Estimación de Riesgo de Fallo en Materias Basado en Similitud y Manejo de Incertidumbre Dada por Datos Académicos Históricos

Aníbal Vásquez Clad[(1)] , Enrique Peláez Jarrín[(2)]
Facultad de Ingeniería en Electricidad y Computación [(1)]
Escuela Superior Politécnica del Litoral (ESPOL)
Campus Gustavo Galindo, Km 30.5 vía Perimetral
Apartado 09-01-5863. Guayaquil-Ecuador
aniavasq@espol.edu.ec[(1)], epelaez@espol.edu.ec[(2)]

Resumen

La deserción en la educación superior es un problema que no solo afecta a los estudiantes, sino también a las universidades, pues al ofrecer educación incompleta se usan ineficientemente los recursos; uno de los factores relacionados es la habilidad interpersonal, la cual le permite a un estudiante entablar relaciones con profesores y compañeros, pues contribuyen a culminar de manera exitosa su carrera; el servicio de Consejerías Académicas que ofrece la ESPOL, permite a los estudiantes entablar una relación con un profesor consejero, quien brinda una guía para mejorar el desempeño mediante recomendaciones; las recomendaciones que brinda son basadas en su experiencia e información académica del estudiante, pero este proceso podría mejorar su precisión al agregar un factor de estimación del riesgo de reprobación que el estudiante afrontará durante el semestre. El siguiente trabajo propone una arquitectura que utiliza herramientas de clustering y prototipado difuso, para clasificación no-supervisada y predicción mediante extracción de variables descriptivas. Se desarrolló un software que implementa el modelo propuesto, el cual posee tres componentes principales: pre-procesamiento, clustering, y clasificación; dicho software permitió validar el modelo para predecir el riesgo de reprobación de estudiantes, basado en su carga académica y rendimiento, con un porcentaje de certeza significativo.

Palabras Claves: *Inteligencia Artificial, Prototipos Difusos, Lógica Difusa, Clasificación No-supervisada, Predicción.*

Abstract

The academic drop-out is not a problem that affect just to students, this affects universities as well, because failing to complete an educational degree is an inefficient use of resources for both students and the university; one of the factors related with the drop-out is the interpersonal competence, which allows a student establish a relationship with him/her teachers and classmates, and this could affect/contribute to the successful culmination of him/her academic career; ESPOL's Academic Counselling service, allows students to establish a relationship with their academic adviser, who provides guidance to improve the academic performance; the help provided by the advisers are based on their expertise and the student's academic information, but this process can be improved by adding "educated" information; for example, an estimation factor of risk for flunking a course in a semester. The following paper proposes an architecture for unsupervised classification based on clustering and fuzzy prototyping, then assessing the risk via extraction of descriptive variables. A prototype was developed implementing the proposed model, which have three principal components: pre-processer, clustering and classifier; this software allowed to validate the model in order to predict the risk, based on the academic load and performance, with a significant certainty performance.

Keywords*: Artificial Intelligence, Fuzzy prototypes, Fuzzy logic, unsupervised classification, Prediction.*

1. Introducción

El servicio de Consejerías Académicas es una herramienta de soporte para la gestión académica de los estudiantes, esta herramienta permite a los estudiantes tener una guía semestre a semestre por parte de los profesores asignados como consejeros, estas recomendaciones proveen al estudiante estrategias en cierto modo, sobre materias que debe tomar o no, que le permitirán culminar con éxito su carrera, el profesor consejero se basa en su experiencia y el perfil académico del estudiante para dichas recomendaciones; información tal como: promedio académico semestral, número de materias tomadas, aprobadas y reprobadas es usada para darle una retroalimentación al estudiante en el periodo de pre-registro para así soportar la toma de decisión del estudiante en materias a cursar; durante este período es posible agregar un factor de predicción que permita estimar el riesgo de reprobación al que un estudiante se afronta durante el semestre, permitiéndole elegir las materias que cursará de acuerdo a su historia académica y la carga académica que representan estas, y así mejorar su desempeño.

