REVISTA + TIC	Junio – Noviembre 2024	Vol. 1, No. 1 ISSN L 3072-9696 pp. 23 – 34

USO Y APLICACIÓN DEL APRENDIZAJE AUTOMÁTICO PARA CREAR MODELOS EN LA TOMA DE DECISIONES EN EL ÁREA DE LA SALUD

USE AND APPLICATION OF MACHINE LEARNING TO CREATE MODELS IN DECISION MAKING IN THE HEALTH AREA.

Denis Cedeño
Universidad de Panamá, Facultad de Informática, Electrónica y Comunicación
denis.cedeno@up.ac.pa https://orcid.org/0000-0002-9640-1284

Recibido: 31-1-2023, Aceptado: 8-5-2023

DOI https://doi.org/10.48204/3072-9696.6355

RESUMEN

La inteligencia artificial (IA) es la disciplina que intenta replicar y desarrollar la inteligencia y sus procesos implícitos a través de computadoras. La IA sintetiza y automatiza tareas que en principio son intelectuales y es, por lo tanto, potencialmente relevante para cualquier ámbito de la actividad intelectual humana, abarca en la actualidad una gran variedad de subcampos, como procesamiento de lenguaje natural (PLN), aprendizaje automático (ML), robótica. Con la IA se están realizando proyectos de análisis de datos trabajando con ML, sin embargo, este tipo de soluciones son muy escasas sobre todo en el área médica de nuestro país. Usar tecnologías innovadoras como la IA aplicada a áreas tan sensitivas como la salud cada día va en aumento. Los nuevos modelos basados en ML en la actualidad están siendo más utilizados, sin embargo, en nuestros países son pocos los estudios relacionados al tema. Por lo tanto sin embargo, en nuestros países son pocos los estudios relacionados al tema, esta investigación tiene como objetivo utilizar diversas técnicas de ML y determinar cómo estos modelos nos pueden ayudar a la solución de problemas en la salud.

PALABRAS CLAVES

cesárea, aprendizaje automático, algoritmos de clasificación, extracción de información.

ABSTRACT

Artificial intelligence (AI) is the discipline that tries to replicate and develop intelligence and its implicit processes through computers. AI synthesizes and automates tasks that are intellectual in principle and is therefore potentially relevant to any field of human intellectual activity, currently encompassing a wide variety of subfields, such as natural language processing (NLP), machine learning (ML), robotics. With AI, data analysis projects are being carried out working with ML, however, these types of solutions are very scarce, especially in

the medical area of our country. Using innovative technologies such as AI applied to areas as sensitive as health is increasing every day. The new models based on ML are currently being used more, however, in our countries there are few studies related to the subject. Therefore, this research aims to use various ML techniques and determine how these models can help us solve health problems.

KEYWORDS

cesarean, machine learning, classification algorithms, information extraction.

INTRODUCCIÓN

La información por sí misma está considerada un bien patrimonial. De esta forma, si una organización tiene una pérdida total o parcial de información esto provoca muchos perjuicios. Es evidente que la información debe ser protegida, pero también explotada. La información asociada a un contexto y a una experiencia como son los procesos médicos se convierte en conocimiento siendo un recurso intangible que aporta verdadero valor a la organización. En Panamá, los sistemas informáticos han ido en constante avance. Una gran parte de la información generada por estos sistemas de información de salud o eHealth es almacenada electrónicamente. Cada día este volumen de información crece. Es tanto el volumen de información de pacientes que actualmente se almacena en diferentes medios o repositorios como las bases de datos, las cuales cada día deben ser más potentes para almacenar dicha información (Cedeno-Moreno & Vargas-lombardo, 2016).

