

Recommender System of Enrollment in Assignments Based on Profiles of Teachers, Students and Subjects in the Professional School of Systems Engineering of the National University of San Agustín

Jerson Erick Herrera Rivera, Estudiante¹, Eveling Castro Gutiérrez, Magíster¹

¹Universidad Nacional de San Agustín de Arequipa, Perú, jherrerar@unsa.edu.pe, ecastro@unsa.edu.pe

Abstract— *One of the main problems faced by university students (specifically the case of the Professional School of Systems Engineering of the National University of San Agustín) is to make the right decision in relation to the subjects to be enrolled based on available information (subjects, syllabus, schedules, content of the subject, teacher, and others), a situation that can be generalized in all universities. Under these circumstances, this research work seeks to develop a Recommendation System based on mining techniques that will give students the support to choose which subjects they should enroll in and obtain better results in the academic field. The recommendation given will be based on the previous experience obtained from each student enrollment. With the information obtained, a student profile, a subject profile and, as far as possible, a teacher profile are created. To reach the objective, data has been analyzed of the students of the aforementioned school, between 2011 and 2016. The results obtained indicate that data mining techniques (algorithms based on rules and tree-based algorithms) do not adequately represent the attributes of teachers or of subjects, unlike the models of recommendation systems (collaborative models and content-based models)*

Keywords—*Recommender System, academic performance prediction, collaborative models, content-based models, hybrid models.*

Digital Object Identifier (DOI):<http://dx.doi.org/10.18687/LACCEI2018.1.1.127>

ISBN: 978-0-9993443-1-6

ISSN: 2414-6390

Sistema de Recomendación de Matrículas en Asignaturas Basados en Perfiles de Docentes, Alumnos y Asignaturas en la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional de San Agustín

Jerson Erick Herrera Rivera, Estudiante¹, Eveling Castro Gutiérrez, Magíster¹

¹Universidad Nacional de San Agustín de Arequipa, Perú, jherrerar@unsa.edu.pe, ecastro@unsa.edu.pe

Abstract– Uno de los principales problemas que enfrentan los estudiantes universitarios (en específico el caso de la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional de San Agustín de Arequipa) es tomar la decisión correcta en relación a las asignaturas a matricularse basado en información disponible (asignaturas, sílabus, horarios, contenido de la asignatura, docente, y demás). Bajo estas circunstancias, el presente trabajo de investigación busca desarrollar un Sistema de Recomendación basados en técnicas de minería que dé soporte a los estudiantes a elegir en qué asignaturas deberían matricularse y obtener mejores resultados en el ámbito académico. La recomendación dada estará basada en la experiencia previa obtenida de cada matrícula de estudiantes; con tal información se crea un perfil de estudiante, un perfil de asignatura y en lo posible, un perfil de docente. Para alcanzar el objetivo se ha analizado data de los estudiantes de la escuela mencionada, entre los años 2011 y 2016. Los resultados obtenidos indican que las técnicas de minería de datos (algoritmos basados en reglas y algoritmos basados en árboles) no representan adecuadamente los atributos de docentes ni los de asignaturas, a diferencia de los modelos de los sistemas de recomendación (modelos colaborativos y modelos basados en contenido).

Keywords– Recommender System, predicción rendimiento académico, modelos colaborativos, modelos basados en contenido, modelos híbridos.

I. INTRODUCCIÓN

La Escuela Profesional de Ingeniería de Sistemas (EPIS) de la Universidad Nacional de San Agustín cuenta con una malla curricular de una cantidad considerable de asignaturas, muchas de los cuales son requisitos unas de otras, considerando además que cuenta con líneas de asignaturas electivas desde cuarto año.

Los estudiantes tratan de escoger las mejores asignaturas para ellos (tanto obligatorias como electivas) (asignaturas interesantes acorde a sus preferencias o asignaturas en las cuales su rendimiento es mayor). Sin embargo encontrar dichos cursos implica tareas que necesitan tiempo para analizarlas y muchas veces escapan de sus posibilidades. Los estudiantes tienen que buscar todos los contenidos de cada curso, examinar cuidadosamente los sílabus de los cursos que

por lo general son muy abstractos y generales, y no están a su disposición sino hasta después del proceso de matrículas, comprender el alcance del curso se vuelve una tarea compleja, incluso tienen que revisar estadísticas de la asignatura o del docente, o pedir consejos a diferentes estudiantes que ya cuenten con la experiencia de la asignatura, aunque los comentarios podrían ser demasiado subjetivos dependiendo de dicha experiencia.

