



Congreso Internacional de Tecnologías de la Información y Computación CITIC 2018



CITIC

29, 30 y 31
octubre 2018

Manta – Manabí
Sede: ULEAM

AUTORES:

Javier H. López Zambrano
jlopez@espam.edu.ec

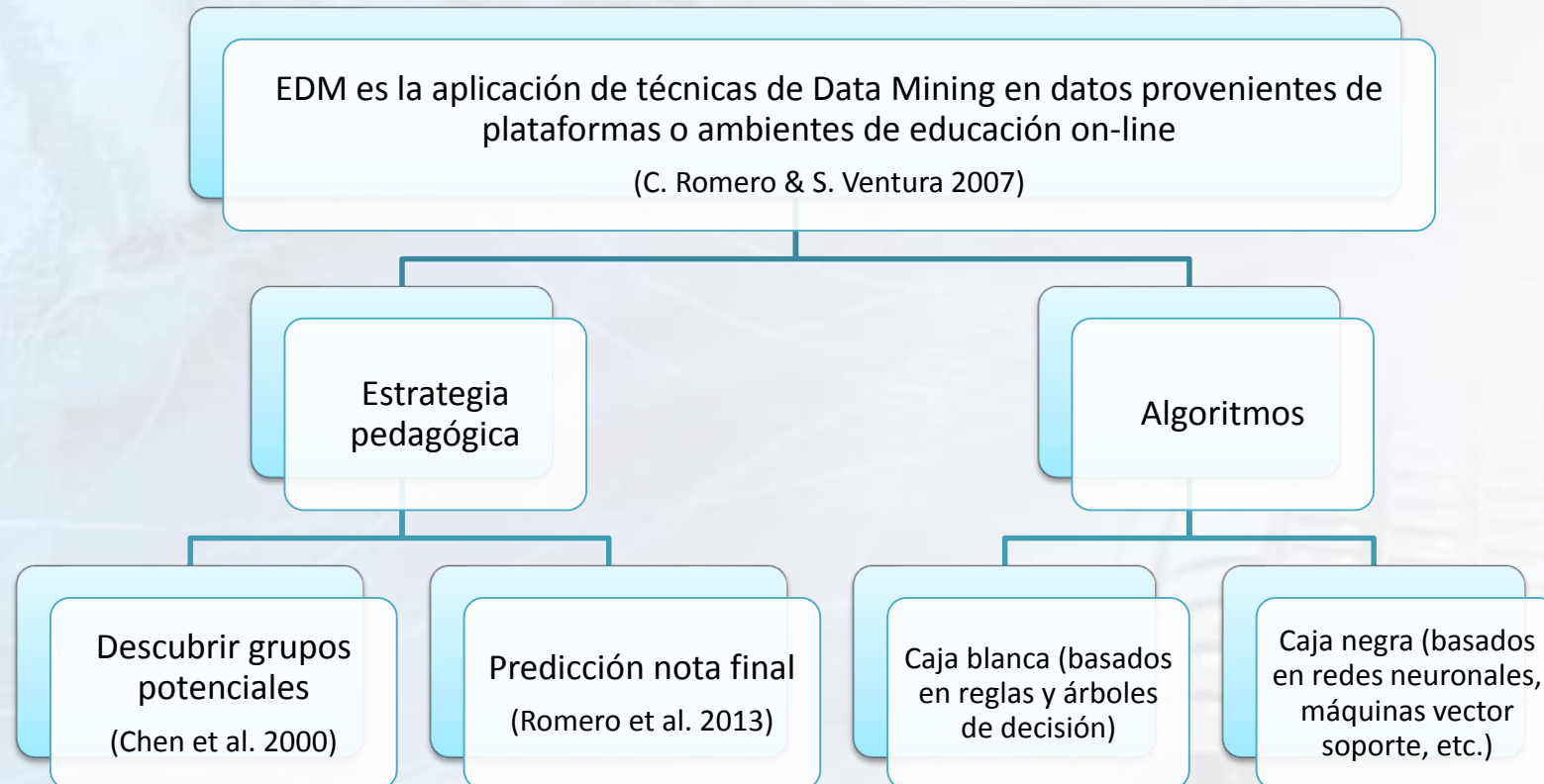
Doctorando en el Departamento de Ciencias de la Computación y Análisis Numérico de la Universidad de Córdoba [España], Magister en Telecomunicaciones por la Universidad Católica Santiago de Guayaquil, Ingeniero en Sistemas Computacionales, Profesor de Ciencias Informática, actualmente docente en la Facultad de Informática de la ESPAM-MFL e integrante del grupo de investigación SISCOM.