Común es hoy día para un especialista de salud registrar los datos del paciente por vía electrónica, esto incluye no solo la información general del paciente, sino también lo relacionado con el diagnóstico de su enfermedad, los resultados analíticos, pruebas funcionales y la medicación entre otros (Maldonado et al., 2012). Con tantos pacientes en los hospitales del país es enorme también el crecimiento de información digital, lo que hace implementar sistemas innovadores que procesen, analicen y exploten dicha información. Muchas veces el tratamiento de esta información requiere tareas de extracción, filtrado y clasificación (Fadly et al., 2007) tareas propias de la IA. La aplicación de la Tecnología de la Información y la Comunicación (TIC) en temas que afectan la atención de la salud está aumentando ya que proporciona considerables ventajas en términos de información, incluida la obtención de diagnósticos alternativos y un control eficiente de los pacientes (Maier, 2007). En la actualidad se genera un gran volumen de documentos y datos por lo cual es importante contar con herramientas tecnológicas que permitan a las personas obtener, procesar y discernir información convertida en conocimiento y que sea útil para ayudar a la toma de decisiones. La extracción de nuevo conocimiento a partir de grandes volúmenes de texto demanda utilizar novedosos mecanismos en el tratamiento de la información como la aplicación de técnicas de IA (Wyner & Peters, 2010). Con esta enorme cantidad de datos que son generados, se pueden realizar modelos predictivos basados en IA.

Les presentamos este trabajo investigativo que tiene como objetivo utilizar diversas técnicas o algoritmos de ML para obtener un modelo y determinar cómo

este nos ayuda a predecir la forma de parto de una mujer embarazada dependiendo de ciertas características que esta presenta (Sebastiani, 2001).

El resto del documento está estructurado de la siguiente manera: Sección 2 presenta los referentes teóricos. Sección 3 los materiales y métodos. Sección 4 los resultados. Sección 5 se describe la discusión y finalmente en la sección 6 las conclusiones y trabajo futuro.

Referentes Teóricos.

La IA es una rama de las ciencias computacionales encargada de estudiar modelos de cómputo capaces de realizar actividades propias de los seres humanos en base a dos de sus características primordiales: **el razonamiento y la conducta** (Ordoñez-Salinas & Gelbukh, 2010). La IA ha evolucionado hasta el punto de presentar diferentes metodologías de ML aplicadas a innumerables áreas del saber (Akusok et al., 2015). El ML tiene por objetivo desarrollar técnicas que permitan que las computadoras aprendan, de forma más concreta, se trata de crear programas o modelos capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos (McCallum et al., 1999). El término ML se definió originalmente como "**generación artificial de conocimiento a partir de la experiencia**". Fue entonces que se realizaron los primeros estudios y se diseñaron los primeros juegos. Las técnicas de ML están experimentando un auge sin precedentes en varios campos, tanto en el mundo académico, empresarial, salud y demás, constituyendo una herramienta de transformación relevante, es una de las áreas de la IA de más rápido crecimiento y se puede considerar una intersección de la informática y las estadísticas, está estrechamente relacionado con la ciencia de datos y el descubrimiento de conocimiento.

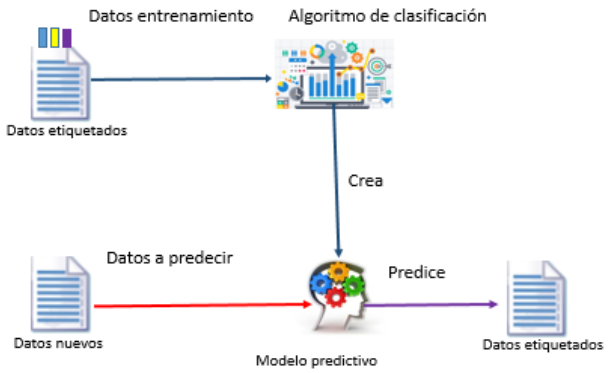
Actualmente diferentes organizaciones, entre ellas en el área de la salud, han optado estas técnicas para evaluar escenarios con pacientes y diferentes enfermedades, para crear modelos que puedan ser implementados en el bienestar del ser humano. En este ámbito es la salud uno de los mayores desafíos. El ML es muy amplio y trata el problema de extraer características de los datos para resolver tareas predictivas, incluido el apoyo a la decisión, el pronóstico, la clasificación (por ejemplo, en el diagnóstico de cáncer), la detección de anomalías (por ejemplo, mutaciones de virus) o el análisis de sentimiento (Pedregosa et al., 2012). Nuestro desafío es descubrir patrones estructurales relevantes o dicho de otra manera conocimiento en los datos, que a menudo están ocultos y no son accesibles para el experto humano. Uno de los mayores problemas es que la mayoría de los conjuntos de datos en el dominio médico están mal estructurado y no están estandarizados. En medicina el ML apenas se ha utilizado, en parte por razones culturales y filosóficas por las que se asume que una computadora nunca será tan capaz como un médico de evaluar una enfermedad mucho menos predecir un diagnóstico. También por el rechazo de algunos médicos a sentirse cuestionados, supervisados o aconsejados por una máquina o por un ingeniero (Verma et al., 2015). En muchos países desarrollados, se realizan proyectos de ciencias biológicas y genómicas que usan ya métodos computacionales avanzados de ML, mientras que los médicos tienen