Tomar malas decisiones a la hora de escoger en qué asignaturas uno se matricula significa una decisión crítica que debería ser analizada concienzudamente y con la suficiente información (en cuanto a cantidad y calidad), ya que de lo contrario la tasa de desaprobación y deserción seguirá siendo alta como puede ser corroborado por las estadísticas obtenidas de la EPIS. Es necesario por tanto una herramienta que sugiera adecuadamente a los estudiantes en qué asignaturas deberían matricularse basados en el perfil del estudiante, en el perfil de la asignatura y en el perfil del docente a cargo, obteniéndose una recomendación objetiva y exacta.

El presente trabajo está organizado como sigue: la sección 2 da un overview de trabajos relacionados aplicando técnicas de data mining y modelos de sistemas de recomendación en el ámbito educativo, principalmente en la predicción de rendimiento académico universitario. En la sección 3 se describe los objetivos que se buscan alcanzar con el desarrollo de la investigación y las contribuciones a la escuela profesional, tanto a estudiantes, docentes y administrativos. La sección 4 detalla los resultados (precisión y eficiencia) obtenidos utilizando los algoritmos basados en reglas y basados en árboles. La sección 5 se explica el funcionamiento y resultados del modelo colaborativo de los sistemas de recomendación. Finalmente, la sección 6 describe las conclusiones alcanzadas.

II. TRABAJOS RELACIONADOS

En “Recommendation in Higher Education Using Data Mining Techniques” [4] del 2009, se identifica como un problema de los estudiantes de universidad el tomar la

Digital Object Identifier (DOI): <http://dx.doi.org/10.18687/LACCEI2018.1.1.127>

ISBN: 978-0-9993443-1-6

ISSN: 2414-6390

decisión correcta en relación a su itinerario académico basado en información disponible (asignaturas, horarios, aulas y docentes). Este trabajo propone el uso de un RS basado en técnicas de data mining [5] para ayudar a los estudiantes en este tipo de decisión. Para ello se trabajó con data real correspondiente a siete años de la Escuela de Ingeniería de Sistemas de la Universidad de Lima. Tras cuatro ensayos se alcanzó una exactitud del 77.3 %. Utilizaron la técnica de Decision Tree [6], los cuales fueron creados a partir de una Base de Datos de la Escuela, para generar reglas. Finalmente, el sistema genera recomendaciones basadas en estas reglas.

En [7] se diseñó un framework de un RS Inteligente que puede predecir el desempeño académico del primer año de estudiantes de educación terciaria, orientando así a la gestión de la institución educativa en su toma de decisiones sobre estrategias de intervención temprana. Usaron data obtenida de los archivos de estudiantes de la Universidad de Babcock, Nigeria. De tales estudiantes, se tomó información relacionada a su familia, desempeño educativo previo a la universidad y resultado del examen de ingreso a la universidad. Utilizaron para el estudio decision trees y Multilayer Perceptron [8] (red neuronal artificial) para generar modelos; alcanzando una exactitud de 96.78 %. Similarmente, en [9], se hace uso de información del background de los estudiantes para analizar su rendimiento en el primer año de estudios.

En [10] se desarrolló un RS para determinar la productividad de aprendizaje del estudiante basado en parámetros fisiológicos, psicológicos y de comportamiento/movimiento. Se genera miles de recomendaciones basadas en Persistencia educacional, motivacional y tablas de aprendizaje social y selecciona la opción más razonable de estas para una situación específica del estudiante. El RS proporciona a un estudiante una evaluación en tiempo real de su propia productividad de aprendizaje e interés en el aprendizaje.

Diferentes estudios han buscado predecir el desempeño académico de los estudiantes para identificar a los estudiantes que necesitan ayuda adicional para mejorar su rendimiento académico y, a su vez, su promedio de calificaciones en la graduación. Un ejemplo de ello, es [11] en donde se evaluó a 127 estudiantes utilizando técnicas de minería de datos a partir de su rendimiento en los dos primeros años de estudio. Este estudio forma parte de Educational Data Mining (EDM).

