Modelo de clasificación para los anuncios en tres portales de empleo de Colombia según la CIUO-08

Laura Nathalia Leon Rocha

Tutor: Mario José Pacheco López

Universidad Santo Tomás

Facultad de Estadística

Trabajo de grado

Bogotá D.C., Colombia

Junio de 2023

Resumen

La Clasificación Internacional Uniforme de Ocupaciones (CIUO) es una herramienta adoptada por la Conferencia Internacional de Estadísticos del Trabajo, que permite agrupar los diferentes tipos de empleos mediante las actividades y tareas de cada uno de ellos. Este trabajo brinda una visión general de los modelos de Machine Learning para procesos de clasificación y del Procesamiento de Lenguaje Natural (NLP) mediante la implementación de la herramienta de análisis textual Topic Modeling en las descripciones de los anuncios en tres portales de empleo en Colombia, con el objetivo de clasificarlos según la CIUO a un dígito. Se hace uso de los métodos de clasificación Ada Boost, Naive Bayes, Random Forest, Knn, Árboles de decisión y Máquinas de Vectores de Soporte para hallar el que se ajuste mejor a los datos y lograr ordenar de forma adecuada los anuncios. El modelo Random Forest fue el que tuvo mayor acierto en los nueve modelos binarios (uno por cada clase de la CIUO), dado que, para un anuncio hay diferentes profesiones que cumplen con los requisitos del cargo.

Palabras clave: Machine Learning, clasificación, CIUO, Topic Modeling, anuncios.

Tabla de Contenido

Lista de tablas

Lista de figuras

1. Introducción

Mediante el uso de métodos estadísticos de clasificación y su implementación en el lenguaje de programación Rstudio con sus últimas versiones, se desea encontrar un modelo de clasificación óptimo para los anuncios en tres portales de empleo en Colombia según la Clasificación Internacional Uniforme de Ocupaciones (CIUO-08). Para lograrlo se utilizaron las variables con el cargo y la descripción que contiene el anuncio, esto con el fin de realizar un proceso de minería de texto que pertenece al área de Procesamiento de Lenguaje Natural (NLP) llamado Topic Modeling, el cual consiste en asignar una clase a cada objeto observado mediante la frecuencia de aparición de palabras. Según Calvo (2016), el Topic Modeling se utiliza cada vez más en la investigación de textos en español, esto puede ser a causa del estudio mencionado por Minoli (2018), donde las empresas estiman estar usando menos del 1% de los datos no estructurados que almacenan los cuales en su mayoría son texto.

Se usaron los modelos de Machine Learning (Ada Boost, Naive Bayes, Random Forest, Knn, Árboles de decisión y Máquinas de Vectores de Soporte) que permitieron aprender de un conjunto de datos previos y así producir modelos más precisos. Adicionalomo se tiene una base de datos desbalanceada, se hace uso de un sobremuestreo aleatorio de las clases minoritorias con la librería ROSE (Random Over-Sampling Examples). Con este trabajo se desea brindar un soporte para la búsqueda del método de clasificación adecuado para este tipo de información, ya que al contar con una gran cantidad de datos se necesita realizar técnicas de muestreo y validaciones que conlleva un mayor tiempo de entrenamiento de los algoritmos de clasificación.

2. Problema de Investigación

¿Cuáles son las ocupaciones más demandadas en el mercado laboral de Colombia según la Clasificación Internacional Uniforme de Ocupaciones (CIUO-08) en el año 2022?

3. Objetivos

3.1. Objetivo General

Crear un modelo óptimo para los anuncios de empleo en Colombia, que a futuro permita realizar la clasificación de ofertas en el grupo correspondiente según la Clasificación Internacional Uniforme de Ocupaciones (CIUO-08).

3.2. Objetivos Específicos

- Identificar cuál de los modelos de clasificación (Adaboost, Árboles de decisión, Knn, Naive Bayes, Random Forest y Máquina de Vectores de Soporte) es el adecuado para cada clase, mediante la comparación de las métricas de evaluación para modelos en el aprendizaje automático.
- Analizar los patrones que se encuentran en las ofertas de empleo en Colombia, con el propósito de identificar el personal requerido según la CIUO-08 para los cuatro trimestres del año 2022.

4. Justificación

El uso del aprendizaje automático en los últimos años ha venido tomando fuerza dado que es de gran utilidad cuando se tienen muchos datos y se desea trabajar con ellos. Un ejemplo de este, son los modelos de clasificación los cuales necesitan de datos previos para su entrenamiento con el fin de que el modelo aprenda y sea capaz de resolver la tarea dada. Se hace uso de los

diferentes métodos de clasificación para comparar los resultados obtenidos, ya que, dependiendo de la información suministrada en los modelos, se pueden ajustar los parámetros de cada algoritmo para identificar cuál se adapta mejor a los datos. Por lo expuesto anteriormente, en este trabajo se quiere encontrar el modelo adecuado que clasifique los anuncios de empleo con la clase de la CIUO-08 a 1 dígito que le corresponde, con el propósito de agilizar el proceso de clasificación para el análisis de la información para cada departamento.

5. Marco Teórico

5.1 Clasificación Internacional Uniforme de Ocupaciones (CIUO-08)

Es una herramienta usada para estructurar los empleos de acuerdo a sus áreas de acción, las actividades que realizan, el nivel educativo requerido y las habilidades necesarias para su desarrollo. La clasificación se da mediante 4 dígitos:

- La clasificación de un dígito hace referencia a 10 grandes grupos por los que se comienza a desagregar las ocupaciones.
- La clasificación de dos dígitos hace referencia a 43 subgrupos que se establecen según la primera desagregación.
- La clasificación de tres dígitos hace referencia a 136 subgrupos que permiten denotar los diferentes campos de acción existente en el mundo de la oferta laboral.
- La clasificación de cuatro dígitos hace referencia a 455 grupos donde se tiene con más detalle el tipo de empleos existentes en el mercado y a los cuales se puede optar.