Ramón J. Moreira Pico
jmoreira@espam.edu.ec

Doctorando por la Universidad de Córdoba [España], Magister en Informática de Gestión y Nuevas Tecnologías por la Universidad Federico Santa María [Chile], desempeñó diferentes cargos en el área de tecnología en: GADM de Portoviejo, GADM de Chone, Ministerio de Transporte y Obras Públicas e IESS, actualmente docente titular en la ESPAM-MFL e integrante del grupo de investigación SISCOM.

Fernando R. Moreira Moreira
fmoreira@espam.edu.ec

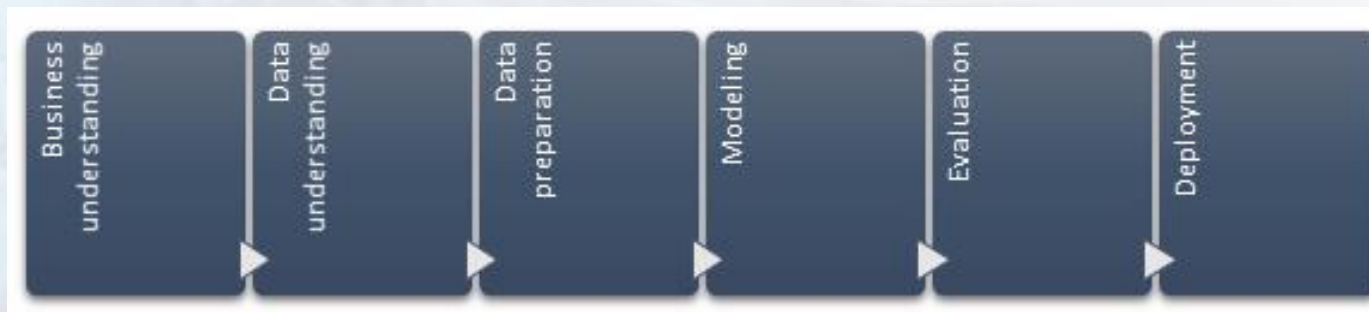
Master en Administración de Empresas por la Universidad Carlos III de Madrid [España], Desempeñó cargos en el área de tecnología en: SDK del Ecuador Cia. Ltda., actualmente docentes Ocasional en la ESPAM-MFL e integrante del grupo de investigación SISCOM.



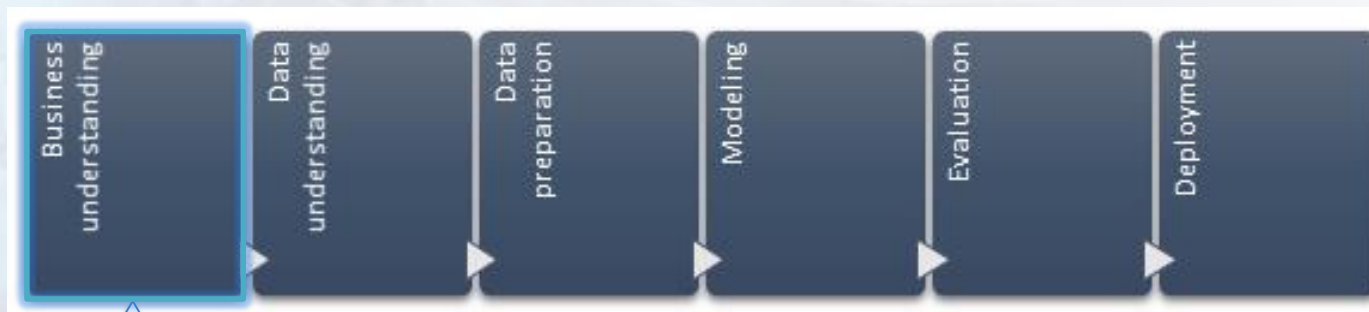
Comparar los Algoritmos-EDM utilizados en la predicción de rendimiento académico.