A diferencia de los trabajos descritos previamente, el presente a descubierto que las técnicas de data mining como las basadas en reglas y las basadas en árboles, no muestran adecuadamente la relevancia que los atributos de asignaturas y de docentes tienen a la hora de predecir el rendimiento académico. Así, la metodología utilizada en el presente trabajo, son exclusivamente propias del enfoque de sistemas de recomendación; haciendo uso de los modelos colaborativos, modelos basados en contenido y posteriormente, su hibridación, según el enfoque que garantice eficiencia y precisión.

III. PROPUESTA DE SOLUCIÓN

Los Sistemas de Recomendación [1] (Recommender Systems - RS) son técnicas y herramientas de software que brindan sugerencias de ítem para ser usados por un usuario. Las sugerencias dadas por un RS [2] están dirigidas a apoyar a los usuarios procesos de toma de decisiones, tales como qué productos comprar, qué música escuchar, o qué noticias leer. RS son medios valiosos para que los usuarios online puedan hacer frente a la sobrecarga de información y ayudarlos a tomar mejores decisiones. Ahora son una de las más poderosas y populares herramientas de descubrimiento de información en la web.

A. *Objetivo General*

Diseñar una arquitectura de Sistema de Recomendación que sugiera adecuadamente a los estudiantes en qué asignaturas deberán matricularse basados en el perfil del estudiante, en el perfil de la asignatura y en el perfil del docente a cargo, obteniéndose una recomendación objetiva y exacta.

B. *Objetivos Específicos*

- 1) Recolectar y analizar documentos científicos sobre RS en educación como parte del estado del arte.
- 2) Recolectar y estructurar la data de estudiantes, docentes y asignaturas.
- 3) Evaluar diferentes algoritmos de recomendación para obtener los mejores resultados.
- 4) Diseñar la arquitectura del RS utilizando los mejores algoritmos.
- 5) Implementar la arquitectura del RS.
- 6) Validar la arquitectura del RS.

C. *Resumen*

Con el desarrollo de esta investigación se profundizará en el estudio de los principales problemas en los sistemas de recomendación basados en el desempeño de los estudiantes, tecnología que actualmente forma parte de Educational Data Mining (EDM) [12][13][14] y que genera interés en la comunidad científica que estudia este campo que involucra a la educación y a la minería de datos. Esto llevará al análisis, diseño e implementación de una arquitectura de recomendación que permitirá a los estudiantes saber en qué asignaturas deberán matricularse y evitar la posibilidad de desaprobar o deserción.

Es importante su desarrollo porque además será un ejemplo de que los RS pueden ser empleados en diferentes circunstancias y pueden ser útiles para resolver problemas en diversos campos (en este caso en el campo de la educación), ya que los RS han sido empleados principalmente para recomendar productos o servicios (por ejemplo: [3], [16], [17], [15], [18]) en el campo del marketing, turismo y ventas.

Algunas contribuciones que se buscan son: a) RS de asignaturas a matricularse para estudiantes de la EPIS, b) Sistema de predicción de notas de asignaturas basado en clasificación y clustering, c) Arquitectura de RS híbrido web en base a modelos de recomendación basado en contenido y modelos de recomendación de filtrado colaborativo.

Fig. 1 muestra la arquitectura propuesta para alcanzar los objetivos planteados anteriormente. Ello involucra una actividad de preprocesamiento y selección de atributos que repercutan significativamente en el proceso de recomendación. Tras esto, se ha realizado dos procesos independientes: 1) aplicar técnicas de clasificación (basada en reglas y basada en árboles) y 2) aplicar técnicas propias de los Recommender System (modelo colaborativo). La finalidad es comparar ambas recomendaciones e identificar cuál brinda recomendaciones más precisas.

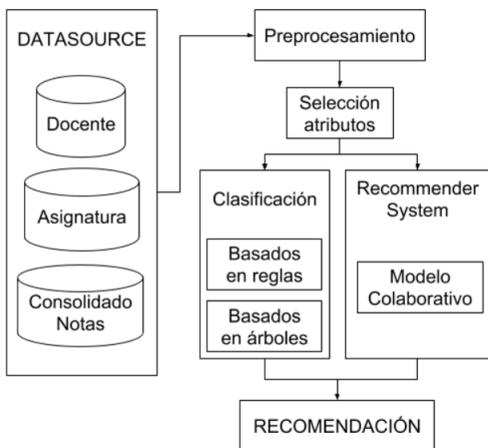


Fig. 1 Arquitectura del Sistema de Recomendación de Matrículas

IV. MODELOS DE PREDICCIÓN

Haciendo uso de la herramienta Weka, se ha evaluado diferentes algoritmos de clasificación agrupados en algoritmos basados en reglas (DecisionTable, JRip, OneR, PART, ZeroR) y basados de árboles (J48, RandomTree).