Para esto se propone un modelo para estimar el riesgo de fallo en un semestre para un estudiante, donde el fallo implica la reprobación de al menos una materia;

dicho modelo es soportado por datos académicos históricos.

Los datos académicos históricos son usados por el modelo para establecer la similitud entre estudiantes, mediante la extracción de características que describen el desempeño académico y la carga académica semestral, los estudiantes son categorizados con estas medidas extraídas, y usando variables descriptivas sobre estas categorías es asociado un valor de riesgo de reprobación dependiendo de la carga académica que representa el semestre.

El modelo propuesto utiliza una arquitectura basada en clustering y prototipado difuso, para la clasificación no-supervisada y predicción mediante asociación de variables descriptivas.

El clustering es una técnica que permite realizar sub-agrupaciones llamadas clusters en un conjunto, permitiendo clasificar los elementos de acuerdo a una similitud entre ellos; esto se puede definir formalmente como [1]:

$$C_r = \{y_1, y_2, \dots, y_m\} \quad y_i \in X, i = 1,2, \dots, m \qquad (1)$$

$$X = \bigcup_{r=1}^{c} C_r \qquad (2)$$

Donde C_r es un cluster tal que $C_r \subseteq X$, y X es el conjunto tal que $X \subseteq \mathbb{R}^p$.

Existe una generalización de esta técnica hacia conjuntos difusos, esta generalización establece que cada cluster puede ser representado como un conjunto difuso [2], permitiendo definir una pertenencia parcial de elementos a un cluster e incluso la pertenencia simultanea de un elemento a más de un cluster a la vez.

El prototipado difuso es un método que permite construir elementos que pueden representar de forma general otros elementos pertenecientes a un conjunto difuso, este método hace uso de las técnicas de clustering para la construcción de los prototipos que describen cada cluster, y en este contexto los clusters representan las diferentes categorías dentro de un conjunto [3].

Este documento está organizado de la siguiente forma: la sección 2 describe el algoritmo de clustering difuso usado, la técnica usa para su validación y la creación de prototipos difusos; la sección 3 describe las variables extraídas en el contexto académico; la sección 4 describe el diseño de la arquitectura propuesta para la estimación de riesgo de reprobación basada en prototipos difusos y clasificación no-supervisada; la sección 5 presenta los resultados obtenidos, y la sección 6 describe las conclusiones.

2. Prototipado difuso y medidas de validación

El prototipado difuso para la construcción de elementos representativos se basa en una noción de tipicidad; esta es una medida asociada a cada elemento que pertenece a un cluster y permite determinar qué tan representativo es, a través de las medidas de semejanza y disimilitud, que se definen en función de las distancias entre el elemento y los demás elementos del cluster, y entre el elemento y los elementos que pertenecen a otros clusters, respectivamente. Rifqi [4] propone un método para determinar el grado de tipicidad mediante la agregación de la semejanza y tipicidad, definido como:

$$t_{ir} = \begin{cases} \Phi\big(R_r(x_i), D_r(x_i)\big), x_i \in C_r \\ 0, x_i \notin C_r \end{cases} \qquad (3)$$

Donde:

$$R_r(x_i) = \frac{1}{|C_r|} \sum_{y \in C_r} \rho\,(x_i, y) \qquad (4)$$

$$D_r(x_i) = \frac{1}{|X/C_r|} \sum_{z \notin C_r} \delta\,(x_i, z) \qquad (5)$$

$R_r(x_i)$ en (4) describe la semejanza asociada al elemento x_i dentro del cluster C_r, y definida por las distancias $\rho(x_i, y)$ entre x_i y los demás elementos del cluster.

$D_r(x_i)$ en (5) describe la medida de dismilitud asociada a un elemento x_i dentro del cluster C_r, definida por las distancias $\delta(x_i, y)$ entre x_i y los demás elementos que no pertenecen al cluster C_r; generalmente esta es una medida calculada como el complemento de la semejanza.