- Revisar investigaciones aplicando EDM
- Determinar los algoritmos más utilizados en la predicción de rendimiento académico
- Realizar pruebas (datos de los alumnos de Computación de la ESPAM-MFL) con los algoritmos determinados para contrastar los resultados de otras investigaciones
- Comparar los resultados en función de las métricas: Precisión de predicción, estadística de kappa, F-Media y el área bajo la curva ROC

Etapas de la metodología Cross Industry Standard Process for Data Mining (CRISP-DM)

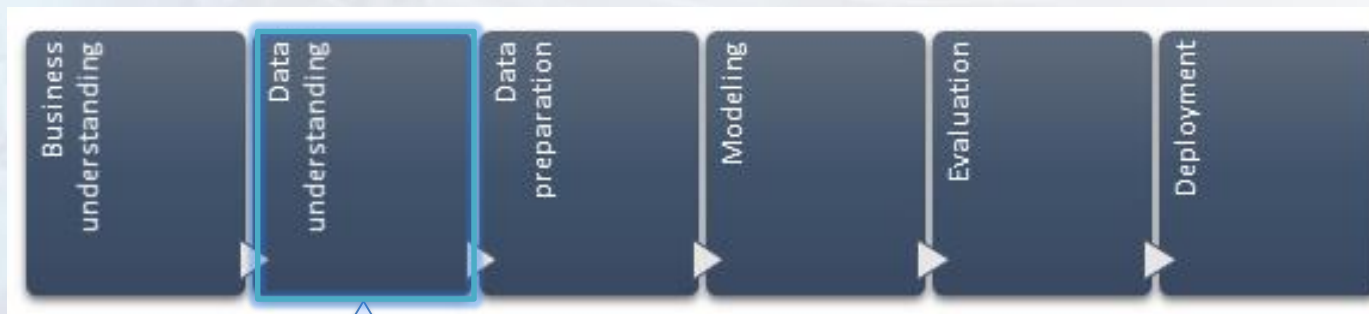


Etapas de la metodología Cross Industry Standard Process for Data Mining (CRISP-DM)



Comprensión de los datos dentro del entorno educativo y la EDM

Etapas de la metodología Cross Industry Standard Process for Data Mining (CRISP-DM)



Obtención de datos de los últimos 5 años, de la DB del sistema de gestión académica respecto a los alumnos de la carrera de Computación de la ESPAM-MFL

Etapas de la metodología Cross Industry Standard Process for Data Mining (CRISP-DM)



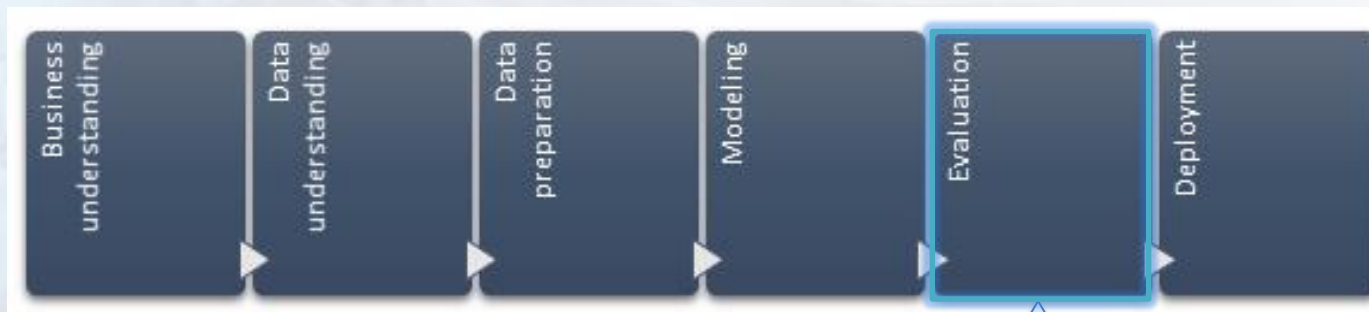
Preparación de los dataset, aplicando normalización, tipificación y discretización de los datos requeridos (variables predictivas y supervisada)

Etapas de la metodología Cross Industry Standard Process for Data Mining (CRISP-DM)



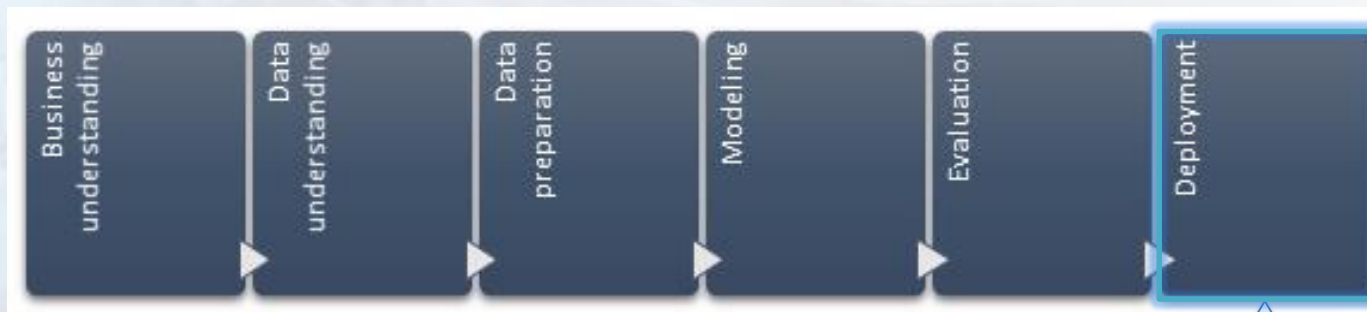
Ejecución de los algoritmos-EDM sobre los dataset obtenidos

