Un estudio comparativo entre algoritmos de aprendizaje automático orientados a la clasificación de personalidad para selección de personal en un contexto hispano.

Adriana Mansilla Hermosillo*, Fausto Abraham Jacques García
Facultad de Informática, Universidad Autónoma de Querétaro,
Av. de las Ciencias S/N, Juriquilla, 76230 Santiago de Querétaro, Qro.
amansilla13@alumnos.uaq.mx, jacques@uaq.edu.mx

Recibido: *29 de agosto de 2018*
Autorizado: *5 de Noviembre de 2018*

RESUMEN

Hoy en día el internet contiene una cantidad impresionante de material útil para explotar con la minería de personalidad, sin embargo, muchas compañías no aprovechan la información de los candidatos en línea y utilizan pruebas psicométricas tradicionales, que, si bien pueden determinar ciertos rasgos de la personalidad de los candidatos, los resultados de estos tienden a estar sesgados por la impresión que los candidatos quieren dar al empleador. El objetivo de esta investigación consiste en resaltar y comparar tres clasificadores hechos con aprendizaje automático que pueden ser utilizados para predecir la personalidad de un candidato a una posición de trabajo basándose en el contenido escrito que se encuentra en sus redes sociales. Para evaluar el desempeño de los clasificadores se emplearon una serie de pruebas con las métricas más importantes como exactitud, precisión, y exhaustividad. Como resultado se obtuvo un mayor grado de exactitud con el clasificador basado en el algoritmo de bosque aleatorio, ya que, este logro un 69.0% de exactitud promedio además de obtener las mejores métricas.

Palabras clave: Aprendizaje Automático, Clasificación automática de personalidad, Indicador Myers-Briggs, Clasificación de texto en español.

ABSTRACT

Nowadays the Internet has an astonishing amount of useful material for personality mining, nevertheless many companies fail to exploit the information and screen job candidates using personality tests that fail to grasp the very information they are trying to gather. The aim of this paper is to highlight and compare the different machine learning classifiers that can be used to predict the personality of a Spanish speaking job applicant based on the written content posted on their social networks. To evaluate the performance of the classifiers a test harness with the most important performance measures such as accuracy, precision and recall was made. Our results show that the random forest classifier outperforms other classifiers in accuracy and most performance metrics.

Keywords: Machine learning, Automatic personality classification, Myers-Briggs, Spanish text classification.

1 INTRODUCCIÓN

La personalidad es una característica crítica en la vida de una persona, ya que explica mucha parte de nuestro comportamiento y de como interactuamos con el mundo exterior. Hasta la fecha se han creado muchos modelos que buscan definir los diferentes tipos de personalidades existentes, entre los más famosos se encuentra el modelo de Myers-Briggs (The Myers & Briggs Foundation, 2018) que está basado en los estudios de Carl Jung y define cuatro ejes principales de la personalidad: Fuente de energía, Modo de recabar información, Toma de decisiones, y Relación con el mundo externo, de acuerdo a esta teoría una persona puede ser: Introvertida o Extrovertida, Intuitiva o Sensorial, Pensadora o Emocional, y Calificadora o Perceptiva, tal que las cuatro dicotomías se combinan para formar 16 tipos de personalidad finales.

De manera tradicional uno puede determinar su personalidad tomando el test Myers-Briggs que consiste generalmente de 30 preguntas que se contestan con una escala que va desde el 1 al 5. El 1 siendo un completo desacuerdo y el 5 un completo acuerdo con la afirmación dada. Sin embargo, en México, el tiempo promedio de conexión al internet es de 8 horas y un minuto (Velasco, 2017), por lo tanto, hay una gran cantidad de información útil para predecir la personalidad de los usuarios utilizando el contenido que publican en línea, sin necesidad de tomar el test tradicional. Esto es posible ya que estudios anteriores han encontrado una fuerte relación entre el lenguaje escrito que utiliza una persona y su personalidad (Oberlander, 2006) (Gill, 2009).

Estos estudios generalmente utilizan una mezcla de técnicas de análisis que incluyen el "Linguistic Inquiry and Word Count" (LIWC) (Pennebaker, 2007) que permite determinar el grado en que autores/hablantes usan palabras que connotan emociones positivas o negativas, auto-referencias, palabras extensas o palabras que se refieren a sexo, comer o religión; Además de varios clasificadores de aprendizaje automático los cuales funcionan con mayor o menor exactitud dependiendo del tamaño de la base de datos, del preprocesamiento de la base de datos y de los parámetros con los que se configuren. En este trabajo se busca clasificar de manera automática la personalidad de los usuarios analizando el texto en sus redes sociales mediante 3 clasificadores conocidos: Maquinas de Soporte Vectorial, Bosque Aleatorio, y el Bayesiano Ingenuo.

Uno de los propósitos de este articulo es comparar el desempeño de diferentes clasificadores para la misma base de datos, ya que esta se compone de textos escritos por usuarios tenemos un alto número de dimensiones para trabajar, por lo que se eligieron clasificadores sencillos que pueden separar linealmente los datos y llegar a una mejor generalización que otros clasificadores.

El artículo presenta la siguiente estructura: la Sección 2 se presenta el estado del arte en el campo de la clasificación automática de personalidad; la Sección 3 plantea el marco teórico de los algoritmos a comparar además de un breve resumen de la teoría sobre Myers-Briggs y la relación entre la personalidad y el desempeño en el trabajo; en la Sección 4, se describe el desarrollo del experimento, la metodología aplicada para la comparación de los algoritmos, y la información sobre la extracción de datos y su preprocesamiento; en la Sección 5, se muestran los resultados del experimento, y finalmente en la Sección 6 expone la conclusión del trabajo.

2 ESTADO DEL ARTE

En las dos últimas décadas el aprendizaje automático y los sistemas de reclutamiento han florecido y muchos no se concentran en la personalidad de los candidatos, pero la cantidad de información disponible sobre estos es una buena oportunidad para ampliar las fronteras de estos sistemas de reclutamiento, además de que, una herramienta como esta puede ser útil para centros de reclutamiento que no pueden probar a todos los candidatos con una entrevista personal.

En (Faliagka et al., 2012) se creó un sistema de reclutamiento en línea que incluía minería de personalidad automática para los candidatos a una posición de trabajo, ellos medían el grado de introversión/extroversión a través de la polaridad de las palabras que un candidato utilizaba en su blog personal, y después los ordenaban de acuerdo con las necesidades del reclutador. El nivel de extroversión lo calculan usando el modelo desarrollado por (Pennebaker, 2007) llamado LIWC (por sus siglas en inglés Linguistic Inquiry and Word Count, Buscador Lingüístico y Contador de Palabras).

