DESARROLLO DE SOLUCIONES DE INTELIGENCIA DE NEGOCIOS PARA EL CENTRO DE INNOVACIÓN EN TECNOLOGÍA Y EDUCACIÓN DE LA UNIVERSIDAD DE LOS ANDES, CONECTA-TE

PROYECTO DE GRADO

HUGO SANTIAGO HERNANDEZ LIMAS

UNIVERSIDAD DE LOS ANDES

ASESOR: HAYDEMAR MARÍA NÚÑEZ CASTRO

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

BOGOTÁ, COLOMBIA

2020

Resumen

Este proyecto contiene el desarrollo de una solución de inteligencia de negocios para el Centro de Innovación en Tecnología y Educación (Conecta-TE) de la Universidad de los Andes. Se tiene como objetivo determinar la efectividad en el proceso de aprendizaje de los diferentes recursos ofrecidos a los estudiantes en los cursos blended.

La finalidad del proyecto es realizar un perfilamiento de los estudiantes de acuerdo con el comportamiento que tiene en la plataforma y la evolución de su proceso académico. A través de esto se apoya los procesos de construcción y formulación de cursos blended enfocados en las necesidades de los estudiantes.

Con lo anterior, se busca brindar apoyo para el diseño, mejora y evaluación de cursos blended con el fin de ofrecer un mejor servicio para los estudiantes que les permita aprender de manera más efectiva y autodidacto alineando el proceso a los pilares de Conecta-TE.

INDICE

1. INTRODUCCIÓN

El grupo de trabajo del curso de Inteligencia de Negocios del departamento de Ingeniería de sistemas y computación ha estado trabajando de la mano del Centro de Innovación en Tecnología y Educación (Conecta-TE) de la Universidad de los Andes para diseñar y desarrollar proyectos que permitan apoyar los procesos de negocio. Gracias a esto, nace la iniciativa de crear proyectos de inteligencia de negocios, como el del presente documento, orientados al análisis de los datos asociados a los estudiantes y a los cursos.

Conecta-Te hace uso de herramientas tecnológicas que le permiten guardar los datos de interacción de los estudiantes en el Learning Management System (LMS) usado por la Universidad de los Andes llamado Blackboard. Esta información está siendo recolectada con el propósito de implementar mejoras en los proyectos actuales de innovación de Conecta-TE. Sin embargo, pese a tener está información, no se han realizado proyectos que les permita aprovechar y generar valor a los datos utilizando técnicas diferentes de análisis que generen conocimiento útil para Conecta-TE.

Es aquí donde se crea la necesidad de este proyecto, que busca crear soluciones de inteligencia de negocios que utilice diversas fuentes de datos provistas por Conecta-TE para analizar el comportamiento de los estudiantes y ayude a la toma de decisiones de mejora en los proyectos de este departamento.

Reconocimientos / Agradecimientos

- **Asesor de tesis:** Haydemar Maria Nuñez Castro
- **Experto Conecta-TE:** Juan Pablo Reyes Gómez, Ingeniero Desarrollador del Equipo Tecnológico de Conecta-TE

2. MARCO TEÓRICO

- **Inteligencia de Negocios (BI):** Es un proceso impulsado por la tecnología para analizar datos y presentar información procesable que ayuda a los ejecutivos, gerentes y otros usuarios finales corporativos a tomar decisiones comerciales informadas. BI abarca una amplia variedad de herramientas, aplicaciones y metodologías que permiten a las organizaciones recopilar datos de sistemas internos y fuentes externas, prepararlos para el análisis, desarrollar y ejecutar consultas contra esos datos y crear informes, paneles y visualizaciones de datos para que los resultados analíticos estén disponibles. a tomadores de decisiones corporativas, así como a trabajadores operativos (1).

- **Business Lifecycle:** Conocido también como Kimball Lifecycle es una metodología desarrollada por Kimball[1] y varios compañeros de trabajo para la creación de data warehouses. Esta metodología define diferentes áreas y una arquitectura de datos sólida para la construcción del data warehouse. A continuación, se muestra un diagrama que ilustra las áreas de esta metodología (2)

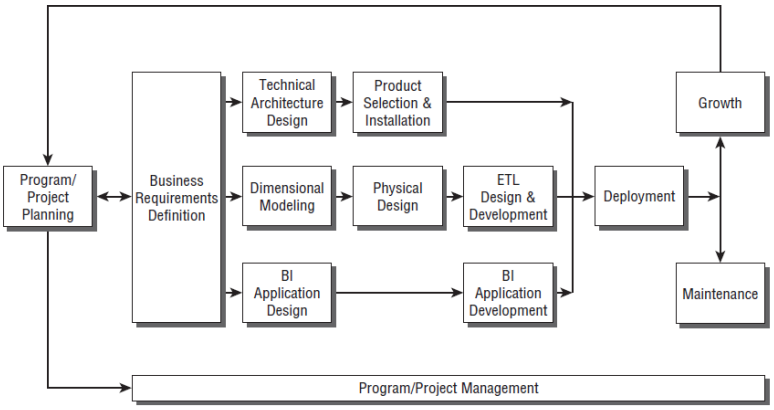


Figura 1: Kimball y Ross (2013). Kimball Lifecycle diagram. [Figura]. Recuperado de The Data Warehouse Toolkit

[1] https://www.kimballgroup.com/

3. DESCRIPCIÓN GENERAL

3.1. OBJETIVOS

- Apoyar el fortalecimiento de la calidad de los recursos disponibles para los estudiantes, desde un enfoque de innovación pedagógica y tecnológica, enriquezcan el proceso de formación del estudiante.

- Apoyar la creación y desarrollo de ambientes de aprendizaje —cursos— en modalidad Blended y nivel educativo, desde un enfoque de innovación pedagógica y tecnológica y en coherencia con los principios de autonomía, flexibilidad e interacción.

- Implementar un algoritmo de predicción supervisado que tome los datos de interacciones de los estudiantes y retorne el resultado final del estudiante en forma categórica relacionado con la aprobación o no del curso.

3.2. ANTECEDENTES

Trabajos y/o proyectos en Conecta-Te que anteceden a este:

- El área de desarrollo de software de Conecta-Te ha elaborado diferentes pruebas usando tableros de control elaborados con la herramienta Tableau. En estos se permite una caracterización de los estudiantes que han tomado el curso común de Colombia, el cual es un curso blended impartido por la Universidad de los Andes y de carácter obligatorio para los estudiantes de pregrado que ingresaron a partir del segundo semestre del 2018.

Trabajos de investigación:

- ✓ En la investigación de Wong, J., Khalil, M., Baars, M., de Koning, B. B., y Paas, F. (2019), comparan y describen el comportamiento de los estudiantes en cursos masivos abiertos en línea (MOOCs por sus siglas en inglés), a partir de los clics de los estudiantes en los recursos disponibles de un curso alojado en Coursera. Para lograr lo anterior, hacen uso del algoritmo para descubrir patrones secuenciales llamado cSPADE. No obstante, para

determinar el camino a seguir de la investigación se toma en consideración diferentes estudios previos a este, en los cuales se recalca de realizar análisis individuales para cada MOOC. Debido a que cada MOOC puede contener videos, lecturas, evaluaciones y más recursos que pueden variar entre MOOCs, luego esta variación hace que cada estudiante cambie su comportamiento entre los distintos cursos en línea en los que se encuentra inscrito.