En este trabajo se quiere examinar la clasificación a 1 dígito, la cual se desagrega de la siguiente manera:

0) Fuerzas militares

1) Directores y gerentes
2) Profesionales, científicos e intelectuales
3) Técnicos y profesionales de nivel medio
4) Personal de apoyo administrativo
5) Trabajadores de los servicios y vendedores de comercios y mercados
6) Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros
7) Oficiales, operarios, artesanos y oficios relacionados
8) Operadores de instalaciones y máquinas y ensambladores
9) Ocupaciones elementales

Para este trabajo no se tiene en cuenta el grupo 0 (Fuerzas militares), dado que en los portales de empleo observados no se encuentran Anuncios relacionadas a este campo.

5.2 Machine Learning

El Machine Learning o aprendizaje automático se centra en el desarrollo de algoritmos y modelos que permitan hacer predicciones y encontrar patrones con el fin de tomar decisiones, esto mediante un proceso donde las máquinas computacionales aprenden automáticamente de una gran cantidad de datos para mejorar su rendimiento.

Su origen se remonta al año 1950 donde el matemático Alan Turing crea el “Test de Turing”, el cual consistía en que una computadora convenciera a un humano de que estaba en frente de otro humano y no de una máquina. Al igual que con Arthur Samuel quién creo el primer algoritmo cuya función era jugar a las damas, el cuál iba mejorando a medida que se generaban más partidas jugadas.

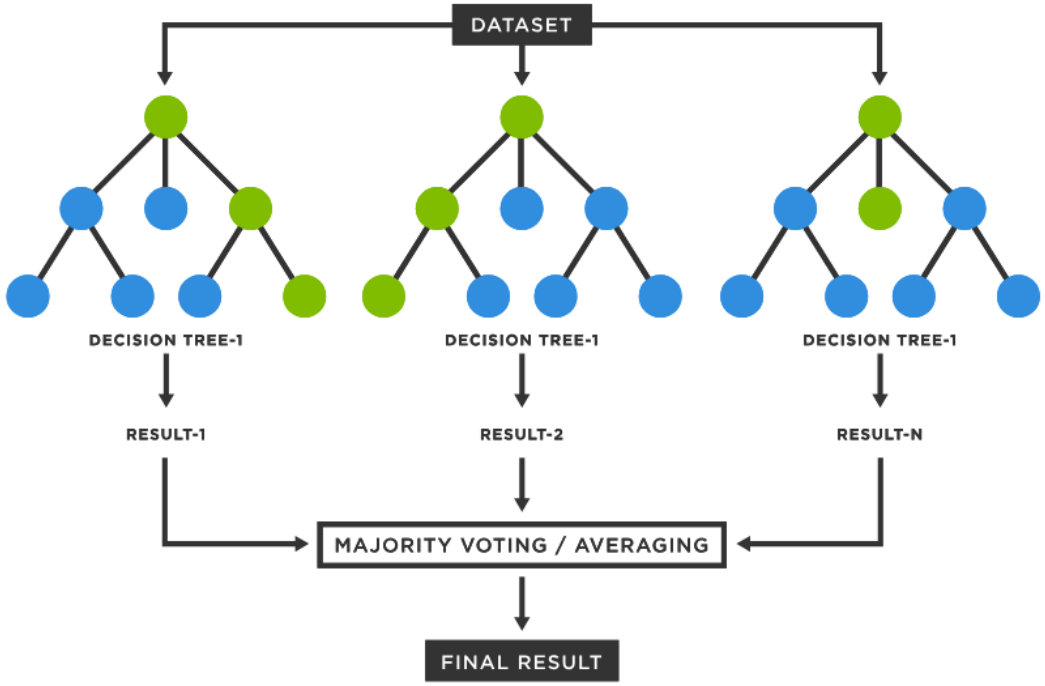
El Machine Learning se divide en 3 grupos:

- El aprendizaje supervisado donde los algoritmos trabajan con datos “etiquetados” (label data), intentado encontrar una función que, dadas las variables de entrada (input data), les asigne la etiqueta de salida adecuada. El algoritmo se entrena con un “histórico” de datos y así “aprende” a asignar la etiqueta de salida adecuada a un nuevo valor, es decir, predice el valor de salida. (Simeone, 2018)
- El aprendizaje no supervisado donde los datos no están “etiquetados” y por ello no se tiene una salida determinada, aquí se puede conocer la estructura que tienen los datos y con esto realizar un agrupamiento para el análisis correspondiente.
- Finalmente, el aprendizaje por refuerzo es donde la máquina va aprendiendo mediante prueba y error, con el fin de ir mejorando a medida que se penaliza el camino incorrecto según el análisis de los datos.

En este trabajo se hará uso del aprendizaje supervisado, en el cual se pueden encontrar modelos de regresión y clasificación. Los algoritmos de clasificación usados para la elaboración de este trabajo son:

5.2.1 Random Forest: es una combinación de modelos predictivos basados en árboles de decisión que se ajustan de forma independiente a diferentes subconjuntos aleatorios del conjunto de datos y luego promedian las predicciones. Es una mejora en relación con el método de un solo árbol debido a que reduce el sobreajuste y proporciona una mayor precisión de las predicciones. (Breiman, 2001)

Figura 1. Clasificación Random Forest.



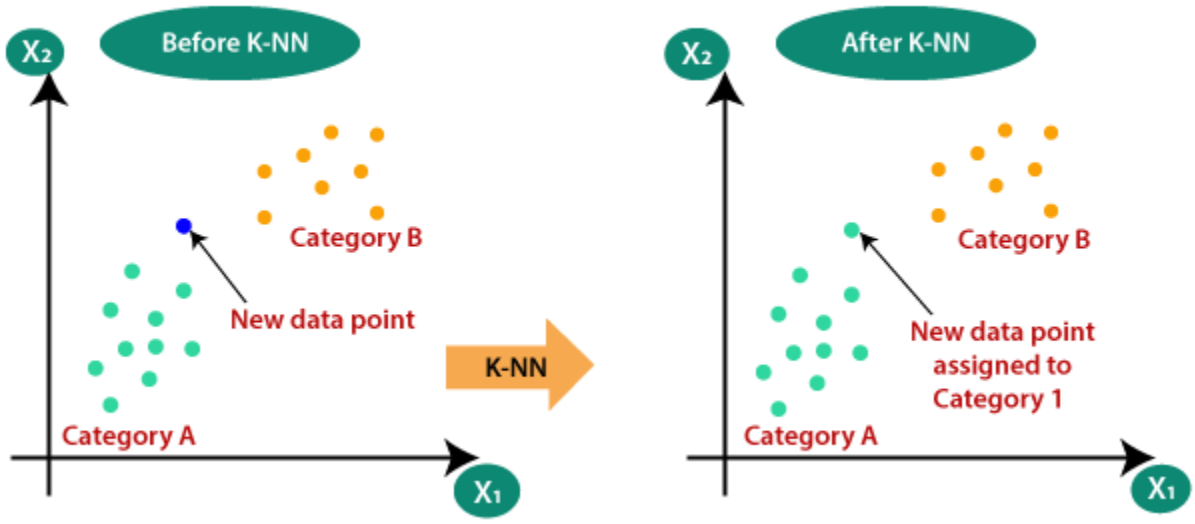
Tomado de TIBCO. https://www.tibco.com/reference-center/what-is-a-random-forest