En otra investigación, (Tandera, et al, 2017) se desarrolló un clasificador de personalidad para usuarios de Facebook utilizando técnicas de Deep Learning y el modelo de los cinco grandes y obtuvieron un 74.14% de exactitud.
Por último, un estudio llevado a cabo por (Ortigosa, et al, 2014) tuvo éxito al encontrar patrones mediante interacciones en Facebook de personas con personalidades similares basándose en el modelo de los cinco grandes alternativo.

Estos estudios enfocados a la clasificación automática de la personalidad utilizando redes sociales y algún modelo de personalidad no se desarrollaron como una herramienta orientada al reclutamiento de personal, sin embargo, estas investigaciones previas sugieren un precedente exitoso en donde aun queda mucho por ahondar en especial en un contexto hispano donde el procesamiento de lenguaje es más escaso que en lenguajes que cuentan con más conjuntos de datos como el inglés.

3 TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

3.1 MÁQUINAS DE SOPORTE VECTORIAL

Una máquina de soporte vectorial es un clasificador discriminativo formalmente definido por un hiperplano separador desarrollado por Cortes y Vapnik (Cortes, 1995). En otras palabras, dado un conjunto de datos etiquetados, el algoritmo regresa un hiperplano o división óptima que categoriza nuevos datos. En un espacio bidimensional este hiperplano es una línea que divide el espacio en dos partes y cada clase esta de un lado.
Para construir un hiperplano, las máquinas de soporte vectorial emplean un algoritmo iterativo de entrenamiento, que minimiza el error de la función.

En caso de que el conjunto de datos no se pueda separar linealmente el algoritmo de máquina de soporte vectorial nos permite aplicar una transformación y agregar otra dimensión, por ejemplo: el eje z, de manera que se puedan separar los puntos dividiéndolos desde la perspectiva del nuevo eje, y después, simplemente se regresa al plano original, estas transformaciones son llamadas “kernels” (Patel, 2017).
En caso de que las clases en el conjunto de datos estén muy cerca unas de otras o que algunos puntos de una clase queden dentro de otra se utiliza el parámetro de tolerancia, que soporta algunos puntos de una clase diferente para no sobreentrenar la frontera. Otra opción es encontrar una frontera perfecta con 0 tolerancia, pero encontrar esta frontera perfecta toma mucho tiempo.

Otro parámetro importante es Gamma, este nos permite definir la influencia de los puntos que están cerca de la frontera, es decir, al tener un gamma bajo los puntos mas lejanos de la frontera se toman en cuenta, al tener un gamma alto los puntos más cercanos de la frontera son los que se consideran. Por último, el margen, que nos sirve para determinar que tan cerca esta la línea divisoria de cada clase, y donde lo mejor es tener una distancia similar entre clases para los dos lados. Adaptar estos parámetros nos permite encontrar una frontera no-lineal con más precisión en menos tiempo. (Tomás, L et al., 2005)

3.2 ALGORITMO DE BOSQUE ALEATORIO

Bosque aleatorio es un algoritmo de clasificación con una estructura simple. Para entender cómo funciona un bosque aleatorio primero tenemos que saber cómo funciona un árbol de decisión. (Saraswat, s.f.)

a. Dado un conjunto de datos, un árbol hace niveles o particiona los datos con reglas (if-else). Es decir, un árbol crea reglas y estas se determinan a partir de la contribución de esa variable a la pureza de los nodos hijos.

b. En la Figura 3.2.1, la variable X1 da como resultado los nodos hijos con mayor pureza, por lo tanto, esa variable se convierte en un nodo raíz. Un nodo raíz es de las variables más importantes en un conjunto de datos.

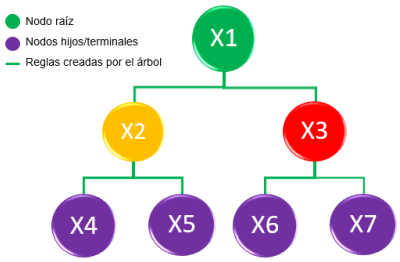


Figura 3.2.1: Árbol de decisión.

c. Para determinar dónde se separa el árbol se usa la entropía (Ecuación 3.2.1) - la entropía es una medida de la impureza del nodo (Saraswat, s.f.).

$$H(X) = -\sum_{x} p(x) \log p(x) \qquad (3.2.1)$$

De manera que un nodo se va a dividir si su impureza es mayor a la del umbral establecido, de otra manera solo es un nodo terminal.

d. Una vez que sabemos cómo funciona un árbol de decisión formamos un conjunto de árboles de decisión que se transforman en un bosque, y este funciona de la siguiente manera:

   i. Se forma un subconjunto (S2, Figura 3.2.2) de entrenamiento con muestras de tamaño n y muestreo de reemplazo aleatorio a partir del conjunto original de muestras (S1).

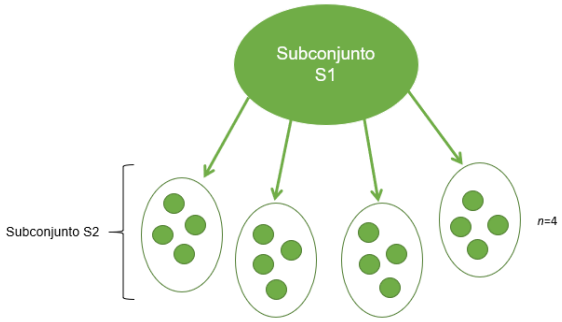


Figura 3.2.2: Subconjunto de entrenamiento.

ii. Si tenemos *M* variables de entrada, un número *m*<*M* se debe de especificar de manera que se seleccionen m variables de *M*. Aquí se emplea el método de división por nivel de impureza como se describe en la sección c para separar el nodo y crear a los nodos hijos con mayor pureza posible. El valor de *m* es constante a lo largo del crecimiento del bosque.

iii. Cada árbol crece sin límite de profundidad.

iv. Finalmente, varios árboles han crecido y se obtiene una predicción final por medio de votación o promedio. Figura 3.2.3

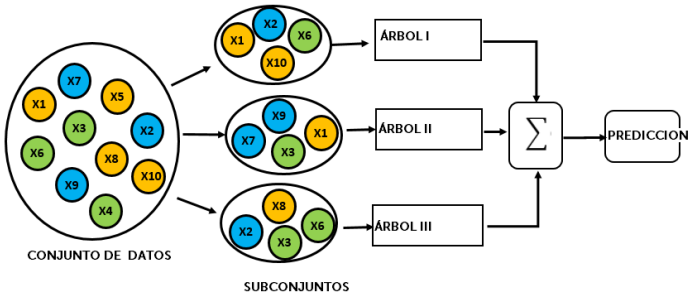


Figura 3.2.3: Formación del bosque aleatorio.