Etapas de la metodología Cross Industry Standard Process for Data Mining (CRISP-DM)



Evaluar y validar los modelos y resultados de las ejecuciones realizadas

Etapas de la metodología Cross Industry Standard Process for Data Mining (CRISP-DM)

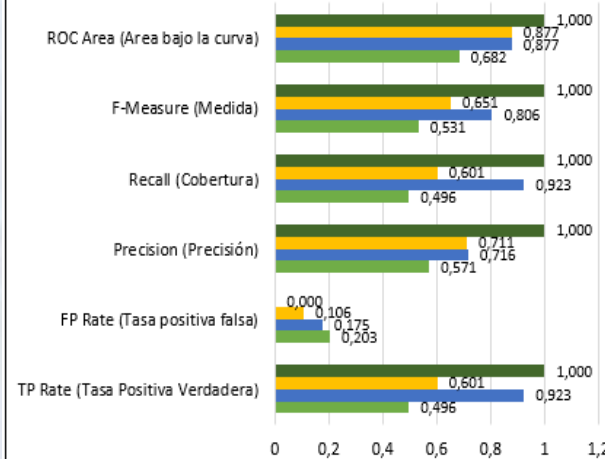


Análisis y preparación de los modelos obtenidos para inferir similitudes y diferencias

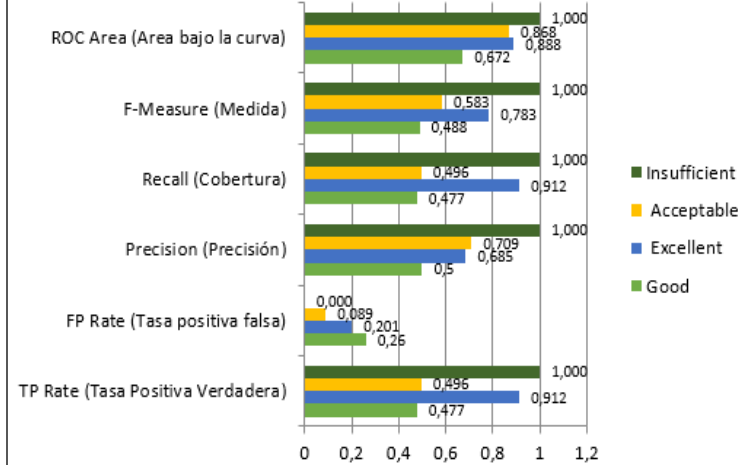
Promedio del área bajo la curva o ROC Area:

J48 0,811
 Naive Bayes 0,808
 Random Forest 0,794
 OneR 0,727

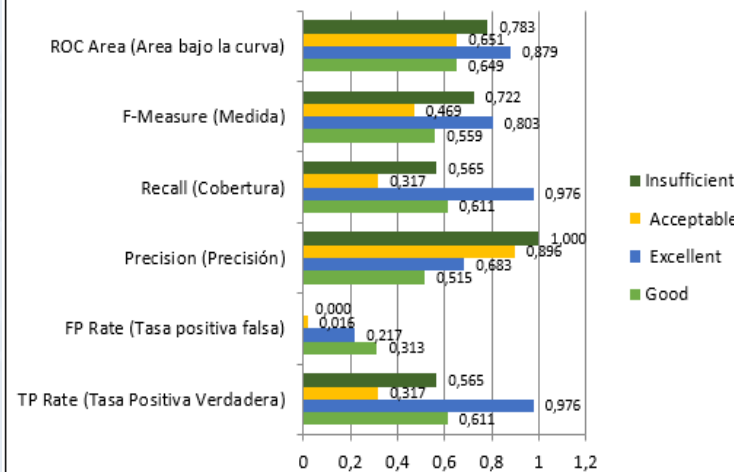
Detalles de la precisión por clase de J48