Inicialmente se ha adecuado el dataset, modificando las clases, generando de este modo 5 datasets distintos:

- 1) Existen 21 clases: desde la nota 0 hasta la nota 20.
- 2) Existen 10 clases agrupados de 2 en 2: (0-2), (3-4), (5-6),..., (17-18), (19-20).
- 3) Existen 4 clases agrupados de 5 en 5: (0-6), (5-10), (11-15), (16-20).
- 4) Existen 4 clases agrupados por letras: AD (17-20), A (14-16), B (11-13), C (0-10).
- 5) Existen 2 clases: aprobados y desaprobados.

Tras el entrenamiento, se evaluó el tiempo que cada algoritmo tarda en generar un modelo de clasificación. En la Fig. 2 se resume los tiempos medidos en segundos; los algoritmos basados en árboles muestran mayor eficiencia ya que ambos no superan los 2 segundos de entrenamiento. Por otro lado, los modelos basados en reglas, tardan algunos segundos más en generar modelos de clasificación, en contraste con el algoritmo PART que tarda 35 segundos.

A pesar, de que los tiempos no alcanzan cercanamente el minuto, es necesario reconocer cuáles son los algoritmos más eficientes, y usarlos cuando el dataset aumente considerablemente, pues será allí cuando la velocidad y

eficiencia pasará a primer plano. Una conclusión adicional, es que mientras menos clases son definidas, menos tiempo tarda el proceso de entrenamiento.

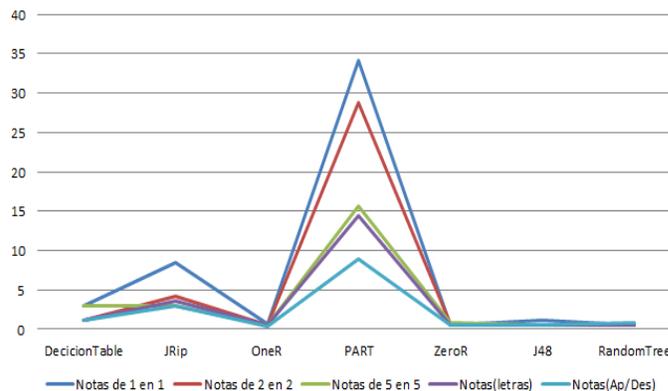


Fig. 2 Tiempo de entrenamiento de cada algoritmo de clasificación (en segundos).

Otra característica evaluada fue la precisión del modelo generado. Fig. 3 muestra un histograma que permite visualizar el porcentaje de precisión utilizando diferentes algoritmos y diferentes dataset. Para el proceso de evaluación se hizo uso de la técnica de cross-validation con 10 folds. Como era de esperarse, los datasets que utilizan mayor cantidad de clases generan modelos menos exactos: al usarse el primer dataset (notas clasificadas de 1 en 1) la precisión es en promedio 22%; usando el segundo dataset (notas agrupadas de 2 en 2), 36%; usando el tercer dataset (notas agrupadas de 5 en 5), 66%; usando el cuarto dataset (notas agrupadas con letras), 57%; y finalmente, el quinto dataset con solo dos clases obtiene una precisión de 77% en promedio.

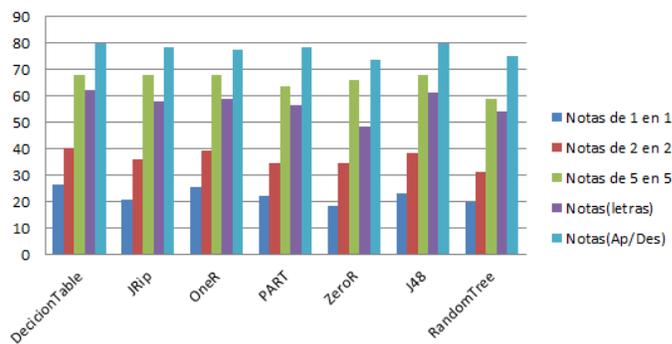


Fig. 3 Precisión del modelo generado por cada algoritmo de clasificación (porcentaje)

En la evaluación entre algoritmos se llega a la conclusión que el orden de los algoritmos más precisos son (en orden creciente): RandomTree, ZeroR, PART, JRip, OneR, J48 y DecisionTable.