3.3 CLASIFICADOR BAYESIANO INGENUO

Este clasificador es muy práctico y su desempeño se ha probado similar al de clasificadores hechos con redes neuronales. Además, se eligió para nuestro artículo por su utilidad para clasificar texto. Este clasificador aplica técnicas de aprendizaje donde cada punto x de un conjunto tiene una función f(x) que puede tomar solo un valor de un conjunto finito V. Primero el clasificador se entrena con un conjunto de datos de entrenamiento, después se le presenta un conjunto de datos nuevos. El clasificador después tiene que clasificar los datos nuevos o predecir su clase. La técnica que usa este clasificador para predecir la clase de los nuevos datos es asignar el valor más probable, u_{MAP} dados los valores pasados $(a_1,a_2,\ldots,a_n)$ que describen a esa instancia. (Mitchell, 1997)

$$v_{MAP} = \underset{v_j \in V}{\text{argmax}} \; \text{P}(v_j \,|a_1, a_2 \ldots a_n) \qquad \textbf{(3.3.1)}$$

Podemos utilizar el teorema de Bayes y la expresión quedaría de esta manera:

$$v_{MAP} = \underset{v_j \in V}{\text{argmax}} \; \frac{\text{P}(a_1, a_2 \ldots a_n|v_j)\text{P}(v_j)}{\text{P}(a_1, a_2 \ldots a_n)} \qquad \textbf{(3.3.2)}$$

$$= \underset{v_j \in V}{\text{argmax}} \; \text{P}(a_1, a_2 \ldots a_n|v_j)\text{P}(v_j)$$

Con la ecuación 3.3.2 prodriamos estimar los dos términos basandonos en los datos de entrenamiento. Es fácil estimar el valor de $\text{P}(v_j)$ contando la frencuenca con la que cada valor esperado ocurre en los datos de entrenamiento. Pero esto solo se puede si se tiene una base de datos grande. El clasificador Bayesiano esta basado en la teoría sencilla de que los valores de los atributos son independientes del valor final esperado. Es decir, dado un valor esperado para una instancia, la probabilidad de observar la conjuncion $a_1,a_2,\ldots,a_n$ es el producto de las probabilidades de los atributos individuales: $\text{P}(a_1,a_2,\ldots,a_n|v_j) = \Pi_i \text{P}(a_i|v_j)$. Si sustituimos esto en la ecuación (3.3.2), tenemos la fórmula que utiliza el clasificador Bayesiano:

$$v_{NB} = \underset{v_j \in V}{\text{argmax}} \; \text{P}(v_j) \prod_i P(a_i|vj) \qquad \textbf{(3.3.3)}$$

Donde v_{NB} es el valor que se espera que regrese el clasificador.

Para resumir, el clasificador Bayesiano es aquel que asume la independencia entre los elementos de un conjunto, por ejemplo, dice que la probabilidad de encontrar las palabra "aprendizaje" seguida de "automático" es la misma que la de encontrar la palabra "pato" y aunque esto no es real es necesario asumir esa probabilidad ya que sin esta el número de probabilidades por términos sería demasiado grande (Mitchell, 1997).

3.4 INDICADOR MYERS-BRIGGS

Este indicador nació para hacer la teoría de los tipos psicológicos descrita por C.G. Jung entendible y útil para la vida de las personas. La esencia de la teoría es que nuestro comportamiento es más consistente de lo que creemos, todos tenemos diferencias básicas en la manera en que preferimos usar nuestra percepción y juicio. Si la gente tiene distintas formas de percibir nueva información y de llegar a conclusiones entonces es lógico que tengan diferentes intereses, reacciones, valores, motivaciones y habilidades. Para dividir los tipos de personalidad diferentes se establecieron cuatro ejes distintos (The Myers & Briggs Foundation, 2018):

Tabla 3.4.1: 4 Ejes principales del Indicador Myers-Briggs.

Mundo favorito:	¿Exterior o interior?	Extroversión(E) /Introversión (I)
Información:	¿Prefieres concentrarte en la información básica que te llega o prefieres agregar significado?	Sensorial(S) / Intuitivo(I)
Decisiones:	¿Al hacer decisiones prefieres primero ver la lógica y después las circunstancias?	Pensar (T) / Sentir (F)
Estructura:	¿Prefieres tomar un lado o permanecer abierto a nueva información?	Juzgar (J) / Percibir (P)

Una vez que se decide la categoría de una persona para cada uno de los 4 ejes se ejemplifica su personalidad con un código de 4 letras. Los 16 tipos posibles de personalidad se muestran a continuación:

Figura 3.4.1: 16 tipos de personalidad según el indicador Myers-Briggs.

De manera tradicional alguien puede tomar esta prueba de personalidad contestando preguntas como las que se muestran en la siguiente figura:

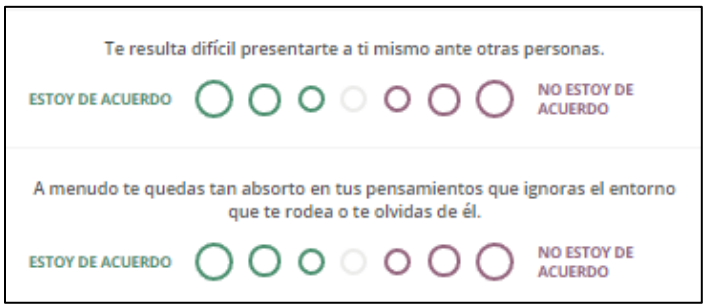


Figura 3.4.2: Ejemplos de preguntas y escala de respuestas del indicador Myers-Briggs.

3.5 PERSONALIDAD Y DESEMPEÑO LABORAL

Los investigadores organizacionales se enfocan actualmente en la relación entre un empleado y su personalidad. Una investigación previa llevada a cabo por (Barrik & Mount, 2006) utilizando el modelo de personalidad de los cinco grandes sugiere que ciertas características de la personalidad están fuertemente ligadas al desempeño de un empleado, por ejemplo, una posición en ventas es perfecta para alguien con un alto nivel de extroversión y juicio (calificador) (Hurtz & Donovan, 2001). El clasificador automático de personalidad tiene la intención de funcionar como una herramienta para reclutadores de recursos humanos que necesitan encontrar una característica específica en un candidato. Esta herramienta reduce la probabilidad de obtener respuestas sesgadas por parte de los candidatos ya que el contenido utilizado para la evaluación proviene de sus redes sociales.

