *dashboards* con gran facilidad. También se puede utilizar, en forma opcional, el componente *Saiku* [12] (versión no *Enterprise,* también plug-in para *Pentaho*), que permite hacer consultas OLAP y mostrar los resultados de forma tabular o mediante gráficas.
- *Pentaho Report Designer* [13] para generar reportes que se publican en la plataforma de *Pentaho*.
- *Community Dashboard Editor* (CDE) [14]. La utilización de esta herramienta es opcional, ya que con la herramienta *Visualizer* pueden obtener las principales funcionalidades que esta provee (la incluimos en la lista por si les interesa explorarla). CDE Permite construir *Dashboards* para destacar indicadores relevantes para la toma de decisiones a nivel gerencial. Para un buen diseño de los *dashboards* se sugiere aplicar conocimientos de HTML, CSS y JavaScript.

3.2.2 Otras herramientas: Para el desarrollo del proyecto, también se podrán utilizar las siguientes herramientas:

- *MongoDB* [15][16]: Base de datos orientada a documentos. Su utilización es opcional.
- *json* [17]: Formato de intercambio de datos, basado en javascript
- *Tableau* [18], *PowerBI* [19]: Herramientas para la visualización y el análisis de los datos. Se deberá utilizar una de estas dos herramientas.
- *Python, Jupyter Notebook* [20]: Posibles herramientas para implementar un modelo de aprendizaje automático para el Requerimiento 4, Opción 2.

4 Fuentes de Datos

Las fuentes de datos necesarias para la realización del proyecto serán provistas o referenciadas en la plataforma EVA del curso.

5 Etapas del Proyecto

El proyecto se realizará en varias etapas y cada una implica las tareas obligatorias que se detallan a continuación.

- **Etapa 1: Diseño conceptual de los Requerimientos**. Se debe realizar el diseño conceptual utilizando el modelo CMDM e incluyendo el análisis de aditividad de las medidas.

    - Entrega: **18 de abril**

- **Etapa 2: Diseño lógico relacional para Requerimientos 1, 2 y 4**. Se debe realizar el diseño lógico del DW relacional para estos requerimientos.

  Para los que hagan la **Opción 2** en el **Requerimiento 4**:
  Se deberá entregar una descripción del modelo de aprendizaje automático que se va a construir, así como de los datos a utilizar para su implementación.

    - Entrega: **30 de abril**

- **Etapa 3: Diseño para el Requerimiento 3.**

  Este diseño lógico deberá construirse en base al artículo *Document-oriented Models for Data Warehouses. NoSQL Document-oriented for Data Warehouses* [21]. Este requerimiento puede ser implementado a través de json o mediante el uso de MongoDB (podrán optar por una de las 2 opciones).

    - Entrega: **9 de mayo**

- **Etapa 4: Carga de los cubos y archivos de datos para el análisis**. De forma obligatoria, en esta etapa, se debe utilizar la herramienta *Kettle (PDI)*.

- **Etapa 5: Implementación del *Front End*.** La implementación debe incluir el uso de las siguientes herramientas de análisis OLAP, creación de *dashboards*, y reportes:

  (1) *Visualizer* o *Saiku* para **Requerimientos 1**, **2 y 4**. Se sugiere mostrar algunos de los datos en el front-end utilizando la visualización en mapas.

  (2) *Power BI* o *Tableau*, para **Requerimiento 3**. Opcionalmente, pueden utilizarse para los otros requerimientos.

  (3) *Pentaho Report Designer* para al menos un reporte de alguno de los requerimientos.

Para el Requerimiento 4, opción 2, se sugiere la utilización de notebooks y Python, con las bibliotecas necesarias para las técnicas de aprendizaje automático elegidas.

- **Etapa 6: Documentación.** Se deberá documentar todo lo realizado según las pautas presentadas en la Sección 6.
  El proyecto e informe serán presentados por los estudiantes en una defensa final.
  - **Entrega del Informe Final: 27 de junio**
  - **Defensa del Proyecto: Fecha a confirmar**

6 Informe Final

El informe final debe presentar una descripción completa de la solución propuesta. Por lo tanto, éste **debe incluir, de forma obligatoria, todas las secciones** que se presentan a continuación:

1. **Introducción**. El informe debe contener una sección de introducción que permita a los lectores saber de qué trata el documento y conocer la realidad del problema para el cual se presenta una solución.

2. **Análisis de requerimientos**. En esta sección se debe presentar una breve descripción de los requerimientos analizados.