que lidiar con bases de datos cada vez más grandes y complejas recurriendo a métodos estadísticos tradicionales.

El objetivo principal del ML es el desarrollo de teorías, técnicas y algoritmos que permitan a un sistema modificar su comportamiento a través de la inferencia inductiva. Esta inferencia está basada en la observación de datos que representan información incompleta sobre un proceso o fenómeno estadístico. Para el ML existen 4 formas distintas de aprender o algoritmos: **aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semi supervisado o aprendizaje por refuerzo**. En el caso particular de los 3 primeros tipos de algoritmos se diferencian en el conocimiento a priori que se tiene en cada uno. Los dos extremos son el supervisado, donde se tiene conocimiento a priori de los datos, y el no supervisado, caracterizado por la ausencia de conocimiento a priori.

- **Aprendizaje supervisado**: El objetivo de este tipo de algoritmo es que mediante unos datos de entrenamiento, deducir una función que haga lo mejor posible el mapeo entre unas entradas y una salida (Khairnar & Kinikar, 2013). Dentro de los algoritmos supervisados se pueden encontrar dos corrientes, los algoritmos de clasificación y los de regresión. Los algoritmos de clasificación se usan cuando el resultado deseado es una etiqueta discreta. Aquí tenemos clasificación binaria, solo se elige entre dos etiquetas y clasificación para múltiples etiquetas. La otra corriente son los algoritmos de regresión, los cuales son útiles para predecir valores que son continuos. Eso significa que la respuesta a su pregunta se representa mediante una cantidad que puede determinarse de manera flexible en función de las entradas del modelo en lugar de limitarse a un conjunto de posibles etiquetas. En la figura 1 se puede ver el modelo estándar del aprendizaje supervisado.

Figura 1. Modelo general del aprendizaje supervisado.



- **Aprendizaje no supervisado**: Se organizan los datos de alguna manera que se pueda describir su estructura. El objetivo del aprendizaje no supervisado es lograr crear un modelo de la estructura o distribución de los datos para aprender más sobre ellos. Sirve tanto para entender como para resumir un conjunto de datos. En términos generales, pueden ser agrupados en algoritmos de clustering y algoritmos de asociación (Mcgregor et al., 2004).