Este algoritmo es importante dado que Random Forest tiene la capacidad de manejar datos complejos y en gran cantidad como los que se están usando en este trabajo, así como la capacidad de detectar patrones complejos dado a el método de ensamble que este modelo de clasificación utiliza, el cual combina varios modelos con el fin de mejorar la precisión de las predicciones y mitigando un sobreajuste. Una limitación de este algoritmo es el tiempo de entrenamiento y el uso de recursos necesarios cuando se tiene un conjunto de datos amplio, dado que, la construcción de diferentes árboles y la combinación de resultados de los mismos, hace que se necesite mayor cantidad de memoria y de tiempo de entrenamiento.

5.2.2. KNN (K-Nearest Neighbors): el clasificador vecino más cercano asigna una instancia a la clase más cercana fuertemente representada entre sus vecinos. Se basa en la idea de que cuanto más similares sean las instancias, más probable es que pertenezcan a la misma clase.

Podemos usar el mismo enfoque para la clasificación siempre que tenemos una medida de similitud o distancia razonable (Chen et al. 2009).

Figura 2. Clasificación KNN

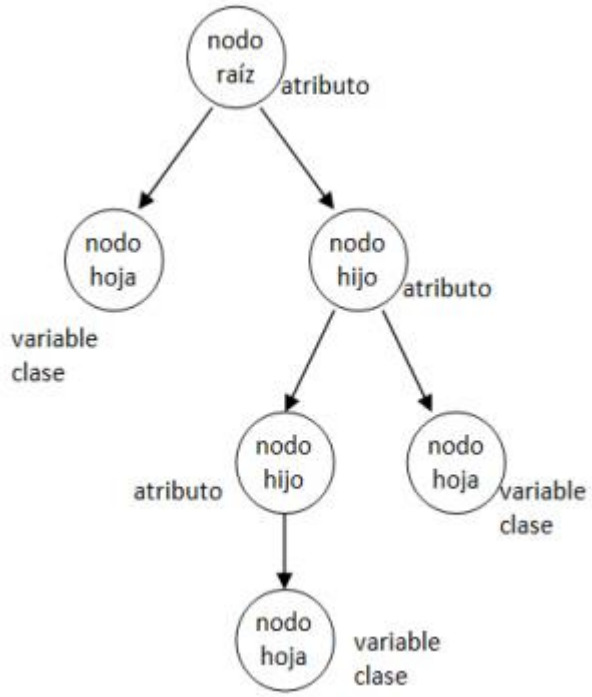


Tomado de java T point. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

El modelo KNN es sencillo de implementar y de entender los resultados al final. Este algoritmo de clasificación se adapta de manera adecuada al momento de incluir datos nuevos y se ajusta muy bien a los cambios. Asimismo, al usar el método de los vecinos más cercanos se puede generar un análisis de recomendaciones sobre los anuncios similares según el interés de las personas. Las limitantes de este algoritmo son la sensibilidad ante la escala y la dimensionalidad, así como la elección de un valor k óptimo, donde no haya un sesgo ni una sensibilidad ante valores atípicos. Adicionalmente, un desequilibrio en la cantidad de clases en los datos de entrenamiento, puede llegar a sesgar hacia las clases con mayor cantidad de ejemplos en la base de datos usada.

5.2.3. Árboles de Decisión: es un modelo de predicción cuyo objetivo principal es el aprendizaje inductivo a partir de observaciones y construcciones lógicas. Son muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva para la solución de un problema. (Martínez et al., 2009)

Figura 3. Clasificación Árboles de Decisión.

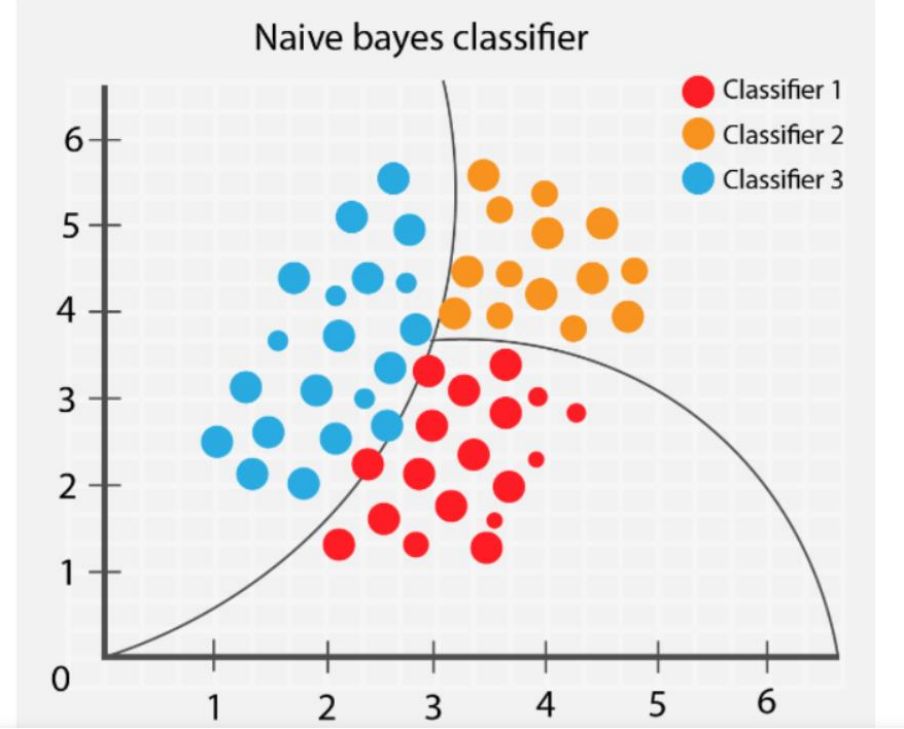


Tomado de (Martínez et al., 2009)

