Informática y comunicaciones

ies azarquiel

Curso especialización en inteligencia artificial y big data

Trabajo Noviembre 2025:Sistema Educativo

Profesor: Sebastián Rubio Valero

Noviembre 2025

Big Data

La actividad consiste en aplicar la metodología KDD a un sistema educativo. Debes realizar un documento pdf con todas las fases de esta metodología, presentado diagramas, análisis, conclusiones, etc Debes entregar lo siguiente:

- Un cuaderno Jypyter Notebook comentado, explicando las partes más importantes
- Un documento pdf

DATOS

Los datos que se proporcionan son sintéticos, no se ajustan a la realidad.
Una vez localizados los datos, realizar un primera inspección,

0.1 Estudiantes.csv

En este fichero tiene la relación de estudiante, con los siguientes campos

- estudiante-id: Identificador único
- edad: 14-17 años (rango realista)
- genero: M/F
- ciudad: Diversas ciudades españolas
- nivel-educativo: Secundaria/Bachillerato
- estrato-socioeconomico: Bajo/Medio/Alto
- tipo-colegio: Público/Privado/Concertado
- idioma-nativo: Varias lenguas cooficiales

0.2 Calificaciones.csv

En este fichero hay la notas de seis materias de cada uno de los estudiantes. Las características son las siguientes:

- materia: 6 materias diferentes
- trimestre: 1 a 3
- calificación: 1-10

- asistencia-porcentaje: 60-100
- participación-clase: Baja/Media/Alta

0.3 Actividades extracurriculares

Otros datos necesarios son las actividades extracurriculares, esto mide la capacidad de implicación del alumno Las características son las siguientes:

- actividad-tipo: 6 tipos de actividades
- horas-semana: 1-6 horas
- participación-anios: 1-4 años
- logros: Logros específicos

1 FASE 1: Selección y compresión del dominio

1.1 Análisis Exploratorio del Dominio

El objetivo es comprender el contexto educativo y las variables disponibles.

1. Analizar la distribución demográfica de los estudiantes
2. Identificar posibles relaciones entre variables socioeconómicas y académicas
3. Documentar hipótesis iniciales sobre factores que afectan el rendimiento

1.2 Definición de Objetivos de Minería

En este punto se da una relación que preguntas que te debes hacer. La respuesta a estas preguntas la encontrarás cuando hayas realizando las fases posteriores. Al final del documento deberás responder a todas ellas(es conveniente que propongas dos preguntas más)

- ¿El estrato socioeconómico predice el rendimiento académico?
- ¿Las actividades extracurriculares mejoran el rendimiento en específicas materias?
- ¿Existen patrones temporales en el rendimiento por trimestres?
- ¿Podemos identificar estudiantes en riesgo de bajo rendimiento?
- Etc.....

2 FASE 2: Preprocesamiento y limpieza

En esta fase tienes que trabajar con los datos, aplicando técnicas vistas en clase

2.1 Integración de Múltiples Fuentes

En este punto debes unir los diferentes dataset de forma que consideres que será útil para el modelo/s de A.A. La operaciones que se suelen realizar, si lo consideras necesario, son las siguientes:

1. Realizar inner/outer joins entre las tablas
2. Verificar consistencia de claves primarias
3. Identificar estudiantes sin datos completos
4. Crear variables derivadas (promedio general, etc.

2.2 Limpieza y Tratamiento de Valores Anómalos

Deberás detectar los valores outliers y realizar el tratamiento adecuado. Debes documentar que valores son outliers, como los has tratado, etc. Las acciones ha realizar son las siguientes:

1. Identificar calificaciones fuera del rango 0-10
2. Detectar asistencias fuera de rangos realistas (0-100
3. Tratar valores missing en participación en clase
4. Normalizar variables numéricas

2.3 Ingeniería de Características

En esta sección se crean o eliminan características . La decisión de eliminar unas características debe estar precedida en un análisis de datos, empleando algunos modelos de A.A (de los vistos en clase). Las acciones a realizar, se procede, son:
Variables a crear:

1. promedio-general (mean de todas las materias)
2. rendimiento-categoria (Bajo/Medio/Alto)
3. numero-actividades (conteo por estudiante)

4. horas-totales-actividades (suma semanal)
5. mejora-trimestral (diferencia trimestre3 − trimestre1)
6. riesgo-academico (flag basado en múltiples criterios)

Variable a eliminar: Aquellas que consideres que no son necesarias para hacer Data Mining. De momento sólo sabemos hacerlo estudiando la correlación.

3 FASE 3: Transformación

Esta no debes hacerla por que aún no se ha estudiado en clase. Nos hace falta estudiar algunos modelos no supervisados, que veremos en la siguiente evaluación. De todas formas, lo dejo enunciado para que lo tengas en cuenta en los próximos proyectos.

3.1 Reducción de Dimensionalidad

1. Análisis de Componentes Principales (PCA) para variables numéricas
2. Selección de características basada en correlación
3. Análisis de importancia de variables con Random Forest
4. Creación de índices compuestos (ej: índice-socioacademico)

3.2 Dicretización y Binnig

Transformación de variables continuas

1. Discretizar edad en rangos (14-15, 16-17, 18)
2. Convertir calificaciones a categorías (Suspenso/Aprobado/Notable/Sobresaliente)
3. Agrupar horas de actividades en niveles (Bajo/Medio/Alto)
4. Crear segmentos de asistencia (Crítica/Regular/Buena/Excelente)

4 FASE 4: Minería de Datos (KDD)

De momento sólo podemos hacer minería de datos con algunos modelos

4.1 Regresión lineal múltiple/polinomial

Tienes que aplicar un modelo multilineal con el objetivo de cuantificar el impacto de la variables en la calificación. El modelo debes presentarlo como un patrón localizado por:

$$calificacion = \beta_0 + \beta_1 * estrato + \beta_2 * asistencia + \beta_3 * actividades$$

En el análisis debes estudiar:

1. Verificar supuestos (linealidad, normalidad, homocedasticidad)
2. Calcular coeficientes y p-valores
3. Interpretar R^2 y R^2 ajustado
4. Identificar variables significativas

4.2 Aplicaciones de K-nn

Debes aplicar K-nn para realizar el siguiente análisis

- Clasificar estudiantes en categorías de rendimiento
- Predecir calificación numérica usando K-NN regresión

Usar K-NN para recomendar actividades extracurriculares basado en:

- Estudiantes con perfiles académicos similares
- Actividades que han tenido impacto positivo en rendimiento
- Preferencias históricas de estudiantes similares

Agrupar estudiantes con características similares para:

- Formar grupos de estudio efectivos
- Compartir estrategias de aprendizaje