- ✓ En cuanto a aproximaciones de modelados de datos, Zorrilla (2019) plantea un modelo dimensional genérico para el almacenamiento y tratamiento de datos de resultados de procesos de aprendizaje en la Universidad de Cantabria. En la siguiente figura se muestra el modelo dimensional propuesto por Zorrilla.

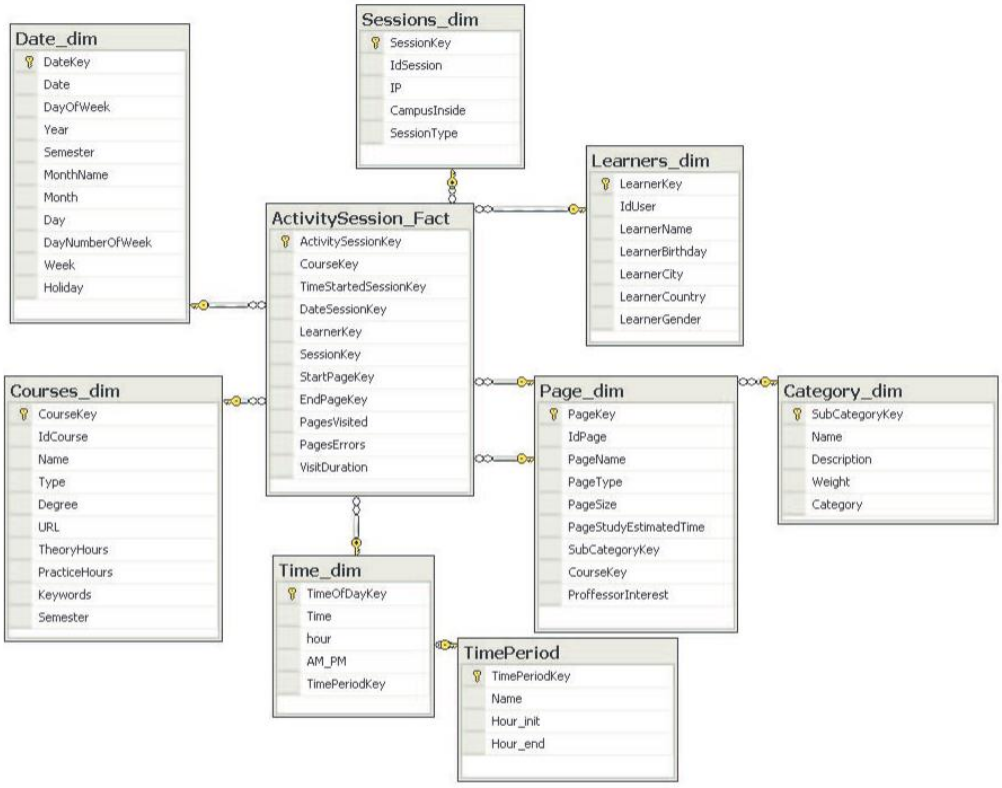


Figura 2: Zorrilla (2009). ActivitySession_Fact schema. [Figura]. Recuperado de Data Warehouse technology for E-learning

Como resultado de la implementación de este modelo se crearon diferentes tableros de control.

3.3. IDENTIFICACIÓN DEL PROBLEMA Y DE SU IMPORTANCIA

Conecta-TE trabaja con diferentes departamentos de la Universidad de los Andes impulsando el uso de tecnologías que apoyen los procesos de aprendizaje en los cursos blended impartidos en la universidad.

Para poder llevar a cabo los procesos de mejora continua de los cursos, se recolectan datos de encuestas de satisfacción al finalizar cada curso y datos referentes a la interacción de los estudiantes con los diferentes materiales y herramientas ofrecidas a los estudiantes en el Learning Managament System.

Sin embargo, pese a tener esta información Conecta-TE no puede utilizarla de manera satisfactoria, debido a que su interacción y uso es complejo. Cada información cuenta con una estructura, un formato y un uso diferente según con lo que se quiera hacer. Además, dada la cantidad de estudiantes, de cursos, profesores, recursos, etc., hacen que el analizar esta información sea una tarea compleja de realizar, dificultando su aprovechamiento y uso en los procesos de mejora y diseño de los cursos.

Actualmente, la universidad está incursionando en la implementación de cursos masivos interdisciplinarios que hagan parte de la formación integral de los estudiantes, esto representa un reto en el diseño del contenido, material y herramientas que permitan ayudar a los estudiantes en el proceso de aprendizaje y que a su vez permita alinear el curso a los pilares institucionales y de Conecta-TE (particularmente el de la autonomía).

Es a la luz de esto que el proyecto toma forma y fundamenta sus objetivos para dar apoyo en el análisis de esta información y apoyar el diseño de herramientas y recursos para cursos masivos.

4. DISEÑO Y ESPECIFICACIONES

4.1. DEFINICIÓN DEL PROBLEMA

Actualmente Conecta-TE tiene datos relacionados con las interacciones de estudiantes en la plataforma, pero la explotación de estos datos es escasa. Por otro lado, otras formas de explotación de datos como algoritmos predictivos o de clasificación no son tenidas en cuenta.

4.2. ESPECIFICACIONES

Requerimiento funcional

Entrada: Un archivo .CSV que contiene la lista de los estudiantes con la información del curso y periodo, el número de interacciones por cada una de las categorías y el resultado final del estudiante en forma categórica (“Fail” si el estudiante perdió la materia o “Pass” si la aprobó)

code_presentation	url	forumng	homepage	oucontent	subpage	resource	sharedsubpage
2014J	143	451	497	1505	143	31	0

forumng	homepage	oucontent	subpage	resource	sharedsubpage	page	questionnaire	ouwiki	htmlactivity	ouelluminate	dataplus	externalquiz	repeatactivity	dualpane	quiz	glossary	oucollaborate	folder	acumneg	acumg	final_result
451	497	1505	143	31	0	0	0	0	0	0	21	0	0	0	0	0	0	0	256	2535	Pass

Salida: Como resultado de la ejecución se espera un archivo .DOT con cada uno de los árboles resultado para cada uno de los periodos.

4.3. RESTRICCIONES

Debido a causas ajenas al proyecto y por temas legales y administrativos, el acceso al conjunto de datos con el que originalmente se pensaba trabajar no fue otorgado en el plazo establecido. Por este motivo se toma la decisión de trabajar con un conjunto de datos diferentes pero que se alinee a los objetivos que se buscan en el proyecto.

Esta restricción trajo consigo otras restricciones relacionadas con el conjunto de datos con el que se pensaba trabajar. Por ejemplo, la diferencia en la duración de los cursos, el tipo de información que se administra por

estudiante, la falta de información descriptiva para algunos campos, el desconocimiento del origen de algunos valores, entre otros.

Por otro lado, se tienen en cuenta las siguientes restricciones:

- Las librerías utilizadas están basadas en Python 3.
- Los algoritmos están hechos en la herramienta de notebooks de Python, Anaconda[2].
- No hay restricciones económicas.
- No hay restricciones de otra índole.

[2] https://www.anaconda.com/

5. DESARROLLO INICIAL

5.1. RECOLECCIÓN DE INFORMACIÓN

Como fuente de datos se utilizó el proyecto Open University Learning Analytics[3] (OULAD) el cual contiene información de cursos, los estudiantes registrados en los cursos y sus interacciones en ambientes de aprendizaje virtuales (VLE). Todos los datos se pueden descargar en archivos CSV en su página (Figura 1).