Los algoritmos más eficaces y eficientes para el problema planteado son J48 y DecisionTable ya que su porcentaje de precisión más alto es 79 y 79.48 %, respectivamente y el

tiempo promedio de entrenamiento es de 0.6 y 1.2 segundos, respectivamente.

V. RECOMENDACIONES COLABORATIVAS

El dataset inicial tuvo que ser modificado de modo de cada columna pueda representar una asignatura (así, se tiene 51 columnas), y cada fila representa las notas que un estudiante a obtenido en cada asignatura (976 estudiantes).

Para predecir las notas de un estudiante ‘x’, es necesario encontrar a los ‘n’ estudiantes que tengan las notas más parecidas al estudiante ‘x’. Para ello, se hizo uso del algoritmo de distancia Euclidiana, para calcular la similitud de rendimiento académico entre cada uno de ellos.

Tras esto, los estudiantes mejor correlacionados son ordenados de mayor a menor, y se realiza un promedio ponderado para calcular la nota probable que el estudiante obtendría en las asignaturas desaprobadas o que no hayan sido cursadas previamente.

TABLA I
ESTUDIANTES CON RENDIMIENTO MUY SIMILAR AL ESTUDIANTE DE PRUEBA

Identificador de estudiante	Distancia Euclidiana
124	0.02582466401386781
179	0.025419046307257744
570	0.025153660726283947
780	0.02417020164295439
515	0.02282398976922232
157	0.0224443900215006
373	0.022398269792852095
310	0.022358114872744683
390	0.022352434862582673
973	0.0222334541410202

Por ejemplo, si se quiere predecir las notas del alumno con código 379, los estudiantes con rendimiento muy similar a éste serían los estudiantes con código: 124, 179, 570, 780, 515, 157, 373, 310, 390, 973 (Tabla I). Posteriormente, el sistema mostrará la nota probable que se obtendría, a través de una suma ponderada. En el ejemplo, el estudiante todavía no se ha matriculado en las asignaturas de cuarto año: “Diseño y Arquitectura de Software”, “Gestión de Proyectos de Software”, “Pruebas de Software”, “Ingeniería de Requerimientos” y la asignatura electiva “Introducción al Desarrollo de Software de Entretenimiento”; y se ha calculado que obtendría las calificaciones 20, 11, 15, 17, 17; respectivamente (Tabla II).

La principal diferencia entre los algoritmos basados en reglas y basados árboles, con el modelo colaborativo es que los primeros predicen que las notas de las cinco asignaturas del ejemplo sería 13, que es equivalente al promedio general de todas las asignaturas del estudiante del ejemplo. Esto significa que los atributos de las entidades ‘docente’ y ‘asignatura’ no tienen la relevancia en el modelo que deberían tener.

Por otro lado, el modelo colaborativo predice notas estrechamente relacionadas con el rendimiento académico de

los estudiantes basados en comportamientos similares, además que los resultados sí muestran la influencia de docentes y asignaturas.

TABLA II
NOTAS DE SALIDA DEL SISTEMA DE PREDICCIÓN QUE FORMA PARTE DEL SISTEMA DE RECOMENDACIÓN

Nota	Identificador de asignatura	Nombre de la asignatura
20	IV_7_DAS	Diseño y Arquitectura de Software
11	IV_8_GPS	Gestión de Proyectos de Software
15	IV_8_PS	Pruebas de Software
17	IV_8_IR	Ingeniería de Requerimientos
17	IV_8_IDSE	Introducción al Desarrollo de Software de Entretenimiento (E)

VI. CONCLUSIONES

Tras evaluar los algoritmos basados en reglas y basados en árboles los porcentajes de precisión (78% en promedio de los 7 algoritmos analizados) y los tiempos de entrenamiento (15 segundos en promedio) son aceptables. Estos resultados no reflejan la influencia correcta de atributos de estudiantes, docentes y asignaturas. Es decir, sólo reflejan el desempeño del estudiante, ya que cada vez que se busca predecir la nota de un estudiante, el resultado es aproximado al promedio del resto de sus notas, por tanto, los atributos de docente y asignatura que influyen en la predicción como deberían hacerlo.