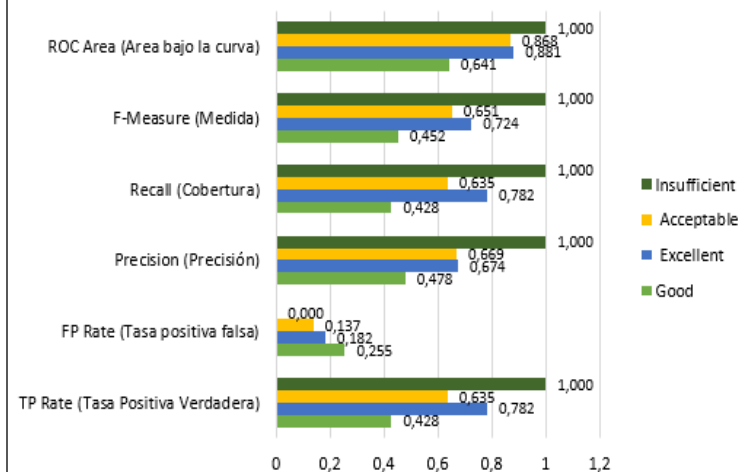
Detalles de la precisión por clase de Naive Bayes



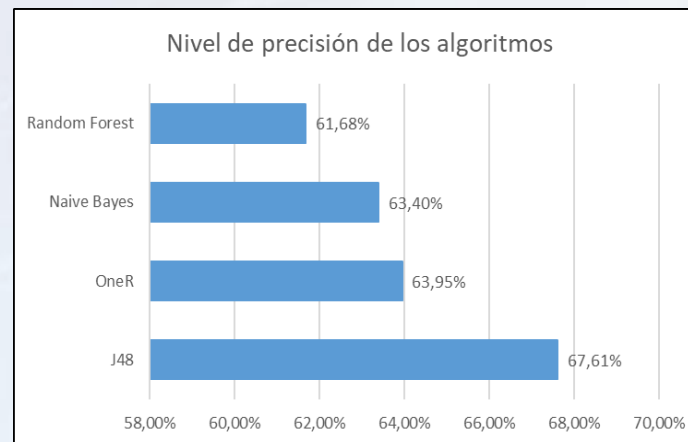
Detalles de la precisión por clase de OneR



Detalles de la precisión por clase de Random Forest



Métricas de Evaluación	J48	Naive Bayes	OneR	Random Forest
Instancias clasificadas correctas	787	738	744	718
Instancias clasificadas incorrectas	377	426	420	446
Precisión de la predicción (%)	67,61	63,4	63,91	61,68
Estadística de Kappa	0,52	0,45	0,46	0,43
Precisión	0,66	0,63	0,69	0,61
Cobertura	0,67	0,63	0,63	0,61
F-Medida	0,66	0,62	0,61	0,61
Área Bajo la curva ROC	0,81	0,80	0,72	0,79



- Generalmente se aplican diferentes técnicas de clasificación para construir un modelo de predicción de rendimiento.
- Según la revisión documental:
 - La frecuencia de uso para estas técnicas fueron: Asociación 8%, Agrupamiento 22%, y Clasificación 70%
 - Los algoritmos más utilizados fueron: J48, Naïve Bayes, OneR y Random Forest
- No existe un mejor o peor algoritmo, ya que esto dependerá de la capacidad predictiva de cada uno de ellos respecto a las características o especificidad de los datos y variables a supervisar

CONCLUSIONES

INVESTIGACIONES	J48	Random Forest	OneR	Naive Bayes
Minería de datos en egresados de la Universidad de Caldas [9]	92,04%		84,83%	81,98%
Predicción del rendimiento académico aplicando técnicas de minería de datos de [11]	68,3%			71,00%
Un enfoque de minería de datos para identificar los factores que afectan el éxito académico de los estudiantes terciarios en Sri Lanka [6]	96,52%	98,89%		91,22%
Minería de datos aplicada a la predicción de Rendimiento Académico [Autores]	67,61%	61,68%	63,95%	63,40%

- Se puede inferir que la capacidad predictivas de un algoritmos puede variar entre una investigación a otra por motivos de especificidad de los datos y variables
- Con la finalidad de contrasta los modelos usualmente en muchos artículos se utilizan más de un algoritmo para realizar análisis de minería de datos.



Congreso Internacional de Tecnologías de la Información y Computación CITIC 2018



CITIC

29, 30 y 31
octubre 2018

Manta – Manabí
Sede: ULEAM