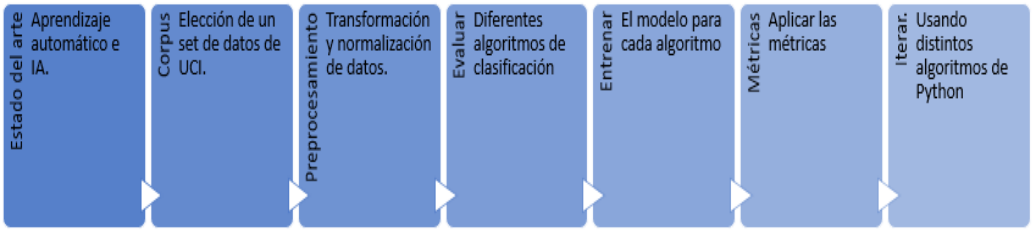
- **Aprendizaje semisupervisado**: El aprendizaje semisupervisado se encuentra a medio camino entre el aprendizaje supervisado y el no supervisado. Ahora lo que tenemos son tantos datos etiquetados como datos no etiquetados, es decir, además de tener tuplas (X,Y), tenemos datos sólo de X de los que no sabemos su respuesta en Y.
- **Aprendizaje por refuerzo**: El objetivo en el aprendizaje por refuerzo es aprender a mapear situaciones de acciones para maximizar una cierta función de recompensa. En estos problemas un agente aprende por prueba y error en un ambiente dinámico e incierto.

Elegir un algoritmo es un paso crítico en el proceso de ML, por lo que es importante que realmente se adapte al caso de uso del problema en cuestión.

MATERIALES Y MÉTODOS

Metodología: La metodología propuesta que se presenta en este artículo ha sido implementada como un caso de estudio en el siguiente orden como lo podemos apreciar en la figura 2.

Figura 2. Metodología implementada.



Corpus o conjunto de datos: *Uno de los retos que se tienen al construir modelos de ML es la obtencion de un corpus lo sufientemente confiable y estable. en nuestro caso el corpus utilizado para el estudio lo hemos obtenido del sitio web de UC Irvine Machine Learning Repository el cual detallo a continuación: (https://archive.ics.uci.edu/ml/datasets/Caesarian+Section+Classification +Dataset). Este conjunto de datos contiene información sobre los resultados de la cesárea de 80 mujeres embarazadas con las características más importantes de los problemas de parto en el campo médico.* El corpus que hemos utilizado es un conjunto de datos para llevar a cabo los experimentos de clasificación, consta de varias variables de predicción médica y la variable objetivo, que es nuestra variable dependiente. En tabla 1, podemos observar la descripción de las variables.

Tabla 1. Descripción de las variables.

Atributo	Características
edad	numérico
partos	numérico
tiempo parto	numérico 0 = oportuno, 1 = prematuro, 2 = tardío
presión	numérico 0 = baja 1 = normal 2 = alta
problema cardiaco	numérico 0 = apto 1 = inepto
cesárea	numérico 0 = No 1 = Sí

Librerías de Python: Decidimos trabajar con la herramienta de programación Python ya que tiene una gran variedad de librerías para el desarrollo de aplicaciones de ML. El lenguaje de programación Python ofrece muchos beneficios para los que desean integrarse en el contexto del ML, ya que sus librerías facilitan las tareas (Ren, 2021). Entre ellas tenemos las que se muestran en la tabla 2.

Tabla 2. Librerías de Python.

Librería	Función
Sklearn	Incluye muchos algoritmos de ML, aquí, estamos usando algunos de sus módulos como train_test_split, RandomForestClassifier y precision_score.
NumPy	Librería con una variedad de módulos numéricos de Python que proporciona funciones matemáticas rápidas para los cálculos. Se utiliza para leer datos y llevarlos a arreglos y para poder manipularlos.
Pandas	Librería para leer y escribir en archivos. La manipulación de datos se puede hacer fácilmente con esta librería.

Preprocesar los datos: Antes de entrenar el modelo, tenemos que dividir el conjunto de datos en conjunto de datos utilizado en entrenamiento y prueba, esta tarea la realizamos utilizando el módulo de **Python** llamado **Scikit-Learn** específicamente el método **train_test_split.** En nuestro caso decidimos usar el método **80/20** es decir entrenamos con el 80% de los datos y el resto para la predicción (Zainuddin & Selamat, 2014). El primer conjunto es utilizado para que el algoritmo "aprenda" de las diversas características de los pacientes y el segundo conjunto sirve para evaluar el rendimiento del modelo obtenido. Dicho modelo, permite clasificar nuevos pacientes.