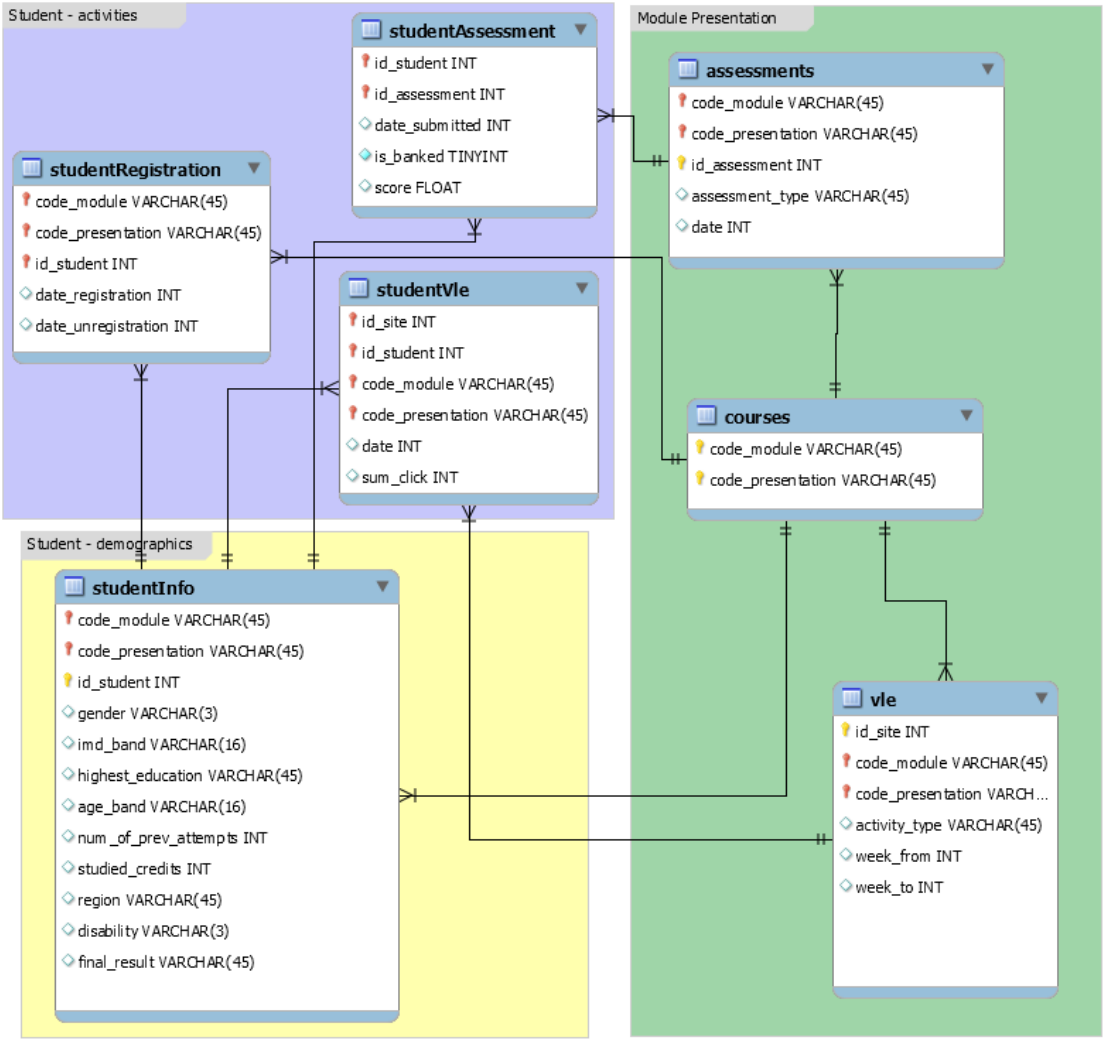


***Figura 1**: Esquema de la base de datos de Open University Learning Dataset (Kuzilek, Hlosta and Zdrahal, 2017)*

[3] https://analyse.kmi.open.ac.uk/open_dataset

5.2. ALTERNATIVAS DE DISEÑO

Teniendo en consideración que el proyectó se llevó acabo haciendo uso de librerías de Python para el procesamiento de datos, aplicación de algoritmos de minería de datos se contemplaron otras dos alternativas de diseño:

- Uso de Weka o KNIME para el procesamiento de datos y aplicación de algoritmos de minería de datos, Tableau para la creación de tableros.
- Uso de otras librerías y de otro lenguaje de programación para el procesamiento de datos y aplicación de técnicas de minería de datos.

El diseño que se propuso tiene en consideración los conocimientos de los integrantes del proyecto y del área de desarrollo de Conecta-TE, ya que cada una de las herramientas mencionadas requieren de un nivel de experticia y para lograrlo es necesario dedicar tiempo considerable para lograrlo. Por esto, se tuvo en cuenta librerías en Python, estas son trabajadas en el curso de Inteligencia de Negocios y son conocidas por ingenieros en Conecta-TE.

6. IMPLEMENTACIÓN

6.1. DESCRIPCIÓN DE LA IMPLEMENTACIÓN

ETAPA I: Comprender el problema y verificar que se puede construir una solución de inteligencia de negocios.

Después de la primera reunión con el experto en Conecta-TE se vio una prueba de concepto realizada por ellos. Esto dio a entender la manera en la que ellos veían viable una mayor exploración utilizando otras herramientas de inteligencia de negocios.

Se fijó como objetivo, el diseñar un esquema que permita ayudar a analizar el comportamiento de los estudiantes de acuerdo con la manera como interactúan con los distintos recursos publicados. Esto teniendo como referencia una visión general del alcance del proyecto y lo que se esperaba lograr a largo plazo con él.

ETAPA II: Análisis de los datos.

Como se mencionó en secciones previas, el conjunto de datos que se utilizaría para realizar los análisis sería el provisto por el proyecto OULAD. Para su utilización, se procede a realizar la respectiva revisión de la documentación oficial publicada en la página. Adicionalmente, se procede a realizar la descarga de los archivos con la información.

Una vez obtenida la información, se realizaron diferentes tablas para perfilar los datos utilizando la librería Pandas[4] y scripts en Python, dichas tablas de

[4] https://pandas.pydata.org/

perfilamiento pueden ser vistas en la sección de anexos. A continuación, se realizará la descripción de la distinta información contenida en los archivos.

Assessment: Este archivo contiene la información de las diferentes evaluaciones realizadas a lo largo del curso. En general, cada curso presenta una serie de exámenes antes de la realización del examen final (este es normalmente en la última semana de presentación). Las columnas contenidas junto con su perfilamiento pueden ser encontrado en la Tabla 2: Perfilamiento de los datos de la tabla Assessments en la sección de anexos.

Courses: En este archivo se encuentra la información respectiva a los diferentes cursos y los periodos en que fueron ofrecidos. Cada curso tiene comienzo en los meses de febrero (identificado con la letra B) o en octubre (identificado con la letra J) y transcurren durante los años 2013 y 2014. Las columnas contenidas junto con su perfilamiento pueden ser encontrado en la Tabla 3: Perfilamiento de los datos de la tabla Courses en la sección de anexos.

StudentAssessment: El archivo contiene la información de las evaluaciones presentadas por los estudiantes junto con el respectivo puntaje obtenido. En caso de que algún estudiante no presentara la evaluación, no aparecerá en el contenido del archivo. Las columnas contenidas junto con su perfilamiento pueden ser encontrado en la Tabla 4: Perfilamiento de los datos de la tabla StudentAssesment en la sección de anexos.