Este algoritmo es de suma importancia para este tipo de clasificación, dado que, permite definir múltiples criterios para realizar la clasificación y a medida que se vaya adquiriendo nueva información se puede ajustar de manera rápida el modelo en aras de mejora. Algunas limitaciones de este modelo son la tendencia al sobreajuste en los datos de prueba y es sensible a las variables que no son significativas para el modelo.

5.2.4. Naive Bayes: Según Mitchell (1997), el teorema de Bayes proporciona una forma de calcular la probabilidad de una hipótesis en base a su probabilidad anterior, las probabilidades de observar varios datos dada la hipótesis, y los propios datos observados. Lo anterior nos da a entender que, en el algoritmo de Naive Bayes se utiliza un modelo probabilístico para conocer las probabilidades de que algunos datos pertenezcan a una clase en específico dadas las características observadas.

Figura 4. Clasificación Naive Bayes.



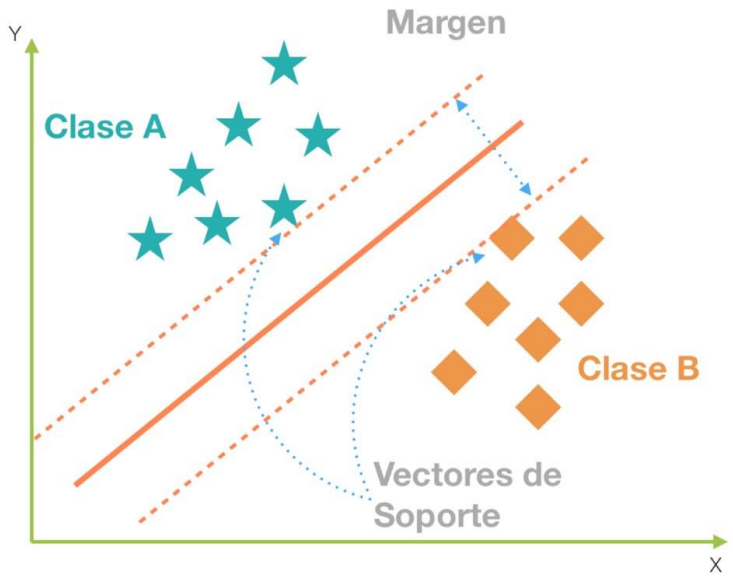
Tomado de Huawei. https://forum.huawei.com/enterprise/es/index

El algoritmo Naive Bayes se usa con frecuencia para la clasificación de texto y de sentimientos, por esto es importante utilizar en este trabajo. Además, computacionalmente es más eficiente y veloz en comparación con otros modelos, es capaz de trabajar con datos textuales y es fácil de interpretar, dado que, arroja una probabilidad de pertenencia en una clase determinada. Algunas limitaciones para este modelo son que, si una característica no tiene un valor en el

conjunto de datos de entrenamiento, se le asignará una probabilidad de cero y esto afecta la calidad de los resultados. Además, es sensible a datos atípicos y a una mala distribución de datos para cada clase en la base de entrenamiento.

5.2.5. Máquinas de Vectores de Soporte (SVM): funciona correlacionando datos a un espacio de características de grandes dimensiones de forma que los puntos de datos se puedan categorizar, incluso si los datos no se puedan separar linealmente de otro modo. Se detecta un separador entre las categorías y los datos se transforman de forma que el separador se puede extraer como un hiperplano. Tras ello, las características de los nuevos datos se pueden utilizar para predecir el grupo al que pertenece el nuevo registro. (IBM, 2019)

Figura 5. Clasificación Máquinas de Vectores de Soporte.



Tomado de Aprende IA. https://aprendeia.com/maquinas-vectores-de-soporte-clasificacion-teoria/

Este algoritmo tiene la capacidad de lograr un rendimiento alto en las tareas de clasificación, dado a su maximización del margen entre las clases encontradas en los datos. Lo

anterior, con el fin de que el hiperplano que separa a las clases esté lo más alejado posible del margen de cada una para una clasificación adecuada de nuevos datos.

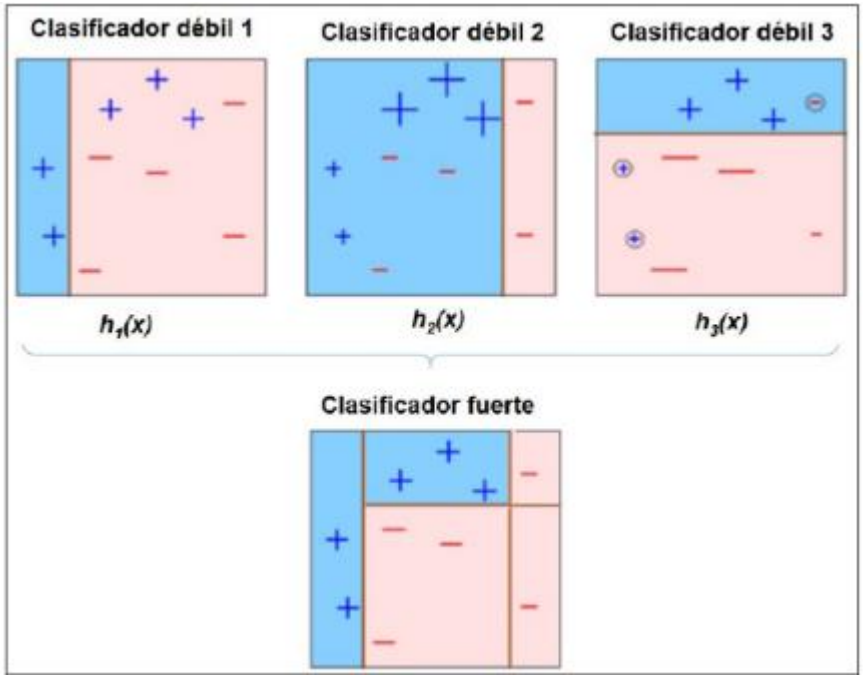
Este modelo también funciona para datos no lineales haciendo una transformación de los datos mediante el uso de la función kernel, la cual admite los siguientes tipos:

- Lineal
- Polinómico
- Función de base radial (RBF)
- Sigmoide

Las limitaciones que presenta este algoritmo son la sensibilidad en la elección de los parámetros como el de regularización y el kernel que depende de los datos de entrenamiento, así como de valores atípicos, al igual que el tamaño de la base dado que puede requerir un gasto computacional considerable si se tiene un conjunto de datos grande.

5.2.6 Adaboost: existen variadas versiones tales como Adaboost, Adaboost.M1, Adaboost.M2, entre otras (Obregón, 2016), cuya metodología consiste en entrenar en forma iterativa una serie de clasificadores débiles tal que cada nuevo clasificador de mayor importancia a los datos mal clasificados en los entrenamientos anteriores, para luego combinar todo el conjunto de clasificadores débiles y obtener un clasificador cuyo rendimiento sea fuerte .Este algoritmo utiliza funciones que ponderan la importancia en relación a cada dato en el proceso del entrenamiento del clasificador, de esta manera, los datos que se han clasificado correctamente pierden peso a favor de los que fueron clasificados erróneamente, intentando conseguir que los nuevos clasificadores se enfoquen en aquellos datos clasificados erróneamente. (Citado por Meza Rodríguez, A. R. en 2020).

Figura 6. Clasificación Adaboost.



Tomado de Uso del algoritmo Adaboost y la regresión logística para la predicción de fuga de clientes en una empresa de telefonía móvil.

El modelo adaboost es de gran importancia debido a su generación de un clasificador basado en errores anteriores, esto genera que tenga una mayor precisión, así como, una capacidad mayor de clasificar de manera adecuada datos nuevos. Las limitaciones de este método de clasificación son la sensibilidad a valores atípicos en los datos y la cantidad de recurso computacional debido a la necesidad de obtener varias iteraciones para dar como resultados un modelo significativo.

5.3 Web Scraping

El Web Scraping, también conocido como extracción web o recolección, es una técnica para extraer datos de la World Wide Web (WWW) y guardarlo en un archivo del sistema o base de datos para su posterior recuperación o análisis. Por lo general, los datos web se obtienen

utilizando el Protocolo de Transferencia de Hipertexto (HTTP) o a través de un navegador web. (Zhao,2017)

Los datos con los que se quiere hacer el análisis y su respectiva clasificación se van a obtener de 3 portales de empleo web en Colombia (El empleo, CompuTrabajo y LinkedIn) mediante el algoritmo de web scraping que nos recopila los datos online y así poder crear la base de datos necesaria para el estudio, mediante el lenguaje de programación R.

5.4 Antecedentes

Guardiola (2019) realizó un trabajo de investigación con el fin de encontrar un clasificador de noticias óptimo a los requerimientos de la empresa que brindó la información. Realizó un proceso de limpieza de datos e implementó los modelos Máquinas Vectoriales de Soporte, Naive Bayes, Árboles de decisión, Regresión logística y K medias. Los resultaron mostraron que el método de Árboles de decisión es el modelo con mayor porcentaje de acierto mientras se va aumentando la cantidad de datos de entrenamiento.

En un estudio más reciente, Alfaro y Allende (2020) realizaron un análisis de sentimiento en el que evaluaron el sentimiento de las personas que usaban la red social twitter durante el pico de COVID-19. Utilizaron un modelo de Representaciones de Codificadores Bidireccionales de Transformers (BERT) y modelos de Machine Learning como Bayes, Regresión logística, Bosque aleatorio y Máquinas de Vectores de Soporte, con el fin de encontrar el mejor. Los resultados mostraron que el modelo con mejor precisión es el BERT con un 84.2%, donde se ve un enfoque positivo y neutral en las personas que tuitean.

6. Hipótesis

El uso de minería de texto en el análisis de las descripciones de los anuncios en los portales de empleo permite identificar patrones y características relevantes para la clasificación efectiva, haciendo uso de metodologías de Machine Learning para la creación de un modelo que permita hacer clasificaciones futuras.

7. Metodología

Para lograr el objetivo de este estudio, se llevará a cabo una serie de actividades importantes con el fin de analizar los datos requeridos y ejecutar los algoritmos de Machine Learning para realizar la clasificación de los anuncios de empleo con la CIUO.

7.1 Datos

La base de aprendizaje para los modelos proviene del Centro de Estudios para la Competitividad Regional (SCORE). Esta se compone de 9 variables con un total de 881.361 Anuncios para el año 2019 con información extraída de los portales de empleo Computrabajo, El Empleo y Linkedin

Tabla 1. Variables base de datos.

Nombre	Descripción
Título_vacante	Descripción breve del tipo de trabajo que la empresa está buscando
nombre_empresa	Empresa que publicó la vacante
Descripcion	Descripción detallada como requisitos, habilidades e información relevante de la vacante
id	Identificador de la vacante
year	Año de publicación de la vacante
sector	Actividad económica a la que pertenece la vacante

educ	Nivel educativo requerido
asal	Remuneración económica por el trabajo
exp	Tiempo de experiencia en el campo requerido

La información utilizada para este trabajo es el título y la descripción de la vacante, las cuales servirán como variables predictoras para la aplicación de los algoritmos de clasificación de Machine Learning.