Entrenamiento de algoritmo: Se usaron técnicas de ML. En este paso se entrenaron los algoritmos de clasificación a través de sus características generados en el pre proceso. La idea es que los algoritmos puedan extraer información útil de los datos que le pasamos para luego poder hacer predicciones con datos nuevos. Aplicamos los algoritmos de **Naive Bayes**, **Máquina de Soporte de Vectores** (SVM), también **Random Forest** y **KNearest Neighbour** (Knn). En la tabla 3 se muestran las características de cada algoritmo usado.

Tabla 3. Características de Algoritmos usados.

Algoritmo	Característica
Naive Bayes	Este clasificador usa el teorema de Bayes, con un supuesto de independencia entre los predictores. En términos simples, un clasificador Bayesiano asume que la presencia de una característica particular en una clase no está relacionada con la presencia de cualquier otra característica (Tanuja et al., 2011).
SVM	Es un algoritmo de aprendizaje automático supervisado que se puede utilizar tanto para tareas de clasificación como de regresión. SVM realiza la clasificación al encontrar el hiper plano que diferencia las clases que trazamos en el espacio n-dimensional (Li & Wu, 2010).

Random Forest	Es un algoritmo que mezcla una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos (Nguyen et al., 2013).
K-nn	La idea básica de este algoritmo es que un nuevo caso que se vaya a clasificar estará en la clase más frecuente a la que pertenecen sus K vecinos más cercanos. El algoritmo es muy usado para tareas de clasificación ya que se fundamenta por tanto en una idea muy simple e intuitiva, además es fácil implementarlo (Trstenjak et al., 2014).

Según la metodología propuesta y los métodos utilizados encontramos buenos resultados que presentamos en el siguiente apartado.

RESULTADOS

En nuestra investigación por las características del conjunto de datos trabajados nos estamos enfrentamos a un problema donde lo mejor que se ajusta o requiere son los algoritmos de clasificación binaria supervisada clásica. Esto lo podemos deducir ya que, dado un número de elementos, todos con ciertas características, queremos construir un modelo de ML para identificar a las mujeres a las que se les practicará un procedimiento de cesárea. Específicamente para las tareas de clasificación usamos los algoritmos: **SVM, Naive Bayes, Random Forest y Knn**, estos algoritmos son de aprendizaje supervisado (Aha et al., 1991).

Pudimos medir el experimento con el módulo de **classification_report** que nos brindó información de las métricas comunes de precisión para cada etiqueta de Si será o No la cesárea con los datos del set de entrenamiento con los distintos algoritmos (Knn, Naive Bayes, Random Forest y SVM) (Sidorov et al., 2013).

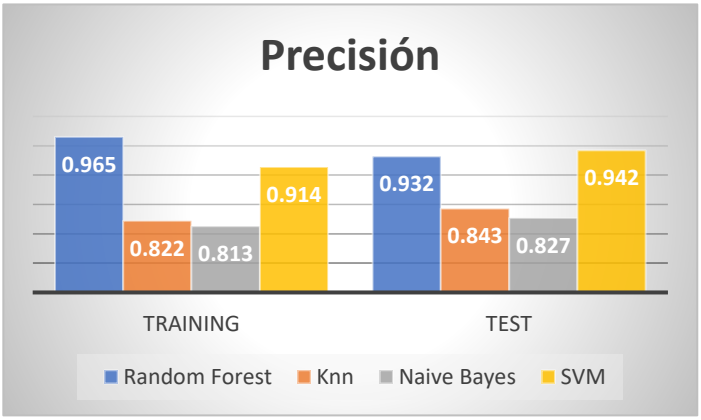
En el experimento obtuvimos como resultados que los algoritmos de Random Forest y SVM presentan mejores resultados que los algoritmos de Naive Bayes y Knn. Esto en gran medida a que siempre ambos algoritmos son considerados mejores para la tarea de clasificación. En tabla 4, podemos observar los resultados de la precisión para datos de entrenamiento y prueba.