StudentRegistration: El archivo contiene la información del registro (y para algunos de la cancelación) de los estudiantes a los diferentes cursos. Las columnas contenidas junto con su perfilamiento pueden ser encontrado en la Tabla 5: Perfilamiento de los datos de la tabla StudentRegistration en la sección de anexos.

StudentInfo: EL archivo contiene la información de los estudiantes, principalmente información demográfica y de los resultados finales obtenidos. El resultado final de cada estudiante es una de las siguientes categorías:

- *"pass"* indicando que el estudiante aprobó el curso.
- "fail" indicando que el estudiante reprobó el curso.
- "withdraw" indicando que el estudiante retiró el curso.
- "distinction" indicando que el estudiante aprobó en curso con una nota muy alta (normalmente es superior a 70)

Las anteriores categorías son suposiciones evidenciadas a la luz de los datos, la documentación oficial no cuenta con la descripción de estás categorías. Las columnas contenidas junto con su perfilamiento pueden ser encontrado en la Tabla 6: Perfilamiento de los datos de la tabla StudentInfo en la sección de anexos.

StudenVle: EL archivo contiene la información de las interacciones del estudiante en el ambiente de aprendizaje virtual. Estas interacciones se miden con el número de clics que realiza el estudiante en los diferentes recursos. Las columnas contenidas junto con su perfilamiento pueden ser encontrado en la Tabla 7: Perfilamiento de los datos de la tabla StudentVle en la sección de anexos.

Vle: El archivo contiene la información de todos los recursos ofrecidos en el ambiente virtual de aprendizaje. Estos recursos normalmente son paginas o documentos PDF. Las columnas contenidas junto con su perfilamiento pueden ser encontrado en la Tabla 8: Perfilamiento de los datos de la tabla Vle en la sección de anexos.

La información de OULAD que está disponible tiene datos desde el primer semestre del 2013 hasta el segundo semestre del 2015. Se utiliza el periodo de tiempo completo de esta información.

ETAPA III: Propuestas de análisis

Teniendo el análisis de los datos y cruzando con los objetivos y el entendimiento del Conecta-TE, se procede con la identificación de los análisis que pueden ser realizados con los datos y que a su vez se ajusten a las necesidades del negocio. En la tabla a continuación pueden verse los temas analíticos propuestos para el desarrollo del proyecto. Este proyecto se enfoca en el requerimiento analítico número 2 con la aplicación de minería de datos con algoritmos de árboles de decisión.

Tabla 1: Descripción de los requerimientos analíticos a realizar en el proyecto

Tema analítico	Análisis requeridos o inferidos	Categoría del análisis - Tablero de control, análisis OLAP, Minería de datos	Procesos de negocio	Fuentes de datos y datos
Análisis del impacto de la interacción de los estudiantes en los cursos apoyados por conecta-TE	Segmentar a los estudiantes en grupos que permitan clasificarlos de acuerdo con los tipos de interacciones que realizan en la plataforma	Minería de Datos - Clustering	Apoyar la toma de decisiones para la creación de contenidos para los cursos	Datos de intereaciones de los estudiantes en ambientes de aprendizaje virtual del proyecto OULAD
	Determinar la probabilidad de aprobar o reprobar un curso a partir de los comportamientos de los estudiantes	Minería de Datos - Decision tree		
	Determinar los tipos y formatos de contenido que son mas relevantes para los estudiantes y el impacto que tienen en la nota final del curso	Tablero de control	Apoyar la toma de decisiones para la creación de contenidos para los cursos	

ETAPA IV: Modelado del Data Mart

Proceso de Negocio: Apoyar la toma de decisiones para la creación de contenidos para los cursos blended.

Tabla de hechos #1: Interacción

Granularidad: Información del acceso de un estudiante a un recurso disponible en la plataforma virtual del curso en el que se encuentra inscrito y la fecha en que ocurre.

Medidas:

Aditivas: Aquellas medidas que por su naturaleza nos permite sumarlas sin importar las dimensiones y sin que el resultado

pierda sentido. Para el modelo propuesto, las medidas que cumplen esta condición es sum_click

Semi-aditivas: Aquellas medidas que se caracterizan por no tener sentido sumar su valor para todas las dimensiones sino solo para algunas en específico. En esta tabla no se cuenta con medidas semi-aditivas

No aditivas: Aquellas medidas que se caracterizan porque su suma no tiene sentido para ninguna dimensión. En esta tabla de hechos, no se cuenta con medidas no aditivas

Dimensiones:

Fecha: Dimensión que representa la fecha.

- anio: Año que se está revisando en números.
- mes: Mes que se está revisando en números.
- dia: Día que se está revisando en números.
- fecha: Fecha completa que se está revisando
- code_presentation: Nombre código del semestre, consiste en el año y la letra 'B' si inicia en el mes de febrero o la letra 'J' si inicia en el mes de octubre.

Curso: Dimensión que contiene todos los cursos disponibles en la fuente de datos.

- code_module: Nombre código del curso

- Estudiante: Contiene la información demográfica del estudiante.
  - Id_student: Identificador único del estudiante en la base de datos de (OULAD)
  - gender: Genero del estudiante
  - region: Ubicación geográfica en donde el estudiante se encuentra viviendo mientras toma el curso
  - highest_education: Nivel más alto de educación cuando ingresa al curso el estudiante
  - imd_band: Índice de desigualdad, aplicado en Reino Unido, para el lugar en donde el estudiante vive mientras el estudiante toma el curso.

- Recurso: Representa la información de los recursos
  - id_site: Número que identifica al recurso
  - activity_type: Tipo de recurso
  -

Modelo dimensional:

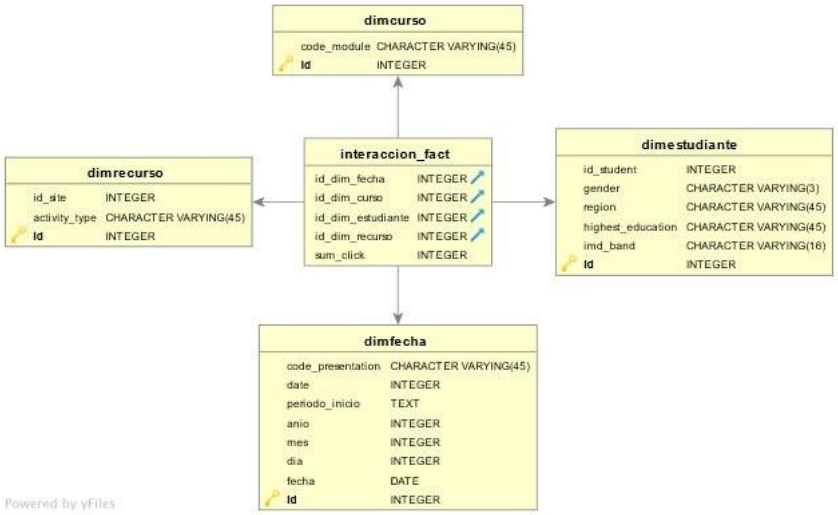


Tabla de hechos #2: Rendimiento

Granularidad: Información de la nota de una evaluación de un estudiante en un curso disponible en la plataforma virtual en el que se encuentra inscrito y la fecha en que ocurre.