Una vez determinadas las medidas de tipicidad de cada elemento, el prototipo difuso para el cluster C_r se construye a través de la agregación sobre todos los elementos cuya tipicidad cumple con la condición de umbral definida como [5]:

$$w_r = \psi(\{x_i / t_{ir} > \tau\}) \qquad (6)$$

Donde τ representa la medida de umbral para la tipicidad en los elementos del cluster C_r.

Existen varios algoritmos de clustering difuso que permiten construir prototipos; Fuzzy C-Means (FCM) es un algoritmo donde los grados membresía asociados a cada elemento, describen la pertenencia a cada cluster y cuan difuso son los límites de este [6], estos grados de membresía son definidos como:

$$u_{ir} = \left(\sum_{s=1}^{c} \left(\frac{d_{ir}}{d_{is}} \right)^{2/(m-1)} \right)^{-1} \qquad (7)$$

Donde x_i pertenece al cluster C_r, $d_{ir} = \|x_i - w_r\|_A$ define la distancia entre x_i y el prototipo w_r del cluster C_r, $d_{is} = \|x_i - w_s\|_A$ define la distancia entre x_i y el prototipo x_s del cluster C_s, m es el exponente de ponderación, y c es el número de clusters. Este algoritmo puede ser visto como un proceso de media ponderada de forma iterativa [7] y es considerado como una optimización de la semejanza.

Para el proceso de clustering existen medidas de validación que permiten determinar el número óptimo de clusters c, y así definir un proceso de categorización no-supervisada [8], estas medidas surgen como una necesidad para validación de clusters en conjuntos multidimensionales, pues para estos casos la verificación por visualización se vuelve complicada. Una de estas medidas es el coeficiente de partición de Dave [9], este es resultado de una modificación del coeficiente de partición de Bezdeck [2] para manejar la monotonicidad, esta medida es definida formalmente como:

$$V_{MPC} = 1 - \frac{c}{c-1}(1 - V_{PC}) \quad (8)$$

Donde c es el número de clusters del proceso a validar y V_{PC} es el coeficiente partición de Bezdeck definido como:

$$V_{PC} = \frac{1}{n}\sum_{r=1}^{c}\sum_{i=1}^{n}(u_{ir})^2 \quad (9)$$

El objetivo del coeficiente de Dave es la optimización del número de clusters c con $\min\{V_{MPC}\}$ [8].

El número de clusters obtenido en el proceso de validación permite definir las categorías en el proceso de clustering, para así construir los prototipos que pueden ser usados como datos de entrenamiento para una Máquina de Soporte de Vectores (Support Vector Machine) [10] y técnicas de defusificación [11], a través de los cuales podemos clasificar nuevos datos y predecir su comportamiento mediante las variables descriptivas asociadas a cada cluster.

3. Estimación de medidas en el contexto aca-démico

Las medidas en contexto académica son producto de la extracción de características del historial académico de estudiantes, estas son usadas en el proceso de clustering y permiten abstraer las variables asociadas a carga académica semestral y rendimiento de un estudiante.

Es posible describir la carga académica semestral mediante medidas relacionadas a qué tan difícil es aprobar las materias dentro de un semestre; Caulkins et. al [12] introdujeron las medidas de dificultad y rigurosidad relacionadas con una materia, estas describen qué tan difícil es aprobar un curso y cuál es la rigurosidad con la que se califica, estas medidas dependen del promedio de calificaciones de los estudiantes y las calificaciones que obtienen en la materia a medir.

La medida de dificultad es definida formalmente como:

$$\alpha_j = \frac{\sum_i GPA_i^2}{\sum_i (r_{ij} \cdot GPA_i)} \quad (10)$$

Así mismo la medida de rigurosidad es definida formalmente como:

$$\beta_j = \frac{\sum_i (GPA_i - r_{ij})}{N_S^j} \quad (11)$$

Donde GPA_i representa el promedio de calificaciones del estudiante i; r_{ij} es la calificación del estudiante i en la materia j.