4 DESARROLLO DEL EXPERIMENTO

4.1 METODOLOGÍA

La figura 4.1.1 muestra el proceso que se siguió para realizar esta investigación. Ya que se utilizó el modelo más conocido para probar algoritmos de aprendizaje automático se puede resumir en los siguientes pasos: recolección de datos, limpieza, vectorización, entrenamiento y pruebas de clasificación.

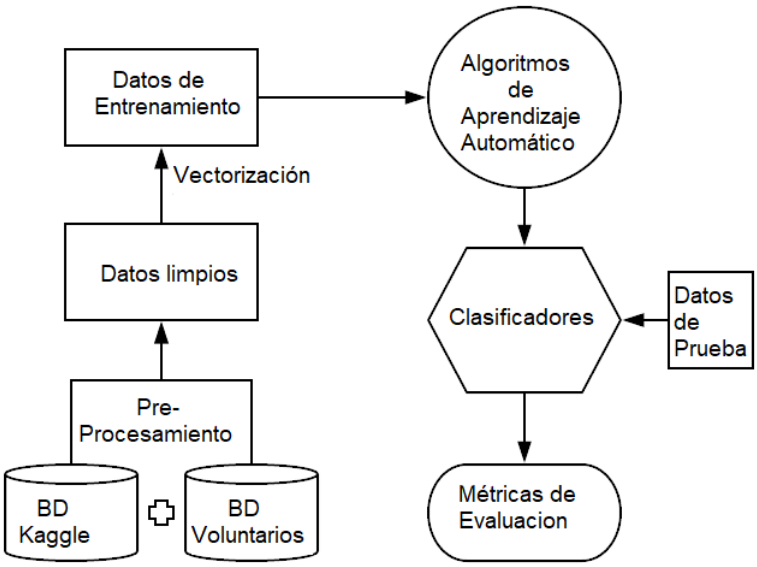


Figura 4.1.1 – Modelo de pruebas para algoritmos de aprendizaje automático.

4.2 CONJUNTO DE DATOS

Los datos para el entrenamiento se obtuvieron por medio de la descarga en el sitio kaggle[1] donde se encuentra una base de datos con los posts de más de 8 mil usuarios y su resultado del indicador Myers-Briggs. De estos más de 8 mil registros se tradujeron en total 852 al español y se usaron como datos de entrenamiento para los 3 clasificadores. Para los datos de prueba se consiguieron 44 estudiantes voluntarios de la facultad de informática de la Universidad Autónoma de Querétaro quienes aceptaron tomar la prueba y compartir el contenido de su perfil de Twitter para nuestro experimento.

En la siguiente figura se muestra un fragmento de la base de datos de entrenamiento y la de prueba:

4	ENTP	"Creí en Dios toda mi vida hasta hace un año. Mi madre creía mucho en Dios y nos alentó a mí y a mis hermanos a ser de la misma manera. Ella no nos forzó en nuestras gargantas o ... ||| Me acecho en todas partes para ser honesto. ||| Tuve un cone...
5	INTP	'Estoy rebotando dentro y fuera de este hilo como lo permite mi temperamento, lo admitiré. Estoy usando mis reservas de temperamento para responder esto. La letra cursiva no es razonable. Eso no va a argumentar en contra. En ... ||| Solo diré que...
6	ISFJ	'Vencí eso buscando a un compañero de casa que trabajaba meses a la vez, y cuando estaba en casa, visitó a sus compañeros después de que terminaron el trabajo, así que estaba solo el 80% del tiempo: D oh, yo no ... | || No hay pegajoso para que s...

Figura 4.2.a - Base de datos de entrenamiento

32	ENTP	Cada vez que tomo decisiones, trato de reducir los riesgos y problemas que puedan existir.???????? Si una persona no tiene sueños no tiene razón de vivir, soñar es necesario aún cuando el sueño va más allá de la re... https://t.co/QBdO2iU68X
33	ENTP	RT @_Cinthhya: Exactamente un mes para mi cumple ! ???? #PalPendiente ???? Pensando en ti!! ???????????? https://t.co/6MfIY7nXtP @lapizito123 Carnal yo solo veía acábatelo por ti! Haaa eres quien hacia reír!!! #Uvimeza @_Cinthhya vamos!...
34	INTP	RT @MegBucher: Heading to bed, hopefully we hit 700 RTs and I can give you guys this #ArcadeMF code in the morning! https://t.co/PdMq7YV0UG RT @cdollarsc: Need 40 more followers! RT @cdollarsc: 34 more followers! RT @freddyomana94: Follo...

[1] https://www.kaggle.com/datasnaek/mbti-type

Figura 4.2.b - Base de datos de prueba

Una vez recabados los datos se hizo un estudio para determinar la distribución de los datos tanto de la muestra obtenida de internet del sitio kaggle como de los datos que se obtuvieron de los participantes voluntarios, en la tabla 4.2.1 se muestra el promedio de edad entre los estudiantes voluntarios que tomaron el indicador Myers-Briggs. En la tabla 4.2.2 se muestra como están divididos sus resultados por eje y por último en las tablas 4.2.3 y 4.2.4 se muestra la distribución de la base de datos de kaggle y de la muestra en total:

Tabla 4.2.1: Características del conjunto de datos de prueba.

Característica	Valor
Tamaño de la muestra (n)	44
Edad promedio (años)	22.72
Distribución de género	35 masculino – 9 femenino
Edad Min - Max	19 – 36

Tabla 4.2.2: Distribución de tipos de personalidad del conjunto de datos de prueba.

Característica	Valor
Introversión/Extroversión	33 / 11
Sensorial/Intuitivo	6 / 38
Pensar/Sentir	17 / 27
Juzgar/Percibir	27 / 17

Tabla 4.2.3: Distribución de tipos de personalidad del conjunto de datos de entrenamiento.

Característica	Valor
Introversión/Extroversión	191 / 661
Sensorial/Intuitivo	165 / 687
Pensar/Sentir	426 / 426
Juzgar/Percibir	523 / 329

Tabla 4.2.4: Distribución de tipos de personalidad del conjunto de datos entero.

Característica	Valor
Introversión/Extroversión	224 / 672
Sensorial/Intuitivo	171 / 725
Pensar/Sentir	443 / 453
Juzgar/Percibir	550 / 346

4.3 PRE-PROCESAMIENTO

Después de un preprocesamiento llevado a cabo con librerías de Python que incluye eliminar signos de puntuación, HTML, emojis, números, urls, y demás caracteres, poner todo en minúsculas y dejar solamente la raíz de las palabras en ambas bases de datos se procedió a entrenar y probar a los clasificadores. Sin embargo, primero se vectorizaron los datos, esto se llevó a cabo por medio de un proceso conocido como extracción de características.