Medidas:

- scored:
- is_banked:
- count:

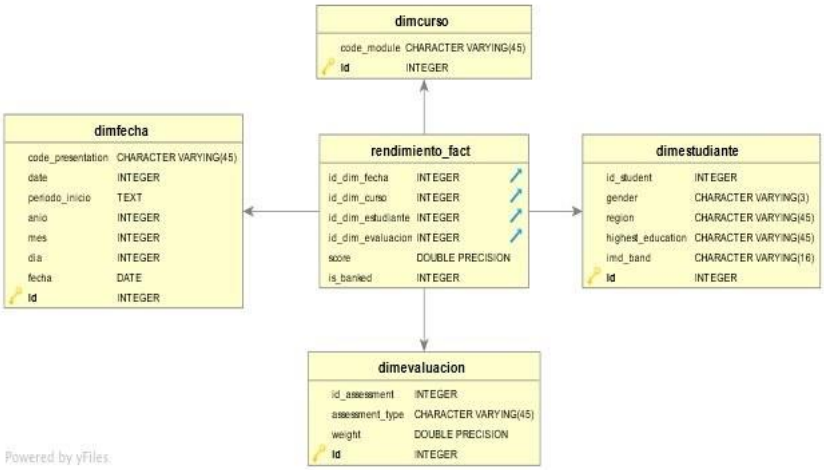
Dimensiones:

Se hace uso de las dimensiones Estudiante, Curso y Fecha anteriormente descritas y de la dimensión:

- Evaluación: Contiene la información de las evaluaciones disponibles por cursos
  - id_assessment: Identificador numérico de la evaluación

- assessment_type: Tipo de evaluación. Existen tres tipos de evaluaciones: evaluación marcada por el tutor (TMA), evaluación marcada por la computadora (CMA) y examen final (FE).
- weight: el peso de la evaluación en porcentaje. En general, el peso de los exámenes es 100% y de la suma de las demás evaluaciones es 100%.

Modelo dimensional



Etapa V: Proceso ETL

Para el proceso de extracción, transformación y carga, se comenzó extrayendo los datos de la fuente descrita en la sección anterior como archivos en formato CSV. Posteriormente, estos datos fueron cargados en una base de datos para su posterior análisis.

Una vez los datos se encuentran cargados, se procede a ejecutar un conjunto de sentencias para poder transformar y extraer los datos según las necesidades de las dimensiones.

Luego de insertar los datos se procede a generar los identificadores de cada dimensión según corresponda y generar las tablas de hechos (Ver Figura 6)

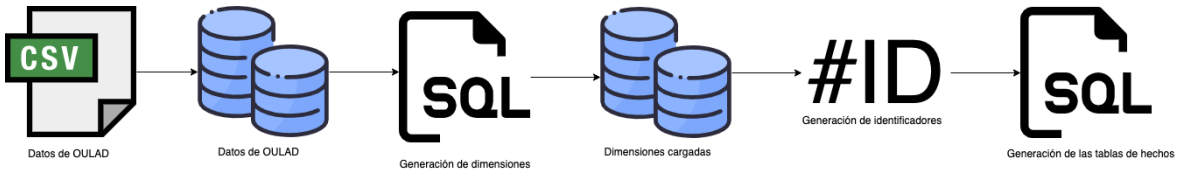


Figura 3: Ilustración del proceso de cargue de datos al data mart

ETAPA VI: Preparación de los datos.

Ya con el data mart implementado, se realizada las consultas con las cuale se esperan obtener los datos necesarios para el entrenamiento del modelo. Combinando los datos de las interacciones y los estudiantes se puede crear el conjunto de datos con los cuales se va a trabajar. Este va a contener los datos de todos los estudiantes con su correspondiente código del curso al que asistieron y el periodo en el cual vieron el curso. Además, los datos de las interacciones que hizo el estudiante por cada uno de los tipos de recurso. Por último, se añade el resultado final del estudiante en forma cualitativa: Fail, si el estudiante reprobó el curso, Pass si lo pasó, Withdrawn si el estudiante lo retiró y por último Distinction si el estudiante obtuvo buenos resultados en el curso.

Tabla 2: Ejemplo de fila en el CSV final

code_presentation	url	forumng	homepage	oucontent	subpage	resource	sharedsubpage	page	questionnaire	ouwiki	htmlactivity	ouelluminate	dataplus	externalquiz	repeatactivity	dualpane	quiz	glossary	oucollaborate	folder
2014J	143	451	497	1505	143	31	0	0	0	0	0	0	21	0	0	0	0	0	0	0
2013J	23	36	184	64	227	70	0	0	0	18	0	0	0	12	0	0	0	0	12	0
2014J	0	2	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

folder	acumneg	acumg	final_result
0	256	2535	Pass
0	81	565	Withdrawn
0	0	10	Withdrawn

ETAPA V: Pruebas de concepto y resultados

Se realizaron 8 pruebas de concepto entrenando 32 árboles de decisión variando múltiples restricciones y modificando el tratamiento de los datos. Todos los árboles son utilizando el algoritmo *DecisionTreeClassifier* de la librería Sci-Kit Learn para Python 3. Por otro lado, todas pruebas se realizan con la búsqueda de los mejores hiperparámetros (profundidad y criterio de selección de atributos) utilizando el algoritmo *RandomizedSearchGrid* de Sci-Kit Learn.

Prueba 1:

- Las columnas con una ocurrencia de ceros mayor al 65% fueron eliminadas.
- Las columnas restantes pasan por un proceso de discretización que consiste en calcular el cuartil correspondiente y este valor pasa a ser el nuevo valor en la columna.
- Estas mismas columnas pasan a ser valores *dummies.*
- Las filas que contienen el valor de Distinction y Withdrawn en la columna final_result son eliminadas. Esta decisión fue tomada dado que el valor de Withdrawn no cuenta con registros completos a lo largo del periodo evaluado. Igualmente, el valor de Distinction al momento de realizar la prueba no era claro si los estudiantes que obtenían este resultado habían terminado o no el curso.
- El conjunto de datos es separado por periodos y se balancea.
- Los sets son separados en proporción 80-20 para entrenamiento y pruebas.
- Con los 4 conjuntos de datos se entrenan 4 árboles de decisión utilizando *RandomizedSearchCV* para buscar los mejores hiperparámetros del árbol.

Resultados obtenidos

- Para el árbol entrenado del periodo 2013B:
- Mejores parámetros
    - Profundidad: None
    - Criterio: gini

	precision	recall	f1-score	support
0	0.86	0.87	0.86	378
1	0.88	0.87	0.87	422
accuracy			0.87	800
macro avg	0.87	0.87	0.87	800
weighted avg	0.87	0.87	0.87	800

- Para el árbol entrenado del periodo 2013J
- Mejores parámetros:
    - Profundidad: None
    - Criterio: gini

	precision	recall	f1-score	support
0	0.81	0.85	0.83	378
1	0.86	0.82	0.84	422
accuracy			0.84	800
macro avg	0.84	0.84	0.84	800
weighted avg	0.84	0.84	0.84	800

- Para el árbol entrenado del periodo 2014B
- Mejores parámetros:
    - Profundidad: None
    - Criterio: gini

	precision	recall	f1-score	support
0	0.82	0.88	0.85	378
1	0.89	0.82	0.85	422
accuracy			0.85	800
macro avg	0.85	0.85	0.85	800
weighted avg	0.85	0.85	0.85	800