Otra media relacionada a las materias es la medida de distribución de notas, esta fue introducida por Méndez et. al [13] y describe la asimetría en la distribución estadística de las distancias entre la calificación que obtiene un estudiante en una materia y su promedio de calificaciones; esta se define como:

$$Sk_j = skewness_i(GPA_i - r_{ij}) \quad (12)$$

Estas medidas que se encuentran relacionadas a las materias son usadas para describir la carga académica semestral, tomando en cuenta que un semestre es un conjunto de materias que el estudiante cursa, se propone la carga académica asociada a un semestre como un vector que está en función de las medidas de dificultad, rigurosidad y distribución de notas de cada materia en el semestre, el número de materias que se cursan, y el número del semestre que cursa el estudiante desde su ingreso.

De la misma forma es posible describir el rendimiento académico de un estudiante; mediante la historia académica de un estudiante se pueden extraer medidas que describan las habilidades que adquiere al aprobar materias; bajo la hipótesis de la existencia de estructuras dentro de una malla curricular, Méndez et. al [13] propone el uso de un Análisis Exploratorio de Factores [14] para develar estas estructuras en la malla de la carrera de Ingeniería en Ciencias Computacionales de ESPOL; este análisis dio como resultado 5 factores que componen el núcleo de la malla curricular, estos son:

1. *Factor de preparación en Ingeniería Básica*, el cual agrupa materias relacionadas a ciencias básicas como: cálculo, física y química; esta permite el desarrollo de la habilidad para entender conceptos abstractos, lógica, matemáticas e interpretación de fenómenos.
2. *Factor de Temas Avanzados de Ciencias Computacionales*, este factor agrupa materias que describen el desarrollo de la habilidad para comprender e interpretar conceptos como Ingeniería de Software, Organización y Arquitectura de Computadores, interacción humano-Computadora o Inteligencia Artificial.
3. *Factor de Interacción con el Cliente*, agrupa materias relacionas a la habilidad para

comunicarse con usuarios finales y clientes, parte del proceso de desarrollo de software.

4. *Factor de Programación*, agrupa materias que ayudan a desarrollar las habilidades de programación, conceptos, estrategias y patrones de diseño necesarios en el proceso de desarrollo de software.
5. *Factor de cursos no estrechamente relacionados a las Ciencias Computacionales*, está formado por materias relacionadas en su mayoría a la ingeniería eléctrica.

Basado en estos factores se propone una medida que describa las habilidades adquiridas mediante el desempeño de un estudiante en las materias relacionadas a un factor, esta medida propuesta se define como:

$$f_{k_i} = \frac{n_{a_k}}{n_{e_k}} \cdot \sum_{j \in F_k} \frac{r_{ij}}{n_{e_k}} \qquad (13)$$

Donde n_{a_k} es le número de materias aprobadas por el estudiante i en el factor F_k, n_{e_k} es el número de materias tomadas en el factor F_k, y r_{ij} es la calificación del estudiante i en la materia j.

La medida de habilidades en F_k está formada por dos componentes multiplicándose, el primero describe la eficiencia del estudiante para aprobar las materias en el factor, y el segundo describe el rendimiento según las calificaciones del estudiante en dichas materias.

Así el rendimiento académico de un estudiante puede ser descrito mediante un vector de características formado por cada una de las medidas en los factores que componen la malla curricular.

La carga académica asociada a un semestre y el rendimiento asociado a un estudiante pueden ser usados en un proceso de clustering, permitiendo identificar las categorías de estudiantes en combinación con semestres, para así poder clasificar y asociar el riesgo de reprobación a un estudiante que desea cursar un semestre similar a los casos ya clusterizados.

4. Diseño

El software diseñado para la estimación del riesgo de reprobación está basado en un arquitectura de componentes, soportada por prototipos difusos y clasificación no-supervisada, esta consta de tres componentes principales: un componente de pre-procesamiento, encargado de la selección y extracción de características en los datos académicos históricos; un componente de clustering, a través del cual se construyen los prototipos y se generan las categorías en un proceso de clsutering; y un componente de clasificación, el cual asocia las variables descriptivas de reprobación académica a cada cluster y categoriza los nuevos datos de estudiantes, asociándolos a un riesgo de reprobación.