4.4 EXTRACCIÓN DE CARACTERÍSTICAS

Este proceso es utilizado en casi todos los ámbitos de aprendizaje automático donde es necesario transformar imágenes o palabras, como en este caso, en vectores. Existen muchas técnicas y algoritmos que se le pueden aplicar a los datos para que esta se convierta en algo más fácil de procesar por una máquina. Estas técnicas

entonces son una manera de simplificar los datos con los que vamos a trabajar para después introducirlos al clasificador y obtener resultados más certeros. En este caso se utilizó la técnica más común que incluye la "tokenización" o separación de palabras, seguido de la obtención de la frecuencia de las palabras, para por último normalizar este vector de manera que las palabras más utilizadas no obtengan mayor importancia a la hora de definir a una clase.

El "vectorizador" en este caso se configuró para tomar en cuenta solo el 80% del total del vocabulario que encontró en toda la base de datos, en la Figura 4.4.1 se puede ver una muestra del vocabulario de nuestra base de datos que se conforma de 432 palabras en total.

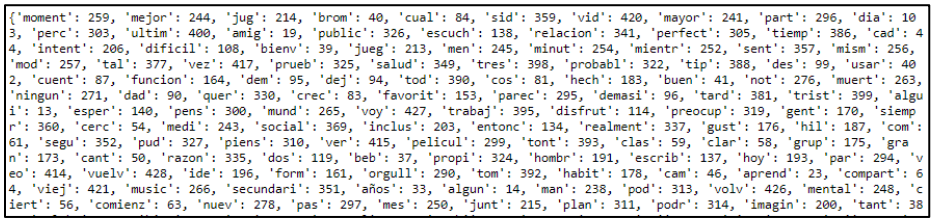


Figura 4.4.1: Muestra del vector de tamaño 432 con el vocabulario recopilado de la base de datos. La primera palabra "moment" seguida por su índice "259" o posición en el vector.

Después de obtener este vector, el siguiente paso es obtener el vector de vocabulario para cada usuario, es decir, mapear las publicaciones de cada usuario con el vector de vocabulario y crear un vector numérico específico para sus registros. En este caso se emplea un algoritmo llamado: frecuencia de término - frecuencia inversa de documento.
La primera parte del algoritmo es la frecuencia de término (Luhn, 1957) y la más fácil de usar es la frecuencia bruta, es decir, contar las ocurrencias de una palabra en el documento, sin embargo, en este caso se utiliza la frecuencia normalizada, para evitar que los posts largos tengan mayor frecuencia de palabras que los cortos. Simplemente se divide la frecuencia bruta de la palabra entre la frecuencia máxima de esa palabra en todos los registros.

$$f_{i,j} = \frac{tf_{i,j}}{\max(tf_{i,j})} \qquad \textbf{(4.4.1)}$$

$f_{i,j}$= frecuencia normalizada
$tf_{i,j}$=frecuencia del término i en el documento j
$\max(tf_{i,j})$=frecuencia máxima del término i en el documento j

La frecuencia inversa de documento (Spärck Jones, 1972) es una medida para reducir el peso de las palabras que se repiten en varios posts, y, por lo tanto, son menos importantes o representativas que las que aparecen menos. Se obtiene dividiendo el número total de documentos por el número de documentos que contienen el término y se toma el logaritmo de ese cociente:

$$idf(t, D) = \log\left(\frac{|D|}{|d \in D: t \in d|}\right) \qquad \textbf{(4.4.2)}$$

|D|= Número de documentos o posts en este caso.
|d ∈ D: t ∈ d |= Número de documentos donde aparece el término.

Aplicando el algoritmo frecuencia de término - frecuencia inversa de documento a los datos obtenemos una matriz con los componentes documento-término. En la Figura 4.4.2 se muestra un ejemplo del vector para el primer usuario.

```
[0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.04366567 0.04916774 0.         0.         0.
 0.         0.04438111 0.         0.         0.         0.13246047
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.10104016 0.         0.
 0.         0.07860859 0.         0.13126557 0.08156512 0.14652168
 0.         0.         0.171365   0.         0.08680048 0.
 0.         0.         0.08156512 0.         0.         0.
 0.06671746 0.         0.06177086 0.         0.08014684 0.15989984
 0.         0.07577311 0.         0.08860757 0.08486753 0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.03860427 0.         0.09291923
 0.13777469 0.         0.         0.10780087 0.         0.
 0.09111488 0.         0.         0.         0.05405699 0.07167368
 0.15516824 0.         0.         0.0664728  0.         0.
 0.         0.11354773 0.         0.         0.         0.
```

Figura 4.4.2: Muestra del vector normalizado entre 0 y 1 para los registros del primer usuario.

5 RESULTADOS

5.1 PRUEBA A - EXACTITUD

En esta primera fase se separa la base de datos de manera que los 852 registros del conjunto de datos obtenidos en línea se utilizan como datos de entrenamiento para el clasificador y los restantes 44 de alumnos voluntarios se utilizaron para las pruebas. En seguida se muestran los cuadros con el resultado de exactitud por clasificador y por cada rubro.

Tabla 5.1.1

Algoritmo	Indicador	Exactitud
SVM		25.00%
Bosque Aleatorio	Introversión/Extroversión	25.00%
Bayesiano Ingenuo		25.00%

Tabla 5.1.2

Algoritmo	Indicador	Exactitud
SVM		84.09%
Bosque Aleatorio	Sensorial/Intuitivo	86.36%
Bayesiano Ingenuo		86.36%

Tabla 5.1.3

Algoritmo	Indicador	Exactitud
SVM		47.73%
Bosque Aleatorio	Pensar/Sentir	45.45%
Bayesiano Ingenuo		45.45%

Tabla 5.1.4

Algoritmo	Indicador	Exactitud
SVM		61.36%
Bosque Aleatorio	Juzgar/Percibir	63.64%
Bayesiano Ingenuo		61.36%

Tabla 5.1.5

Algoritmo	Indicador	Exactitud
SVM		54.55%
Bosque Aleatorio	PROMEDIO	55.11%
Bayesiano Ingenuo		55.04%

5.2 MATRICES DE CONFUSIÓN – PRUEBA A

Las matrices de confusión como su nombre lo sugiere son matrices que describen el desempeño de un modelo. Debido a que en esta fase se hicieron las pruebas de cada rubro por separado tenemos cuatro matrices de confusión. Estas se agregan como complemento visual de manera que también se aprecien los datos a partir de los cuales se calcularon las métricas de la sección 5.3.

