ANÁLISIS COMPARATIVO DE TÉCNICAS DE MACHINE LEARNING PARA PREDECIR LA DESERCIÓN DE ESTUDIANTES EN VARIOS NIVELES DE ESTUDIOS

CARLOS ALBERTO PALACIOS ROJAS
MAGÍSTER EN GESTIÓN DE OPERACIONES

RESUMEN

Más del 50% de los matriculados en Chile en educación superior, no termina sus estudios en la primera carrera que se matricula. Esto genera importantes pérdidas de eficiencia para el Estado, las familias de los alumnos y las Universidades. Por este motivo se presenta un análisis comparativo de diversos algoritmos de Machine Learning para predecir la deserción en varios niveles y establece cuales son las variables significativas para los modelos. El estudio se dividió en dos etapas, la primera determina la deserción de los estudiantes, sin importar el tiempo en que suceda. La segunda considera la deserción en tres diferentes niveles por separado: Primer, Segundo y Tercer año. Los análisis muestran que el método Random Forest es el que mejor desempeño presenta. Los atributos más significativos de acuerdo a Information Gain resultaron ser las Notas de Educación Media e Índice de Pobreza Comunal, factores que de acuerdo al estado del arte no han sido aplicados en otros estudios de Minería de Datos aplicada a la Educación. Otro aporte de esta investigación, es la respuesta a una interrogante planteada por Arrau and Loiseau (2003) respecto de la deserción por quintiles de Ingreso económico. Palabras claves— Retención estudiantil, Random Forest, Minería de datos, Dashboard.

ABSTRACT

Over 50% of those enrolled in higher education in Chile, did not finalize studies in the first career they entered. This creates significant efficiency losses for the government, families of students, and universities. Therefore a methodology that assesses various Machine Learning algorithms to predict attrition and specifies what are the significant variables. The study was divided into two stages, the first determines the dropout of students, regardless of the time that passed. The second considers the desertion in three separate levels: First, Second and Third Year. Analyses show that the method, Ensemble Random Forest Classifier, delivers the best results. With respect to the attributes found within the Secondary education marks, the Poverty Index of communities, factors in the state of the art research had not previously been applied in other studies of Educational Data Mining. Another contribution of this research result is its response to a question raised by Arrau and Loiseau (2003) regarding the abandonment by economic quintiles. Keywords— Student retention, Random Forest, Data Mining, Dashboard.