7.1.1 Método de Clasificación: actualmente, el método usado para la clasificación de los anuncios en SCORE es de forma manual. Un grupo de personas se dedica a leer las descripciones de los anuncios y mediante las definiciones, tareas y ocupaciones de las 9 clases utilizadas de la Clasificación Internacional Uniforme de Ocupaciones a 1 dígito, se asigna una clase a cada vacante. Esto genera un reto operacional, dado que, al tener que clasificar las nuevas Anuncios en los portales de empleo se vuelve desgastante debido a la cantidad de datos. Es por esto que se quiere optimizar y agilizar este proceso, con el fin de ahorrar tiempo y dar mayor eficiencia en el análisis de esta información.

Esta clasificación puede implicar algunos errores y se espera realizar un ajuste de esta mediante la ejecución del algoritmo del modelo.

7.1.2 Datos Análisis: se recopila la información de los anuncios de empleo para el año 2022, publicadas en los portales de empleo ya mencionados obteniendo el cargo, la descripción, el salario, la empresa y el departamento de la oferta.

7.2 Preprocesamiento de texto

Para las bases de datos utilizadas en este trabajo, se crea un corpus de una columna que contiene el cargo y la descripción de la vacante, este sirve de ayuda para manipular de forma

eficiente documentos de texto en R. Luego de tener el corpus se comienza a realizar la limpieza de los datos para facilitar el procesamiento y análisis de texto, se eliminan los artículos, pronombres y preposiciones del texto (stopwords) con el fin de utilizar solamente las palabras clave del texto, también se eliminan números y los símbolos de puntuación como comas, puntos y signos de interrogación o exclamación, etc. Adicional, se convierte el texto a minúsculas para que haya una consistencia en la redacción de las palabras y estas no interfieran por su escritura.

Por último, se realiza un proceso de tokenización el cual consiste en separar por palabras cada descripción y dejándolas como fila con una columna de id para identificar la vacante a la que pertenece, luego se crea una columna utilizando la función ‘wordStem’ con la cual se obtiene la raíz de las palabras, esto con el fin de agruparlas de manera correcta sin importar las diferentes formas de escritura de una misma palabra. Al final se tiene una columna con la descripción conformada de las palabras lematizadas.

7.3 Algoritmos de Clasificación

Al realizar el preprocesamiento de los datos, se obtiene una matriz con 154 palabras predictoras y una columna con la clasificación a un dígito del anuncio, adicional, se eliminan los duplicados y los anuncios que no tienen la clasificación CIUO, consiguiendo una matriz con 531.090 anuncios. Con esta matriz se obtienen 3 muestras y se realiza la ejecución de los modelos con cada una de ellas, con esto se realiza una validación de la eficiencia de los modelos con diferentes datos.

Para la ejecución de los modelos se hace uso del lenguaje de programación R, utilizando las siguientes librerías:

Tabla 2. Librerías para la ejecución de los modelos.

Librería	Función	Parámetros
e1071	svm	kernel: linear, polynomial, radial y sigmoid
	naiveBayes	Predeterminados
ada	ada	type: discrete, gentle y real
kknn	train.kknn	kmax=8
rpart	rpart	Predeterminados
randomForest	randomForest	Predetermiandos

Luego de ejecutar los algoritmos, se obtienen las predicciones de los modelos y se genera la matriz de confusión para evaluar el rendimiento de cada uno. Con estos resultados se calculan las métricas de evaluación y así se puede definir cuál es el modelo adecuado para cada clase.

Tabla 3. Tasa de aciertos (Accuracy).

Modelos	Clase 1*	Clase 2	Clase 3	Clase 4	Clase 5	Clase 6*	Clase 7	Clase 8	Clase 9*
Ada_discrete	100,0%	89,9%	90,3%	87,2%	89,0%	100,0%	91,9%	98,7%	100,0%
Ada_gentle	100,0%	88,2%	88,3%	84,2%	86,5%	100,0%	90,7%	99,0%	100,0%
Ada_real	100,0%	90,4%	90,5%	88,2%	89,5%	100,0%	92,0%	97,7%	100,0%
Árboles de decisión	96,5%	85,4%	87,2%	82,5%	84,6%	92,3%	89,8%	96,2%	94,3%
KNN	100,0%	94,7%	92,2%	93,0%	93,2%	100,0%	94,5%	97,1%	100,0%
Naive Bayes	94,4%	71,0%	64,2%	73,4%	75,8%	86,8%	67,3%	67,8%	93,1%
Random Forest	100,0%	99,3%	99,6%	99,5%	99,2%	100,0%	98,9%	99,6%	100,0%
SVM_linear	80,9%	88,3%	86,6%	84,6%	86,3%	73,3%	89,8%	96,1%	78,5%
SVM_polinomial	98,6%	91,9%	93,8%	92,2%	92,5%	94,5%	93,8%	97,5%	96,0%
SVM_radial	96,8%	93,6%	91,4%	91,9%	93,3%	95,8%	90,8%	96,4%	95,5%
SVM_sigmoid	74,7%	85,4%	86,0%	79,1%	81,8%	63,8%	89,2%	96,0%	72,6%

** Clases con muestras balanceadas realizando un sobremuestreo con ROSE*

Al sacar la tasa de aciertos promedio de las 3 muestras tomadas, el que cuenta con mayor porcentaje acierto es el modelo Random Forest. Para los modelos de las clases 1, 6 y 9 se realiza un sobremuestreo para equilibrar las clases y que no haya un sesgo en los resultados del modelo, dado que, hay poca información de estas clases y las predicciones no estaban clasificando a ningún anuncio. Esto se hace mediante la librería ROSE que permite tener un conjunto de datos

equilibrado, por esta razón, en el cuadro anterior se puede observar una mejor clasificación en los modelos Ada Boost, Knn y Random Forest al compararlos con los demás, pero ese balanceo puede estar generando un sobreajuste (overfitting) en los modelos de estas clases.

7.4 Aplicación del Modelo y Análisis a Nuevos Datos

Para el año 2022, se tiene la información de los anuncios por mes (no se tiene diciembre) y se agrupan por trimestres para realizar el análisis de la aplicación del modelo elegido. Teniendo en cuenta la matriz de palabras utilizada para la calibración de los modelos, se crea otra matriz donde se evidencie la presencia de estas palabras en la descripción de los anuncios nuevos y así obtener la misma estructura para la ejecución del algoritmo del modelo.