Por otro lado, las recomendaciones obtenidas mediante el enfoque colaborativo de los sistemas de recomendación, reflejan adecuadamente la relevancia de los atributos de estudiante, docente y de asignatura, por cuanto, las recomendaciones están influenciadas adecuadamente por los perfiles de estudiantes, docentes y asignaturas, y las tasas de precisión alcanzan a lo menos 84%.

REFERENCIAS

- [1] F. Ricci, L. Rokach, and B. Shapira, Recommender Systems Handbook, vol. 54. Second ed., 2015.
- [2] A. S. Lampropoulos and G. A. Tsihrintzis, Machine Learning Paradigms Applications in Recommender Systems. 2015.
- [3] M. D. Dussan Sarria, Sistema de recomendación web basado en la actividad de los usuarios de la Universidad Nacional de Colombia. PhD thesis, Universidad Nacional de Colombia, 2012.
- [4] C. Vialardi, J. Bravo, L. Shafiq, and A. Ortigosa, “Recommendation in Higher Education Using Data Mining Techniques,” Proceedings of the 2nd International Conference on Educational Data Mining, pp. 191-199, 2009.
- [5] D. Brockmann, L. Hufnagel, and T. Geisel, Data Mining and Knowledge Discovery Handbook. 2006.
- [6] L. Rokach and O. Maimon, Data Mining with Decision Trees Theory and Applications. World Scientific, 2nd edition, 2015.
- [7] M. Goga, S. Kuyoro, and N. Goga, “A Recommender for Improving the Student Academic Performance,” Procedia - Social and Behavioral Sciences, vol. 180, no. November 2014, pp. 1481-1488, 2015.
- [8] S. K. Pal, S. Member, S. Mitra, and S. Member, “Multilayer Perceptron, Fuzzy Sets, and Classification,” vol. 3, no. 5, pp. 683-697, 1992.

- [9] J. Mahesh, B. Parag, M. Aniket, and W. Pawan, "Analyzing Students performance Using Academic Analytics," 2016 International Conference on ICT in Business Industry Government (ICTBIG), pp. 1-4, 2016.
- [10] A. Kaklauskas, E. Zavadskas, M. Seniut, V. Stankevic, J. Raistenskis, C. Simkevicius, T. Stankevic, A. Matuliauskaite, L. Bartkiene, L. Zemeckyte, R. Paliskiene, R. Cerkauskiene, and V. Gribniak, "Recommender System to Analyze Student's Academic Performance", *Expert systems with Applications*, vol. 40, no. May, pp. 6150-6165, 2013.
- [11] A. Tekin, "Early prediction of students' grade point averages at graduation: A data mining approach," vol. 14, pp. 207-226, 02 2014.
- [12] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4 PART 1, pp. 1432-1462, 2014.
- [13] R. Jindal and M. D. Borah, "A Survey on Educational Data Mining and Research Trends", *International Journal of Database Management Systems (IJDBMS)*, vol. 5, no. 3, pp. 53-73, 2013.
- [14] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12-27, 2013.
- [15] H.-c. Chen and A. L. P. Chen, "A music recommendation system based on music data grouping and user interests", *CIKM '01 Proceedings of the tenth international conference on Information and knowledge management*, pp. 231-238, 2001.
- [16] E. R. Nuñez Valdez, *Sistemas de Recomendación de Contenidos para Libros Inteligentes*. PhD thesis, Universidad de Oviedo, 2012.
- [17] S. Landa Tinco and F. J. Ticona Pacheco, *Un Modelo de Recomendación de Productos utilizando Redes Sociales implementado en Tecnologías de la Web Semántica*. PhD thesis, Universidad Nacional Mayor De San Marcos, 2015.
- [18] F. M. Hsu, Y. T. Lin, and T. K. Ho, "Design and implementation of an intelligent recommendation system for tourist attractions: The integration of EBM model, Bayesian network and Google Maps", *Expert Systems with Applications*, vol. 39, no. 3, pp. 3257-3264, 2012.