3. **Diseño conceptual**. En esta sección se debe presentar el diseño conceptual completo para cada uno de los requerimientos del proyecto, por lo tanto, esto incluye el análisis de los problemas de aditividad.

4. **Diseño lógico**. En esta sección se debe presentar el diseño lógico correspondiente a cada uno de los Requerimientos.

5. **Implementación**.
   a. Para los Requerimientos 1, 2 y 4 (opción 1) se debe presentar la implementación de las dimensiones y de las relaciones dimensionales sobre *Pentaho BI Server* (archivos .xml generados).

   b. Para el Requerimiento 3, se debe presentar la implementación de las dimensiones y de las relaciones dimensionales en el modelo de datos documental, utilizando las herramientas específicas que se definirán en el transcurso del proyecto.

   c. Para el Requerimiento 4, opción 2, se debe presentar los detalles acerca del modelo de aprendizaje automático implementado y los resultados obtenidos.

6. **Proceso de carga**. Para todos los Requerimientos, en esta sección se debe presentar una breve descripción del proceso de carga, destacando los problemas abordados que resultaron más interesantes.

7. **Front-End.** Documentación sobre los distintos front-end utilizados, incluyendo imágenes ilustrativas que deben ser descritas en el cuerpo del informe.

8. **Capacidad de soportar de nuevas cargas (opcional)**. Esta sección debe presentar un breve análisis de la capacidad que presenta, la solución implementada, de soportar nuevas cargas de datos.

9. **Calidad de datos**. En esta sección se debe presentar una descripción de los problemas de calidad de datos encontrados durante la resolución del proyecto. Además, se debe incluir las soluciones propuestas para la resolución de dichos problemas de calidad.

10. **Testing (opcional)**. Descripción del plan de testeo realizado sobre la solución implementada.

11. **Conclusiones**. Se deben presentar las conclusiones que se consideren relevantes acerca de la solución propuesta. Las conclusiones pueden incluir reflexiones correspondientes a cada una de las etapas del proyecto, como así también conclusiones generales que surjan del análisis de los datos. Por otro lado, también pueden ser incluidas conclusiones sobre las herramientas, el proyecto y el curso.

7 Referencias

[1] https://catalogodatos.gub.uy/
[2] https://ceibal.edu.uy/plataformas-y-programas/crea/
[3] https://www.hitachivantara.com/content/dam/hvac/pdfs/datasheet/lumada-dataops-pentaho-business-analytics-datasheet.pdf
[4] http://mondrian.pentaho.com/documentation/olap.php
[5] https://mondrian.pentaho.com/documentation/mdx.php
[6] https://mondrian.pentaho.com/documentation/workbench.php
[7] https://docs.hitachivantara.com/r/en-us/pentaho-data-integration-and-analytics/9.4.x/mk-95pdia003
[8] https://docs.hitachivantara.com/r/en-us/pentaho-data-integration-and-analytics/9.4.x/mk-95pdia003/use-a-pentaho-repository-in-pdi/unsupported-repositories
[9] PostgreSQL: https://www.enterprisedb.com/downloads/postgres-postgresql-downloads
[10] https://www.pgadmin.org/
[11] https://master.dl.sourceforge.net/project/datafor-visualizer/workshop/Plugins/Visualizer%20Setup%20Manual.pdf?viasf=1
[12] https://eva.fing.edu.uy/mod/url/view.php?id=107013
[13] https://docs.hitachivantara.com/r/en-us/pentaho-data-integration-and-analytics/9.4.x/mk-95pdia008/using-pentaho-report-designer
[14] https://docs.hitachivantara.com/r/en-us/pentaho-data-integration-and-analytics/9.4.x/mk-95pdia006/ctools/general-overview-of-pentaho-and-ctools/cde-community-dashboard-editor
[15] MongoDB: https://www.mongodb.com/try/download/community
[16] Tutorial de MongoDB: https://www.mongodb.com/docs/manual/administration/install-community/#std-label-install-mdb-community-edition
[17] JSON: https://www.json.org/json-es.html
[18] https://www.tableau.com/es-es/products
[19] https://powerbi.microsoft.com/es-es/downloads/
[20] https://jupyter.org/
[21] Chevalier, M.; El Malki, M.; Kopliku, A.; Teste, O. and Tournier, R. (2016). Document-oriented Models for Data Warehouses - NoSQL Document-oriented for Data Warehouses. In Proceedings of the 18th International Conference on Enterprise Information Systems - Volume 1: ICEIS, ISBN 978-989-758-187-8, pages 142-149. DOI: 10.5220/0005830801420149. http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0005830801420149