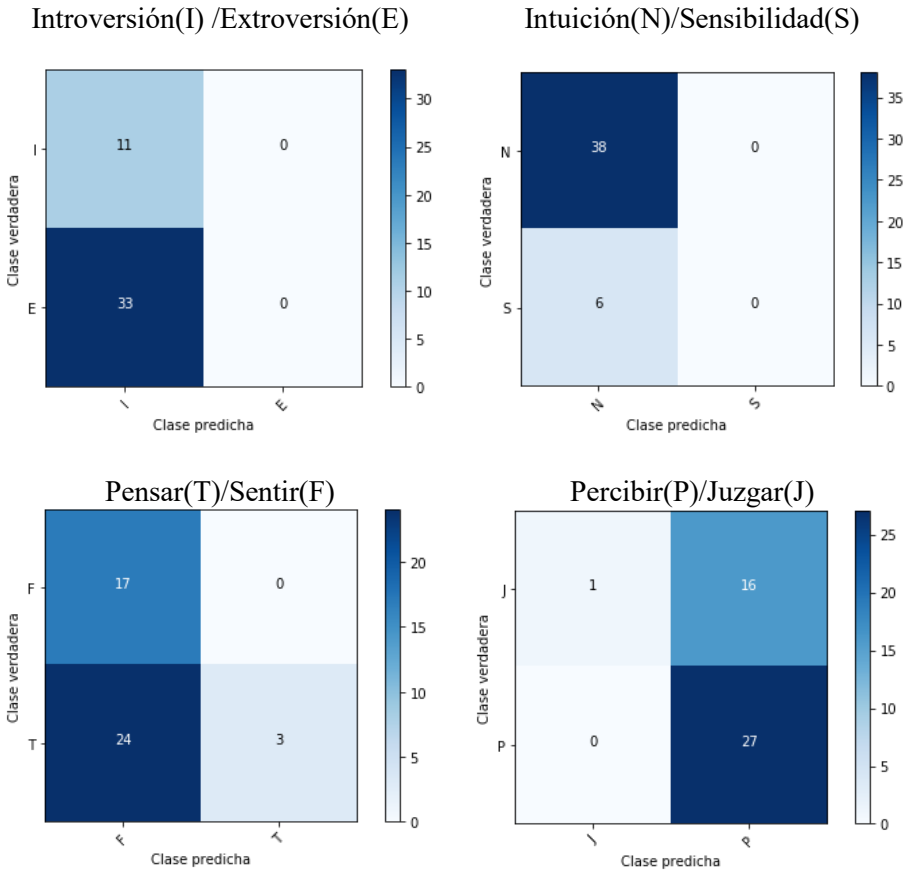


Figura 5.2.1: Matrices de confusión para la prueba A con el clasificador de Bosque Aleatorio.

Estas matrices se pueden interpretar de la siguiente manera, los verdaderos positivos y verdaderos negativos son las observaciones que se predijeron de manera correcta. Los que se deben minimizar son los falsos positivos y falsos negativos. Por ejemplo, el número de verdaderos positivos para la matriz de confusión de Introversión/Extroversión en la figura 5.2.1 son 11 mientras que 33 personas fueron incorrectamente clasificadas como Introvertidas cuando eran Extrovertidas.

Introversión(I) /Extroversión(E)
Intuición(N)/Sensibilidad(S)

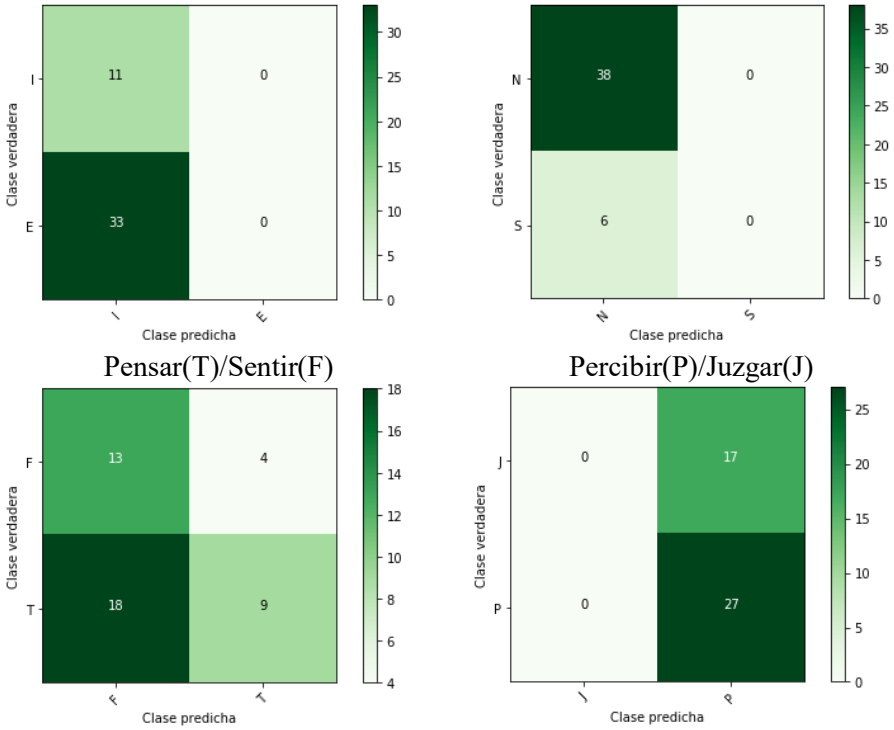
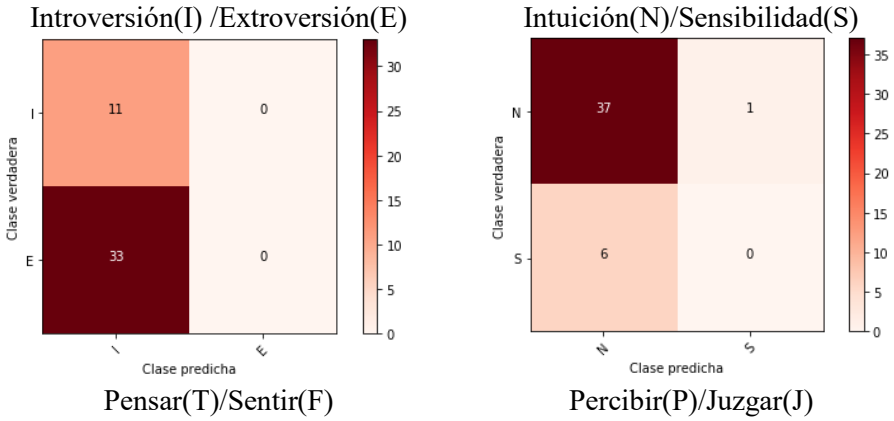


Figura 5.2.2: Matrices de confusión para la prueba A con el clasificador Bayesiano Ingenuo.