- Para el árbol entrenado del periodo 2014J
- Mejores parámetros:
    - Profundidad: None
    - Criterio: entropy

	precision	recall	f1-score	support
0	0.86	0.87	0.87	378
1	0.89	0.88	0.88	422
accuracy			0.88	800
macro avg	0.87	0.87	0.87	800
weighted avg	0.88	0.88	0.88	800

Prueba 2:

- Las columnas con una ocurrencia del cero mayor al 85% fueron transformadas a variables binarias (1 si el recurso fue utilizado, 0 si no).
- Las columnas con una ocurrencia del cero menor al 85% pasan por un proceso de discretización que consiste en calcular el cuartil correspondiente y este valor pasa a ser el nuevo valor en la columna.
- Estas mismas columnas pasan a ser valores *dummies*.
- Las filas de que tengan el valor “Distinction” en final_result se agrupan con el valor “Pass” ya que en una exploración se concluyó que los estudiantes que obtienen este resultado final aprueban la materia.
- Las filas que contienen un valor de “Withdrawn” en la columna final_result son eliminadas.
- El set de datos es separado por periodos y es balanceado.
- Los sets son separados en proporción 80-20 para entrenamiento y pruebas.
- Con los 4 set de datos se entrenan 4 árboles de decisión utilizando RandomizedSearchCV.

Resultados obtenidos

- Para el árbol entrenado del periodo 2013B:
- Mejores parámetros
    - Profundidad: 19

- Criterio: gini

	precision	recall	f1-score	support
0	0.89	0.92	0.90	378
1	0.93	0.89	0.91	422
accuracy			0.91	800
macro avg	0.91	0.91	0.91	800
weighted avg	0.91	0.91	0.91	800

- Para el árbol entrenado del periodo 2013J
- Mejores parámetros:
  - Profundidad: 19
  - Criterio: entropy

	precision	recall	f1-score	support
0	0.80	0.85	0.82	378
1	0.86	0.81	0.83	422
accuracy			0.83	800
macro avg	0.83	0.83	0.83	800
weighted avg	0.83	0.83	0.83	800

- Para el árbol entrenado del periodo 2014B
- Mejores parámetros:
  - Profundidad: None
  - Criterio: gini

	precision	recall	f1-score	support
0	0.82	0.89	0.85	378
1	0.89	0.82	0.86	422
accuracy			0.85	800
macro avg	0.85	0.86	0.85	800
weighted avg	0.86	0.85	0.85	800

- Para el árbol entrenado del periodo 2014J
- Mejores parámetros:
  - Profundidad: 19
  - Criterio: gini

	precision	recall	f1-score	support
0	0.87	0.92	0.90	378
1	0.93	0.88	0.90	422
accuracy			0.90	800
macro avg	0.90	0.90	0.90	800
weighted avg	0.90	0.90	0.90	800

Prueba 3, 4, 5:

- Para la prueba 3, las columnas con una ocurrencia mayor al 90% fueron transformadas a variables binarias (1 si el recurso fue utilizado o 0 si no).
- Para la prueba 4, las columnas con una ocurrencia mayor al 80% fueron transformadas a variables binarias (1 si el recurso fue utilizado o 0 si no).
- Para la prueba 5, las columnas con una ocurrencia mayor al 70% fueron transformadas a variables binarias (1 si el recurso fue utilizado o 0 si no).
- Para todas las pruebas, las variables que cumplen con la restricción del porcentaje son normalizadas utilizando MinMaxScaler de Pandas.
- Para todas las pruebas, las variables que cumplen con la restricción del porcentaje son reemplazadas con el cálculo del cuartil al que pertenecen.
- Estas mismas columnas pasan a ser valores *dummies*.
- Las filas de que tengan el valor “Distinction” en final_result se agrupan con el valor “Pass” ya que en una exploración se concluyó que los estudiantes que obtienen este resultado final aprueban la materia.
- Las filas que contienen un valor de “Withdrawn” en la columna final_result son eliminadas.
- El conjunto de datos es separado por periodos y es balanceado.
- Los sets son separados en proporción 80-20 para entrenamiento y pruebas.
- Con los 4 set de datos se entrenan 4 árboles de decisión utilizando *RandomizedSearchCV*.

Consideraciones:

- ***Accuracy 1 y Recall 1*** corresponde a la exactitud y recall de la prueba 3
- ***Accuracy 2 y Recall 2*** corresponde a la exactitud y recall de la prueba 4
- ***Accuracy 3 y Recall 3*** corresponde a la exactitud y recall de la prueba 5

Resultados obtenidos:

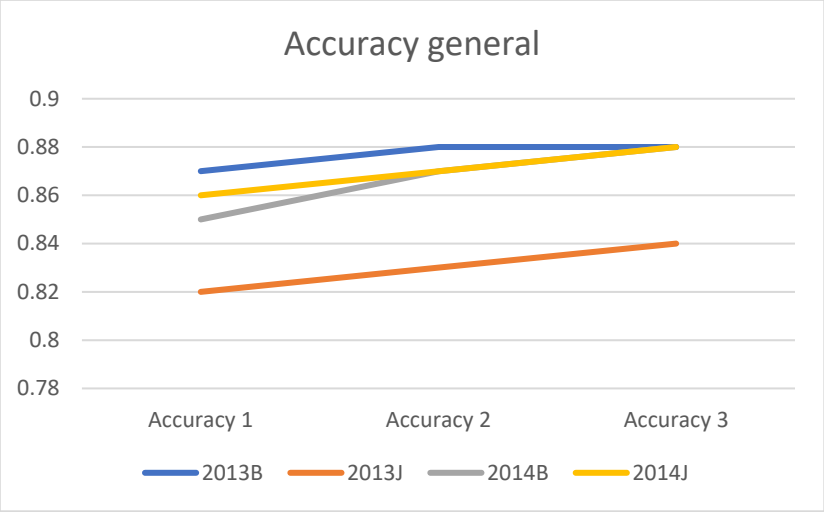


Fig 7. Exactitud de los árboles por periodo de las pruebas 3,4 y 5

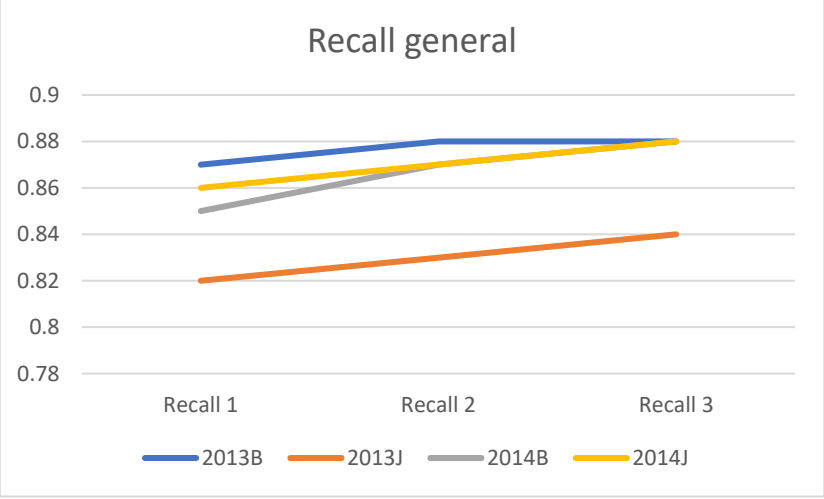


Fig 8. Recall de los árboles por periodo de las pruebas 3,4 y 5

Mejores hiperparámetros:

- Árbol correspondiente al periodo 2013B

Prueba	**Criterio**	**Profundidad**
Prueba 3	Entropy	19
Prueba 4	Entropy	None
Prueba 5	Gini	15

- Árbol correspondiente al periodo 2013J

Prueba	**Criterio**	**Profundidad**
Prueba 3	Entropy	17
Prueba 4	Gini	None
Prueba 5	Gini	None

- Árbol correspondiente al periodo 2014B

Prueba	**Criterio**	**Profundidad**
Prueba 3	Gini	18
Prueba 4	Entropy	20
Prueba 5	Entropy	20

- Árbol correspondiente al periodo 2014J

Prueba	**Criterio**	**Profundidad**
Prueba 3	Gini	20
Prueba 4	Gini	18
Prueba 5	Entropy	16

Las pruebas 3, 4 y 5 cumplen con características muy similares siendo la única diferencia el porcentaje sobre el cual las variables pasan a ser binarias o no. Los resultados son comparados con el fin de detectar cuál es la mejor manera de tratar las columnas con un alto número de incidencia del cero.

Por lo que se puede observar en las figuras 1 y 2, una disminución en el porcentaje sobre el cual las columnas pasan a binario ayuda a un incremento en la exactitud y recall general de los árboles.

Prueba 6, 7, 8:

- Las pruebas 6, 7 y 8 tienen las mismas condiciones que las pruebas 3, 4, 5 correspondientemente, con la excepción que las clases no están balanceadas.

Consideraciones:

- ***Accuracy 1 y Recall 1*** corresponde a la exactitud y recall de la prueba 6
- ***Accuracy 2 y Recall 2*** corresponde a la exactitud y recall de la prueba 7
- ***Accuracy 3 y Recall 3*** corresponde a la exactitud y recall de la prueba 8

Resultados obtenidos

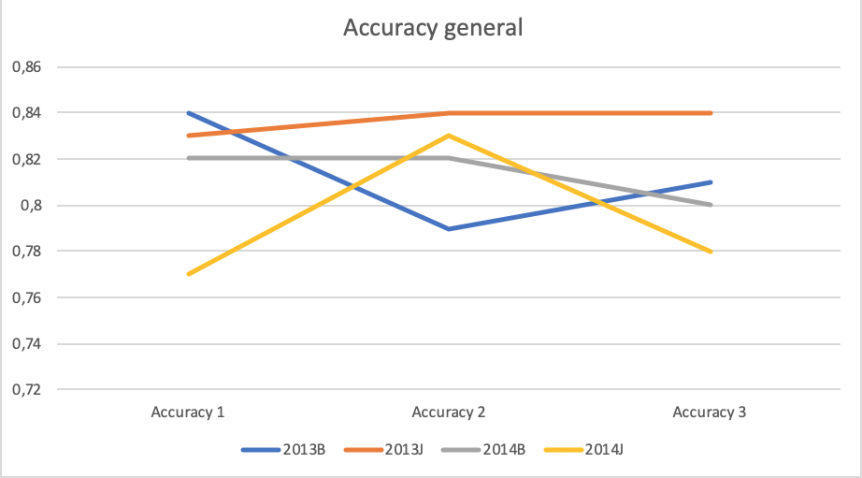


Fig 9. Exactitud de los árboles por periodo de las pruebas 6,7 y 8

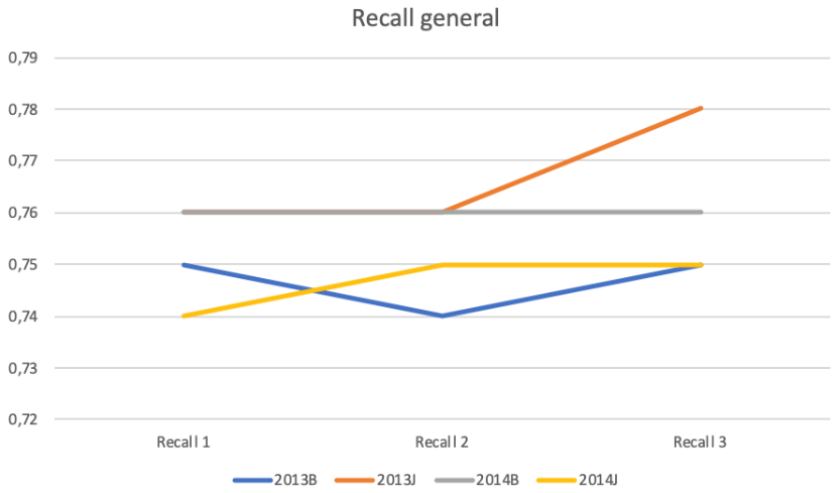


Fig 10. Recall de los árboles por periodo de las pruebas 6,7 y 8

Mejores hiperparámetros:

- Árbol correspondiente al periodo 2013B

Prueba	**Criterio**	**Profundidad**
Prueba 3	Entropy	6
Prueba 4	Gini	6
Prueba 5	Gini	6

- Árbol correspondiente al periodo 2013J

Prueba	**Criterio**	**Profundidad**
Prueba 3	Entropy	8
Prueba 4	Gini	6
Prueba 5	Gini	7

- Árbol correspondiente al periodo 2014B

Prueba	**Criterio**	**Profundidad**
Prueba 3	Entropy	7
Prueba 4	Gini	6
Prueba 5	Entropy	5

- Árbol correspondiente al periodo 2014J

Prueba	Criterio	Profundidad
Prueba 3	Entropy	11
Prueba 4	Gini	6
Prueba 5	Entropy	6

Las pruebas 6, 7 y 8 cumplen con características muy similares siendo la única diferencia el porcentaje sobre el cual las variables pasan a ser binarias o no. Los resultados son comparados con el fin de detectar cual es la mejor manera de tratar las columnas con un alto número de incidencia del cero.

Por lo observado en las figuras 9 y 10, en las pruebas con la clase no balanceada no es posible afirmar una mejora con respecto al criterio del porcentaje con el cual las variables se transforman a binaria. Sin embargo, el hiperparámetro de profundidad es menor en todos los árboles, en todas las pruebas realizadas con respecto a las pruebas 3,4 y 5. Por otro lado, la exactitud en todos los casos fue un poco menor a las pruebas con las clases balanceadas, esta diferencia no llega a superar el 8% de exactitud.

7. VALIDACIÓN

7.1. MÉTODOS

Los modelos son validados teniendo en cuenta las métricas obtenidas después de la realización de las pruebas. Con las métricas de exactitud y recall es posible definir un nivel de confianza sobre el modelo.

8. CONCLUSIONES

8.1. DISCUSIÓN

Con el objetivo de ampliar las pruebas de concepto realizadas por Conecta-TE se obtienen resultados positivos. Si bien los datos utilizados en las pruebas no son los con los que cuenta Conecta-TE, la estructura y el modelo dimensional fueron pensados de una manera general para cualquier conjunto de datos relacionados con ambientes de aprendizaje virtuales. Con esta

nueva aproximación, Conecta-TE puede darle un valor agregado a los datos con los que cuenta. Una vez implementado el modelo, le permitirá entender cómo el uso de los diferentes recursos disponibles para los estudiantes afecta su proceso educativo.