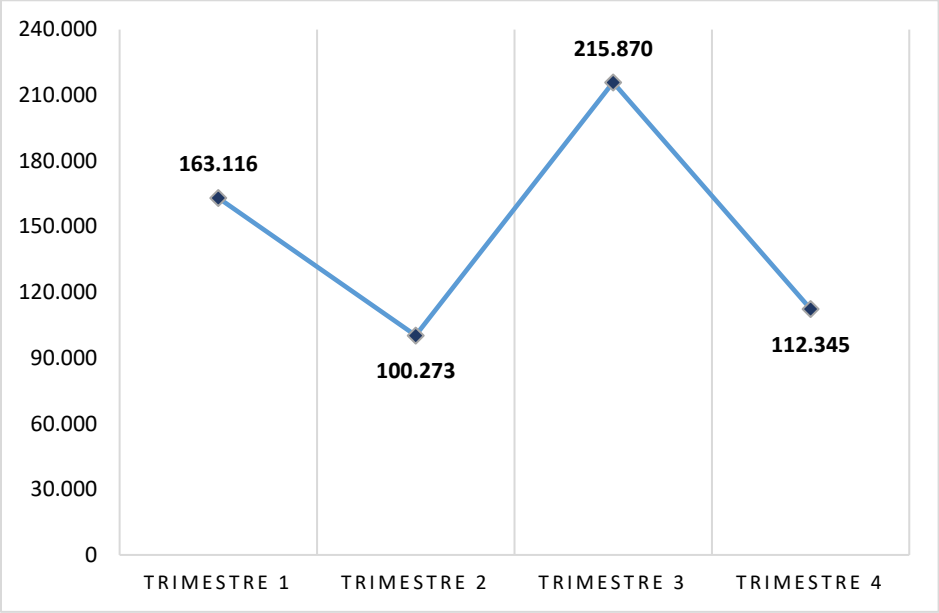
Con las predicciones para el año 2022, se realiza un análisis de la información obtenida según la clasificación de los anuncios con la CIUO. Con esto se extraen estadísticas descriptivas para cada clase y se puede observar con mayor detenimiento la eficiencia y ajuste del modelo a los datos.

8. Resultados

Como se mencionó anteriormente, al momento de observar la métrica de evaluación *accuracy* se pudo identificar que el modelo óptimo para las nueve clases de la CIUO es el Random Forest. Por lo cual, es el modelo con el que se trabaja para realizar el análisis por trimestre del año 2022.

El gráfico 1 muestra el número de anuncios que se tienen por trimestre.

Figura 7. Anuncios por trimestre.



Elaboración propia mediante Excel

En el tercer trimestre del año que lo conforma el mes de julio, agosto y septiembre, es donde se tiene mayor número de anuncios en los portales de empleo.

Por otro lado, la Tabla 4 nos indica cuántos anuncios cuentan con una asignación a una clase de la CIUO y en cuántos no se logró identificar la clase a la que puede pertenecer el anuncio.

Tabla 4. Porcentaje de clasificación.

	T1	**T2**	**T3**	**T4**
Si	99,29%	99,75%	99,78%	99,76%
No	0,71%	0,25%	0,22%	0,24%
Total	163116	100273	215870	112345

El trimestre en el que se tiene mayor número de anuncios sin clasificar es el primero con 1.166 y el que tiene mayor número de anuncios clasificados según la cantidad de ellos por trimestre, es el tercero con 215.401.

Dado que se realizó un modelo para cada una de las clases, un anuncio puede quedar clasificado más de una vez. Para la Tabla 5, se tiene la cantidad de anuncios de los 4 trimestres que se clasificaron en cada una de las 9 clases observadas de la CIUO, obteniendo el porcentaje del total de anuncios que pertenece a dicha clase por trimestre. En el total se suma la cantidad de clases a la que pertenece un mismo anuncio, es decir, si se clasificó en 2 clases, en el total de anuncios este aparece 2 veces.

Tabla 5. Porcentaje de ocupaciones por trimestre.

Clase	**Ocupaciones**	**T1**	**T2**	**T3**	**T4**
1*	Directores Y Gerentes	25,87%	24,52%	21,74%	21,68%
2	Profesionales, Científicos E Intelectuales	8,16%	5,75%	5,19%	4,95%
3	Técnicos Y Profesionales De Nivel Medio	0,29%	0,18%	0,16%	0,14%
4	Personal De Apoyo Administrativo	1,28%	0,95%	0,90%	0,93%
5	Trabajadores De Los Servicios Y Vendedores De Comercios Y Mercados	4,36%	3,05%	5,03%	5,00%
6*	Agricultores Y Trabajadores Calificados Agropecuarios, Forestales Y Pesqueros	26,17%	25,69%	26,95%	27,02%
7	Oficiales, Operarios, Artesanos Y Oficios Relacionados	0,00%	0,00%	0,01%	0,00%
8	Operadores De Instalaciones Y Máquinas Y Ensambladores	20,83%	22,27%	22,32%	22,32%
9*	Ocupaciones Elementales	13,04%	17,60%	17,71%	17,97%

**Clases con muestras balanceadas realizando un sobremuestreo con ROSE*

Se puede observar que para las clases a las cuales se les realizó un sobremuestreo son las que tienen un mayor porcentaje de anuncios clasificados por trimestre, excepto la clase de Operadores de instalaciones y máquinas y ensambladores.

Como criterio de validación se realizan las nubes con los cargos para verificar si son acorde a la clase que se está observando

Clase 1 Clase 2 Clase 3

Figura 8. Nube de palabras clase 1, 2 y 3.

Clase 4 Clase 5 Clase 6

Figura 9. Nube de palabras clase 4, 5 y 6.

Clase 7 Clase 8 Clase 9

Figura 20. Nube de palabras clase 7, 8 y 9.