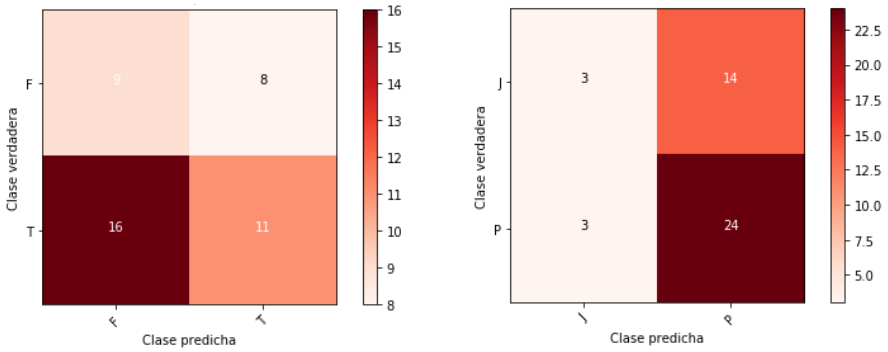


Figura 5.2.3: Matrices de confusión para la prueba A con el clasificador de Máquinas de Soporte Vectorial.

5.3 PRECISIÓN, ÍNDICE DE RECUPERACIÓN Y F1 - PRUEBA A

La exactitud de los resultados, es la métrica más común para medir el desempeño de un modelo, ya que, es igual a la proporción de las observaciones predichas de manera correcta entre las observaciones totales. Aunque esta métrica es buena para medir nuestros resultados de manera general en este caso se emplearon más mediciones ya que nuestra muestra de datos no está balanceada, es decir, nuestros valores de falsos positivos y falsos negativos no vienen en una cantidad similar. Por lo tanto, agregamos otros parámetros para evaluar el desempeño de los clasificadores.

Tabla 5.3.1

SVM	Intro/Extro	Sens/Intuit	Pensar/Sentir	Juzg/Perc	Promedio
Precisión	6.00%	74.00%	51.00%	58.00%	47.25%
Ind. Recuperación	0.00%	84.00%	48.00%	61.00%	48.25%
F1	10.00%	79.00%	48.00%	55.00%	48.00%

Tabla 5.3.2

Bayesiano	Intro/Extro	Sens/Intuit	Pensar/Sentir	Juzg/Perc	Promedio
Precisión	6.00%	75.00%	51.00%	38.00%	42.50%
Ind. Recuperación	25.00%	86.00%	45.00%	61.00%	54.25%
F1	10.00%	80.00%	45.00%	47.00%	45.50%

Tabla 5.3.3

Bosque Aleatorio	Intro/Extro	Sens/Intuit	Pensar/Sentir	Juzg/Perc	Promedio
Precisión	6.00%	75.00%	77.00%	77.00%	58.75%
Ind. Recuperación	25.00%	86.00%	45.00%	64.00%	55.00%
F1	10.00%	80.00%	35.00%	52.00%	44.25%

Los mejores resultados obtenidos para la prueba A en cuanto a clasificaciones correctas (exactitud) fueron de 25.00% para Introversión/Extroversión, Tabla 5.1.1, 86.36% para Sensorial/Intuitivo, Tabla 5.1.2, 47.73% para Pensar/Sentir, Tabla 5.1.3 y 63.64% para Juzgar/Percibir, Tabla 5.1.4. El clasificador con el mejor desempeño fue el creado con el algoritmo de Bosque Aleatorio que obtuvo un promedio de exactitud de 55.11% ,Tabla 5.1.5, este algoritmo también obtuvo las mejores métricas de precisión e índice de recuperación, Tabla 5.3.3.

5.4 EXACTITUD – PRUEBA B

Para esta prueba se mezclaron los datos de prueba y entrenamiento de manera aleatoria para promover un índice más alto de exactitud. Los datos se separaron con el 80% como datos de entrenamiento y el resto de prueba. Los resultados aparecen en las siguientes tablas.

Tabla 5.4.1

Algoritmo	Indicador	Exactitud
SVM		70.09%
Bosque Aleatorio	Introversión/Extroversión	77.23%
Bayesiano Ingenuo		76.34%

Tabla 5.4.2

Algoritmo	Indicador	Exactitud
SVM		77.23%
Bosque Aleatorio	Sensorial/Intuitivo	77.23%
Bayesiano Ingenuo		77.23%

Tabla 5.4.3

Algoritmo	Indicador	Exactitud
SVM		62.05%
Bosque Aleatorio	Pensar/Sentir	64.73%
Bayesiano Ingenuo		64.29%

Tabla 5.4.4

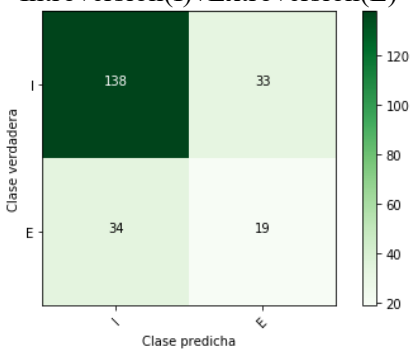
Algoritmo	Indicador	Exactitud
SVM		57.59%
Bosque Aleatorio	Juzgar/Percibir	57.14%
Bayesiano Ingenuo		57.14%

Tabla 5.4.5

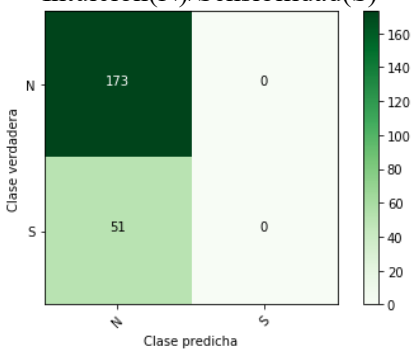
Algoritmo	Indicador	Exactitud
SVM		66.74%
Bosque Aleatorio	Promedio	69.08%
Bayesiano Ingenuo		68.75%

5.5 MATRICES DE CONFUSIÓN – PRUEBA B

Introversión(I) /Extroversión(E)