Con respecto a los árboles de decisión resultado de las pruebas se concluye que los mejores resultados se obtienen con las condiciones de la prueba 5, alcanzando una exactitud y recall del 88%. Por otro lado, con resultados generales que superan el 77% de exactitud y el 74% de recall se considera que satisfactorio el desarrollo del presente proyecto. En el Anexo se resalta uno de los árboles de decisión resultado y un cuadro con las variables más importantes de dicho árbol.

8.2. TRABAJO FUTURO

Para trabajos a realizar en el futuro se puede considerar:

- Adaptar los datos que posee Conecta-TE al modelo propuesto
- Implementar un algoritmo de regresión logística regularizada a L1 con el fin de definir los recursos de mayor relevancia.

9. REFERENCIAS

- Kimball, R., & Ross, M. (2013). The data warehouse toolkit: The definitive guide to dimensional modeling (Tercera edición). John Wiley & Sons.
- Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset Sci. Data 4:170171 doi: 10.1038/sdata.2017.171 (2017).
- Wong, J., Khalil, M., Baars, M., de Koning, B. B., & Paas, F. (2019). Exploring sequences of learner activities in relation to self-regulated learning in a massive open online course. Computers & Education, 140, 103595.
- Zorrilla, M. E. (2009). Data warehouse technology for e-learning. In Methods and Supporting Technologies for Data Analysis (pp. 1-20). Springer, Berlin, Heidelberg.

ANEXOS

Perfilamiento de los datos de OULAD

Tabla 2: Perfilamiento de los datos de la tabla Assessments

Assessments							
variable	tipo	cantidad	media	std	min	max	listado_valores
code_module	string	206					AAA,BBB,CCC,DDD,EEE,FFF,GGG
code_presentation	string	206					2013J,2014J,2013B,2014B
id_assessment	int64	206	26474	10098.6	1752	40088	
assessment_type	string	206					TMA,Exam,CMA
date	float64	195	145.01	76	12	261	
weight	float64	206	20.87	30.38	0	100	

Tabla 3: Perfilamiento de los datos de la tabla Courses

Courses							
variable	tipo	cantidad	media	std	min	max	listado_valores
code_module	string	22					AAA,BBB,CCC,DDD,EEE,FFF,GGG
code_presentation	string	22					2013J,2014J,2013B,2014B
module_presentation_length	int64	22	255.55	13.65	234.0	269.0	

Tabla 4: Perfilamiento de los datos de la tabla StudentAssesment

StudentAssessment							
variable	tipo	cantidad	media	std	min	max	listado_valores
id_assessment	int64	173912	26553.80	8829.78	1752.0	37443.0	
id_student	int64	173912	705150.72	552395.19	6516.0	2698588.0	
date_submitted	int64	173912	116.03	71.48	-11.0	608.0	
is_banked	int64	173912	0.01	0.10	0.0	1.0	
score	float64	173739	75.80	18.80	0.0	100.0	

Tabla 5: Perfilamiento de los datos de la tabla StudentRegistration

StudentRegistration							
variable	tipo	cantidad	media	std	min	max	listado_valores
code_module	string	32593					AAA,BBB,CCC,DDD,EEE,FFF,GGG
code_presentation	string	32593					2013J,2014J,2013B,2014B
id_student	int64	32593	706687.67	549167.31	3733.0	2716795.0	
date_registration	float64	32548	-69.41	49.26	-322.0	167.0	
date_unregistration	float64	10072	49.76	82.46	-365.0	444.0	

Tabla 6: Perfilamiento de los datos de la tabla StudentInfo

StudentInfo							
variable	tipo	cantidad	media	std	min	max	listado_valores
code_module	string	32593					AAA,BBB,CCC,DDD,EEE,FFF,GGG
code_presentation	string	32593					2013J,2014J,2013B,2014B
id_student	int64	32593	706688	549167	3733	2,72E+10	
gender	string	32593					M,F
region	string	32593					EastAnglianRegion,Scotland,NorthWesternRegion,...
highest_education	string	32593					HEQualification,ALevelorEquivalent,LowerThanAL...
imd_band	string	31482					90-100%,20-30%,30-40%,50-60%,80-90%,70-80%,nan...
age_band	string	32593					55<=,35-55,0-35
num_of_prev_attempts	int64	32593	0.16	0.48	0	6	
studied_credits	int64	32593	79.76	41.07	30	655	
disability	string	32593					N,Y
final_result	string	32593					Pass,Withdrawn,Fail,Distinction

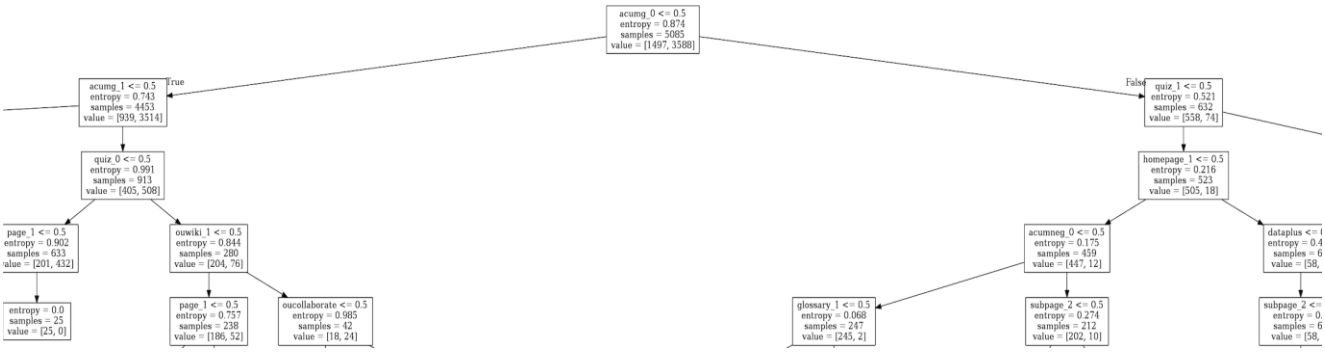
Tabla 7: Perfilamiento de los datos de la tabla StudentVle

StudentVle							
variable	tipo	cantidad	media	std	min	max	listado_valores
code_module	string	10655280					AAA,BBB,CCC,DDD,EEE,FFF,GGG
code_presentation	string	10655280					2013J,2014J,2013B,2014B
id_student	int64	10655280	733333.57	582705.98	6516.0	2698588.0	
id_site	int64	10655280	738323.42	131219.62	526721.(	1049562.0	
date	int64	10655280	95.17	76.07	-25.0	269.0	
sum_click	int64	10655280	3.72	8.85	1.0	6977.0	

Tabla 8: Perfilamiento de los datos de la tabla Vle

Vle							
variable	tipo	cantidad	media	std	min	max	listado_valores
id_site	int64	6364	726099	128315	526721	1,08E+11	
code_module	string	6364					AAA,BBB,CCC,DDD,EEE,FFF,GGG
code_presentation	string	6364					2013J,2014J,2013B,2014B
activity_type	string	6364					resource,oucontent,url,homepage,subpage,glossa...
week_from	float64	1121	15.2	8.79	0	29	
week_to	float64	1121	15.21	8.78	0	29	

Árbol de decisión parcial resultados de la prueba 6



	feature	importance
53	acumg_0	0.390
54	acumg_1	0.115
36	page_1	0.093
4	folder	0.083
43	quiz_1	0.052
57	acumg_4	0.048
42	quiz_0	0.047
21	oucontent_1	0.025
30	resource_0	0.018
56	acumg_3	0.015
5	url_0	0.013