9. Conclusiones y Recomendaciones

- El modelo escogido para el análisis con la mejor tasa de acierto para la clasificación en todas las clases y para tener una consistencia en los resultados de las predicciones, fue el Random Forest.
- Al realizar un equilibrio en los datos para la clase 1, 6 y 9 se evidenció un posible sobreajuste en los modelos Ada, KNN y Random Forest. Lo que indicaría un error en la clasificación de estas clases en la base de datos, por esta razón, se sugiere una revisión de la base de entrenamiento para verificar si el modelo si se sobre ajusta a los datos.
- Para obtener una buena clasificación de los anuncios, no es suficiente el uso solamente de la descripción, se podrían utilizar las variables como salario, nivel educativo y ubicación. para brindar más criterios de clasificación y obtener mejores resultados.
- Para agilizar el proceso de validación y predicción de los modelos, se recomienda tener en cuenta los resultados de este trabajo para descartar los modelos con una tasa de acierto baja y ejecutar los algoritmos de los modelos que sean significativos, dado que, el tiempo de entrenamiento de los modelos es muy alto.
- A partir de la generación de las nubes de palabras, se puede evidenciar que el modelo tiene aciertos al observar los cargos encontrados en las 9 clases observadas de la clasificación CIUO, sin embargo, para la clase 6 (Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros) si es necesaria una calibración de la base para ajustar los datos de entrenamiento y evitar un agrupamiento erróneo de anuncios para esta clase.
- Los resultados de este trabajo ayudaron a identificar una mala clasificación de los anuncios.

9. Bibliografía

Alpaydin, E. (2014). *Introduction to Machine Learning*. Massachusetts Institute of Technology. https://dl.matlabyar.com/siavash/ML/Book/Ethem%20Alpaydin-Introduction%20to%20Machine%20Learning-The%20MIT%20Press%20(2014).pdf

Aprende IA. (s.f). *Máquinas Vectores de Soporte Clasificación Teoría.* https://aprendeia.com/maquinas-vectores-de-soporte-clasificacion-teoria/

Arquez, M. (30 de marzo de 2020). *Random Forest*. https://rpubs.com/arquez9512/592295

Barrientos et al. (2009). *Árboles de decisión como herramienta en el diagnóstico médico*. http://www.soporte.uv.mx/rm/num_anteriores/revmedica_vol9_num2/articulos/arboles.pdf

Breiman, L. (2001, Octubre). *Random Forests*. https://doi.org/10.1023/A:1010933404324

Burgues, C. (1998). *A Tutorial on Support Vector Machines for Pattern Recognition*. https://doi.org/10.1023/A:1009715923555

Calvo, J. (1 de diciembre de 2016). *Topic Modeling: ¿Qué, cómo, cuándo?*. http://www.morethanbooks.eu/topic-modeling-introduccion/

Carmona, E. (2016). *Tutorial sobre Máquinas de Vectores Soporte (SVM)*. https://www.researchgate.net/publication/263817587_Tutorial_sobre_Maquinas_de_Vectores_Soporte_SVM

Darad, S. y Krishnan, S. (2023). *Análisis de sentimiento de los datos de twitter de COVID-19 utilizando modelos de aprendizaje profundo y aprendizaje máquina.* https://doi.org/10.17163/ings.n29.2023.10

Guardiola González, C. (2020). *Clasificador de texto mediante técnicas de aprendizaje automático.* [Trabajo Fin de Grado, Escola Tècnica Superior d'Enginyeria Informàtica Universitat Politècnica de València]. https://riunet.upv.es/bitstream/handle/10251/133840/Guardiola%20-%20Clasificador%20de%20textos%20mediante%20t%C3%A9cnicas%20de%20aprendizaje%20autom%C3%A1tico.pdf?sequence=1

Hastie et al. (2008, Agosto). *The Elements of Statistical Learning*. Springer. https://hastie.su.domains/Papers/ESLII.pdf

IBM. (17 de agosto de 2021). *Funcionamiento de SVM*. https://www.ibm.com/docs/es/spss-modeler/saas?topic=models-how-svm-works

Java T point. (s.f.). *K-Nearest Neighbor(KNN) Algorithm for Machine Learning*. *https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning*

Martínez Ribón, J. G. T. (2011). *Propuesta de metodología para la implementación de la filosofía Lean (construcción esbelta) en proyectos de construcción* [Tesis de Maestría, Universidad Nacional de Colombia]. http://bdigital.unal.edu.co/10578/

Meza, A. y Chue, J. (2020). *Uso del algoritmo Adaboost y la regresión logística para la predicción de fuga de clientes en una empresa de telefonía móvil.* Natura@economía. https://doi.org/10.21704/ne.v5i2.1610

Minoli, M. (5 de julio de 2018). *Análisis de datos no estructurados usando Topic Models.* https://www.linkedin.com/pulse/an%C3%A1lisis-de-datos-estructurados-usando-topic-models-mariano-minoli/?originalSubdomain=es

Mitchell, T. (1997). *Machine Learning*. Book News. https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf

Nalda, V, (29 de septiembre de 2020). *Machine Learning: Los orígenes y la evolución*. https://www.futurespace.es/machine-learning-los-origenes-y-la-evolucion/#:~:text=Los%20or%C3%ADgenes%20del%20Machine%20Learning&text=Por%20moderno%20que%20pueda%20parecer,en%20vez%20de%20un%20ordenador

Organización Internacional del Trabajo. (30 de enero de 2005). *Estructura de la CIUO-08 y concordancias previas con la CIUO-88.* https://www.ilo.org/public/spanish/bureau/stat/isco/isco08/index.htm

.Recuero, P. (2021, Diciembre). *Tipos de aprendizaje en Machine Learning: supervisado y no supervisado*. https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/

Simeone, O. (2018, Noviembre). *A Very Brief Introduction to Machine Learning With Applications to Communication Systems*. https://arxiv.org/

TIBCO. (s.f.). What is a Random Forest?. https://www.tibco.com/reference-center/what-is-a-random-forest

Tovar, J. (11 de mayo de 2022). *Método Supervisado – Clasificación – Naive Bayes*. https://forum.huawei.com/enterprise/es/m%25C3%25A9todo-supervisado-clasificaci%25C3%25B3n-naive-bayes/thread/667228966991314944-667212895009779712

Zelada, C. (10 de mayo de 2017). *Evaluación de modelos de clasificación*. https://rpubs.com/chzelada/275494

Zhao, B. (2017). *Web Scraping*. DOI:10.1007/978-3-319-32001-4_483-1. https://www.researchgate.net/publication/317177787_Web_Scraping