Tabla 4. Precisión para Training y Test.

Algoritmo	Training	Test
Random Forest	0.965	0.972
Knn	0.822	0.843
Naive Bayes	0.813	0.827
SVM	0.914	0.932

En figura 3, podemos observar los resultados de la precisión para datos de entrenamiento y prueba.

Figura 3. Reporte de clasificación.



En cuanto a la precisión del conjunto de datos de entrenamiento, los resultados del algoritmo de Random Forest presentan un 96% el mejor porcentaje obtenido de los cuatro. Y para los datos de prueba el mejor porcentaje es el algoritmo de SVM. Es bueno señalar que por encima del 90% se considera un resultado bueno. Esto se debe a que el aprendizaje automático es en realidad un conjunto de muchos métodos diferentes que son especialmente adecuados para responder a diversas preguntas sobre un contexto específico.

DISCUSIÓN

Después de realizar los experimentos sobre el conjunto de datos de mujeres embarazadas en donde aplicamos distintos algoritmos de ML para lograr obtener un modelo predictivo que nos ayudara a la tarea de clasificación del tipo de parto que pudiese tener (parto por cesárea o parto normal) dependiendo de sus características, las cuales fueron expuestas al inicio (edad, cantidad de partos, presión arterial, problemas cardiacos) consideramos que el rendimiento global obtenido es bastante positivo, ya que cada algoritmo mostró métricas superior al 80% en exactitud.

Hemos realizado una experimentación que se ha basado principalmente en el enfoque de ML, donde hemos combinado resultados de diferentes algoritmos de clasificación y también hemos podido utilizar métricas para ver su precisión. Aplicamos varios algoritmos de ML para clasificación, en conjunto los cuatro obtuvieron buenos resultados sin embargo para nuestro experimento nos resultó mejor el algoritmo de Random Forest, lo que marca un precedente para otros trabajos similares.

La precisión aún puede mejorarse, consideramos que haciendo una selección más cuidadosa del corpus y logrando un mejor pre proceso de la data, puesto que en el corpus utilizado puede notarse ciertas ambigüedades como por ejemplo en algunos casos la clasificación no sigue las normas generales. Por lo tanto, este tipo de desafíos se pueden resolver utilizando enfoques innovadores.

CONCLUSIONES Y TRABAJOS FUTUROS

Aunque la región centroamericana está en desarrollo, Panamá cuenta con una importante infraestructura tecnológica y de comunicaciones que permite el desarrollar soluciones innovadoras para la población de forma tal que avancemos y adaptemos nuevos servicios para la salud.

Existe también una gran demanda de servicios de salud pública y el número de pacientes en diferentes niveles de atención hospitalaria está creciendo rápidamente. El uso de tecnologías de salud como salud electrónica y la implementación de la IA puede desempeñar un papel importante en las estrategias de salud del país.

Se ha presentado una investigación que hace experimentos sobre un corpus de datos que estaba etiquetado para determinar si una paciente dará a luz o no por cesárea, aplicando diferentes algoritmos de ML supervisado y realizando tareas de clasificación. Nuestra meta fue analizar y validar a través de métricas cual algoritmo era más eficiente para nuestros propósitos.

Para ello creamos una metodología que fue explicada, utilizada y validada obteniendo buenos resultados. Consideramos que esta investigación fue positiva y eficiente en cuanto a la metodología propuesta y por otro lado se pudo ver las ventajas que ofrecen el aprendizaje automático en cuanto a la clasificación en este caso de una patología clínica.

En cuanto a trabajos futuros, seguir encaminados en esta área de la IA pues nos parece muy interesante y proponer otros estudios con otros conjuntos de datos para ver el comportamiento de los algoritmos de clasificación.