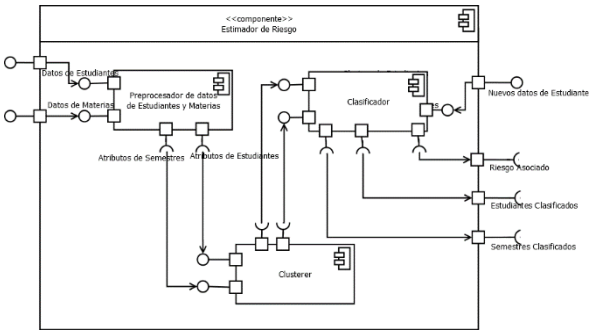

Figura 1. Diagrama de componentes del software de estimación de riesgo de reprobación

5. Resultados

Para analizar el nivel de precisión de las estimaciones, se elaboraron pruebas usando los datos de los estudiantes de Ingeniería en Ciencias Computacionales registrados en los años de 1978 y 2013.

Uno de los experimentos realizados permitió determinar el número adecuado de categorías de estudiantes en el proceso de clustering, la maximización del índice de Dave indica que el número de clusters para estudiantes basado en su rendimiento es de 5 con un índice de validación igual a 0.860776; la figura 2 se puede observar un gráfico de radar con las medidas en las habilidades de los estudiantes para cada uno de los clusters obtenidos.

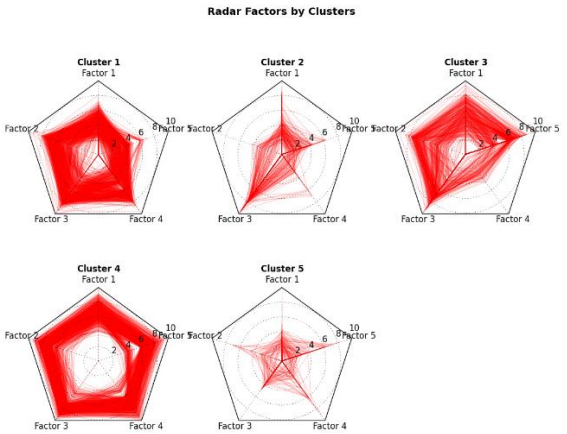

Figura 2. Gráfico de radar para las medidas de habilidades en los clusters de estudiantes

El primer cluster está formado por estudiantes que tienen una deficiencia en el factor de materias no estrechamente relacionadas con ciencias computacionales; en el segundo cluster se encuentran estudiantes que poseen un desempeño aceptable en materias relacionadas con la interacción con el cliente, pero no poseen un buen desempeño en las demás medidas; el tercer cluster muestra a estudiantes que no poseen un buen desempeño en materias relacionadas con la programación; el cuarto cluster está formado por

estudiantes que poseen un desempeño equilibrado en las medidas de habilidades; y en el quinto cluster se encuentran estudiantes que no poseen un buen desempeño en los factores medidos.

Para analizar el nivel de confiabilidad se llevó a cabo un análisis mediante el Score de Brier [15], este experimento es realizado para probar el nivel de certeza que puede manejar le modelo implementado; la verificación mediante el Score de Brier permite medir la precisión de un modelo de predicción, mediante el contraste entre la probabilidad predicha y la frecuencia relativa observada, por lo que la verificación en este caso se realizó sobre el pronóstico de riesgo de reprobación. El Score de Brier puede ser descompuesto en tres términos [16] que describen: la incertidumbre inherente de que ocurra el evento, la confiabilidad dada por qué tan cercana es la probabilidad pronosticada de la real, y la resolución con la que difieren la probabilidad de los diferentes pronósticos y el promedio; estos términos son usados para determinar el score de skill de Brier el cual mide la diferencia entre el score de la predicción y el score de la predicción de un modelo no cualificado; en los casos en los que el score de skill alcanza valores negativos se considera al modelo como no cualificado para predicción.

En este experimento se entrenó el modelo con los datos de los estudiantes hasta el año 2012, usado como modelo estándar no cualificado, la estimación se realizó para los términos I y II del año 2013 y se contrastó con la observación de reprobación real en dichos términos.

Para el análisis del I término del 2013, se encontró un score Brier de 0.2164, cercano a cero, un score de skill positivo de 0.1304; con una confiabilidad de 0.01001, una resolución de 0.04256 y un porcentaje de incertidumbre de 24.88%; esto quiere decir que para el primer semestre del 2013 se pudo realizar una estimación de riesgo de reprobación con una certeza del 75.12%.

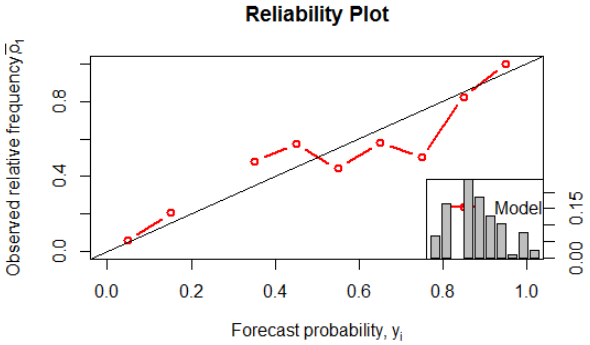

Figura 3. Confiabilidad para la estimación de riesgo de reprobación en estudiantes Ciencias Computacionales en el término académico 2013 I

Mientras que en el análisis del segundo término 2013 se encontró un score de Brier de 0.2422, un score de *skill* igual a 0.03113, también positivo, con una confiabilidad de 0.01486, una resolución de 0.02264, y una incertidumbre de 24.99%; esto quiere decir que para el segundo término se encontró que la estimación en el riesgo de fallo posee una certeza del 75.01%.

Estos resultados nos dicen que el modelo propuesto permite estimar el riesgo de reprobación, para estudiantes de Ingeniería en Ciencias Computacionales, con una certeza que bordea el 75%, además podemos decir que el modelo posee una confiabilidad significativa en ambos términos del 2013, también se puede apreciar que hay una mayor resolución en el primer término que la estimación del riesgo en el segundo término, por lo que el pronóstico en la primer término difiere menos que el pronóstico en el segundo término.

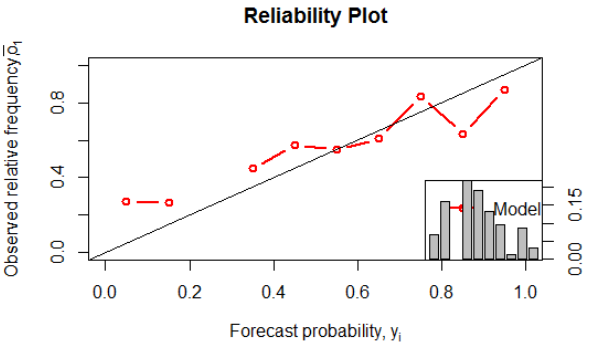

Figura 4. Confiabilidad para la estimación de riesgo de reprobación en estudiantes Ciencias Computacionales en el término académico 2013 II

6. Conclusiones

En este trabajo se ha propuesto un modelo de manejo de incertidumbre que permita estimar el riesgo de fallo en materias durante el semestre, este modelo está basado en prototipos y clasificación no supervisada, soportada por algoritmos de clustering, los cuales permiten construir prototipos dentro de un conjunto de datos, a través de valores de tipicidad asociados a cada elemento; estos prototipos representan los elementos más distintivos de cada cluster; además, pueden ser usados como datos de entrenamiento para la clasificación no-supervisada de nuevos datos

El proceso de clustering debe de ser validado en los casos en los que no se conozca de manera previa el número de categorías que posea el conjunto de datos, junto con los parámetros estimados por el proceso de validación.

La selección de las características que serán usadas en el proceso de clustering juegan un papel muy importante, ya que deben de aportar a la semántica que proveen los prototipos obtenidos en este proceso.