REFERENCIAS BIBLIOGRÁFICAS

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-Based Learning Algorithms. *Machine Learning*, *6*(1), 37–66. https://doi.org/10.1023/A:1022689900470

Akusok, A., Bjork, K.-M., Miche, Y., & Lendasse, A. (2015). High-Performance Extreme Learning Machines: A Complete Toolbox for Big Data Applications. *IEEE Access*, *3*, 1011–1025. https://doi.org/10.1109/ACCESS.2015.2450498

Cedeno-Moreno, D., & Vargas-lombardo, M. (2016). Towards A Model Of Knowledge Extraction Of Text Mining For Palliative Care Patients In. *International Journal of Scientific & Technology Research*, *5*(07).

Fadly, A. El, Daniel, C., Bousquet, C., Dart, T., Lastic, P.-Y., & Degoulet, P. (2007). Electronic Healthcare Record and Clinical Research in Cardiovascular Radiology. HL7 CDA and CDISC ODM Interoperability. *AMIA 2007 Symposium Proceedings*, 216–220. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655824/pdf/amia-0216-s2007.pdf

Khairnar, J., & Kinikar, M. (2013). Machine Learning Algorithms for Opinion Mining and Sentiment Classification. *International Journal of Scientific and Research Publications*, *3*(6), 1–6. www.ijsrp.org

Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, *48*(September), 354–368. https://doi.org/10.1016/j.dss.2009.09.003

Maier, R. (2007). Knowledge Management Systems: Information and Communication Technologies for Knowledge Management. *Knowledge Management*, *2*, 720. https://doi.org/10.1007/978-3-540-71408-8_10

Maldonado, J. A., Costa, C. M., Moner, D., Menárguez-Tortosa, M., Boscá, D., Miñarro Giménez, J. A., Fernández-Breis, J. T., & Robles, M. (2012). Using the ResearchEHR platform to facilitate the practical application of the EHR standards. *Journal of Biomedical Informatics*, *45*(4), 746–762. https://doi.org/10.1016/j.jbi.2011.11.004

McCallum, a., Nigam, K., Rennie, J., & Seymore, K. (1999). Building domain-specific search engines with machine learning techniques. *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, *1999*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.54.4708&rep=rep1&type=pdf

Mcgregor, A., Hall, M., Lorier, P., & Brunskill, J. (2004). *Flow Clustering Using Machine Learning Techniques*. 205–214.

Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering*, *06*(05), 551–560. https://doi.org/10.4236/jbise.2013.65070

Ordoñez-Salinas, S., & Gelbukh, A. (2010). Information retrieval with a simplified conceptual graph-like representation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6437 LNAI*(PART 1), 92–104. https://doi.org/10.1007/978-3-642-16761-4_9

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2

Ren, Y. (2021). Python Machine Learning:Machine Learning and Deep Learning With Python. *International Journal of Knowledge-Based Organizations*,

11(1), 67–70.

Sebastiani, F. (2001). *Machine Learning in Automated Text Categorization*. https://doi.org/10.1145/505282.505283

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A., & Gordon, J. (2013). Empirical study of machine learning based approach for opinion mining in tweets. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7629 LNAI*(PART 1), 1–14. https://doi.org/10.1007/978-3-642-37807-2_1

Tanuja, S., Acharya, D., & Shailesh, K. R. (2011). Comparison of different data mining techniques to predict hospital length of stay. *Journal of Pharmaceutical and Biomedical Sciences*, *07*(07).

Trstenjak, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF based framework for text categorization. *Procedia Engineering*, *69*, 1356–1364. https://doi.org/10.1016/j.proeng.2014.03.129

Verma, V. K., Ranjan, M., & Mishra, P. (2015). *Text Mining and Information Professionals*.

Wyner, A., & Peters, W. (2010). Lexical semantics and expert legal knowledge towards the identification of legal case factors. *Frontiers in Artificial Intelligence and Applications*, *223*, 127–136. https://doi.org/10.3233/978-1-60750-682-9-127

Zainuddin, N., & Selamat, A. (2014). Sentiment analysis using Support Vector Machine. *I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology, Proceedings*, *May 2016*, 333–337. https://doi.org/10.1109/I4CT.2014.6914200