En este caso se pudo describir el rendimiento de un estudiante gracias a las medidas que representan las habilidades ganadas al aprobar materias, estas medidas permitieron definir la similitud entre estudiantes en el proceso de clustering difuso; además la carga académica semestral pudo ser descrita mediante las

medidas de dificultad, rigurosidad y distribución de notas que están asociadas a una materia, y así determinar las categorías de semestres a las que un estudiante se puede afrontar.

Con el diseño propuesto para el prototipo de estimación, se obtuvo un porcentaje de certeza significativo en resultados experimentales sugiriendo que el modelo planteado, posee un nivel de respuesta aceptable en la estimación del riesgo de reprobación en al menos una materia durante el semestre para estudiantes de la carrera de Ingeniería en Ciencias Computacionales de la ESPOL; esto quiere decir que si dicho módulo es implementado e implantando en el sistema de Consejerías Académicas podría proveer de información para retroalimentar a los profesores consejeros con respecto al riesgo de falla durante el periodo de pre-registro.

Para trabajos futuros se sugiere explorar la generalización de este modelo de estimación basado en prototipos difusos, ya que el caso de estudio en este trabajo está orientado al contexto académico, pero sería de gran interés aplicarlo a casos que no necesariamente se encuentren relacionados al ámbito académico. Además los resultados obtenidos utilizando el modelo propuesto muestran un porcentaje de certeza significativo, a pesar de que los datos provistos para entrenamiento solo poseían información académica, por lo que una estrategia a futuro es la incorporación de otros factores relacionados con carga extra-universitaria, datos socioeconómicos o información demográfica, para así mejorar el nivel de precisión de la estimación del riesgo de reprobación.

Referencias

[1] Kaufman, L., & Rousseeuw, P. J. Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons (2009).

[2] Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers (1981).

[3] Lesot, M. J. Similarity, typicality and fuzzy prototypes for numerical data. In 6th European Congress on Systems Science, Workshop Similarity and resemblance Vol. 94. (2005) 95-96.

[4] Rifqi, M., Berger, V., & Bouchon-Meunier, B. Discrimination power of measures of comparison. Fuzzy sets and systems, 110(2), (2000) 189-196.

[5] Lesot, M. J., Rifqi, M., & Bouchon-Meunier, B. Fuzzy prototypes: From a cognitive view to a machine learning principle. InFuzzy Sets and Their Extensions: Representation, Aggregation and Models. Springer Berlin Heidelberg (2008) 431-452.

[6] Bezdek, J. C., Ehrlich, R., & Full, W. FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2) (1984) 191-203.

[7] Lesot, M. J., Mouillet, L., & Bouchon-Meunier, B. Fuzzy prototypes based on typicality degrees. In Computational Intelligence, Theory and Applications (pp. 125-138). Springer Berlin Heidelberg (2005).

[8] Wang, W., & Zhang, Y. On fuzzy cluster validity indices. Fuzzy sets and systems, 158(19), (2007) 2095-2117.

[9] Dave, R. N. Validating fuzzy partitions obtained through c-shells clustering. Pattern Recognition Letters, 17(6), (1996) 613-623.

[10] Suykens, J. A., & Vandewalle, J. Least squares support vector machine classifiers. Neural processing letters (1999) 293-300.

[11] Krishnapuram, R., & Keller, J. M. A possibilistic approach to clustering. Fuzzy Systems, IEEE Transactions on (1993) 98-110.

[12] Caulkins, J. P., Larkey, P. D., & Wei, J. Adjusting GPA to reflect course difficulty (1996).

[13] Méndez, G., Ochoa, X., & Chiluiza, K. Techniques for data-driven curriculum analysis. In Proceedins of the Fourth International Conference on Learning Analytics And Knowledge. ACM (2014) 148-157.

[14] Fabrigar, L. R., & Wegener, D. T. Exploratory factor analysis. Oxford University Press (2011).

[15] Brier, G. W. Verification of forecasts expressed in terms of probability.Monthly weather review, 78(1), (1950) 1-3.

[16] Murphy, A. H. A new vector partition of the probability score. Journal of Applied Meteorology, 12(4), (1973) 595-600.