Intuición(N)/Sensibilidad(S)



Pensar(T)/Sentir(F)

Percibir(P)/Juzgar(J)

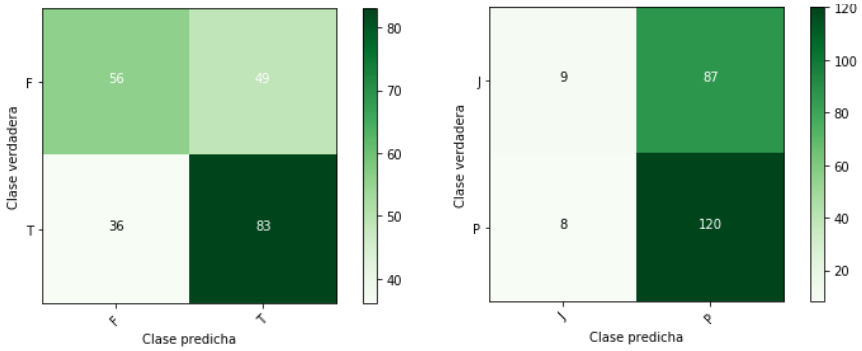


Figura 5.5.1: Matrices de confusión para la prueba B con el clasificador de Maquinas de Soporte Vectorial.

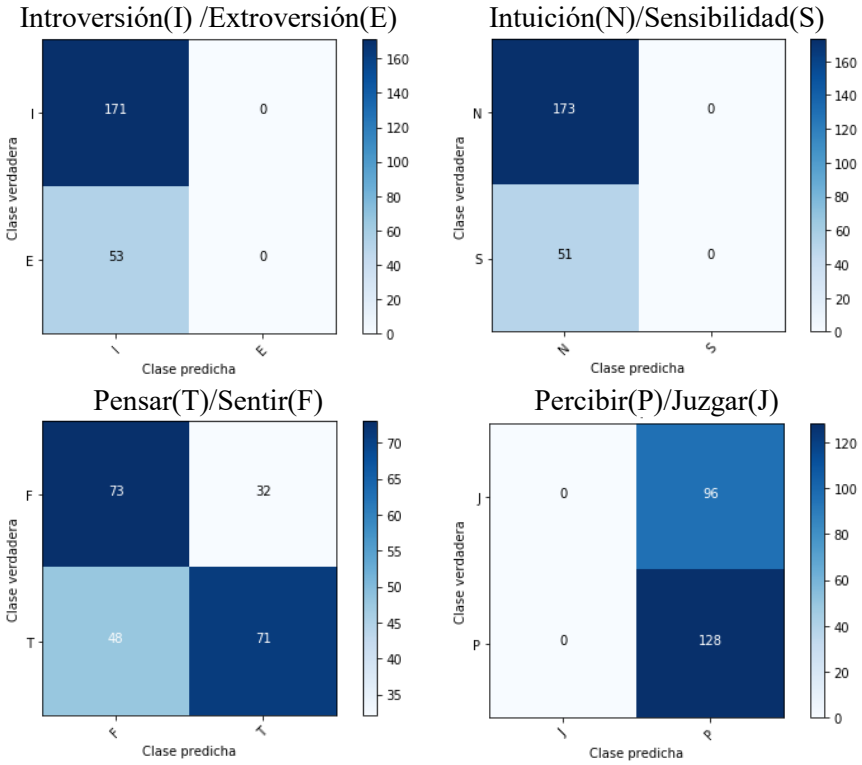
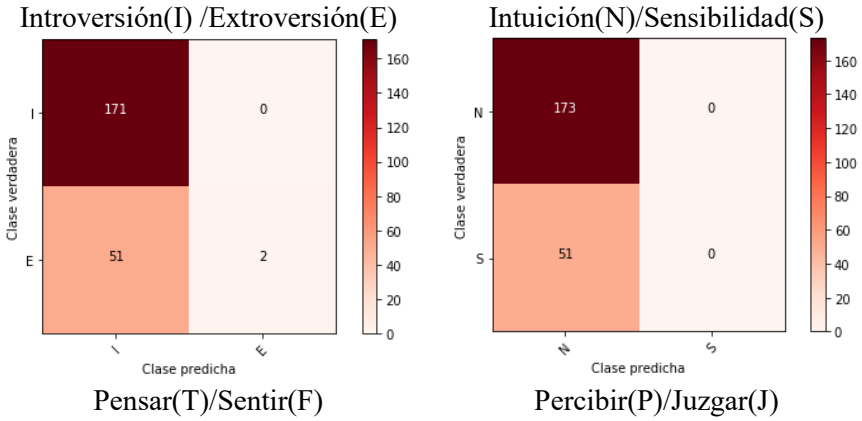


Figura 5.5.2: Matrices de confusión para la prueba B con el clasificador de Bayesiano Ingenuo.



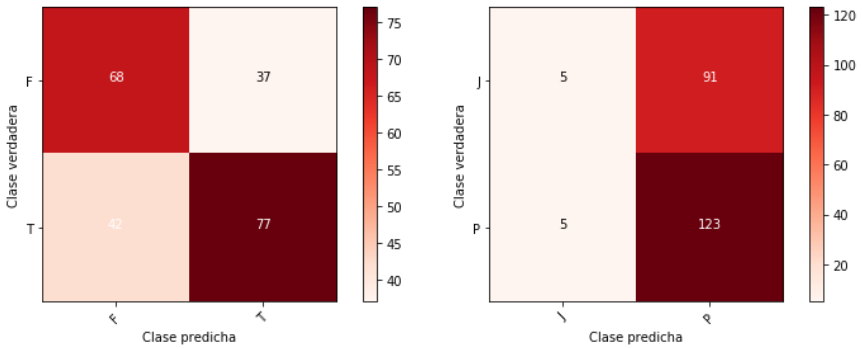


Figura 5.5.3: Matrices de confusión para la prueba B con el clasificador de Bosque Aleatorio.

5.6 PRECISIÓN, ÍNDICE DE RECUPERACIÓN Y F1 - PRUEBA B

Tabla 5.4.1 Precisión, índice de recuperación y f1 para SVM

SVM	Intro/Extro	Sens/Intuit	Pensar/Sentir	Juzg/Perc	Promedio
Precisión	70.00%	60.00%	62.00%	56.00%	62.00%
Ind. Recuperación	70.00%	77.00%	62.00%	58.00%	66.75%
F1	70.00%	67.00%	62.00%	48.00%	61.75%

Tabla 5.4.2 Precisión, índice de recuperación y f1 para Bosque Aleatorio

BosqueAleatorio	Intro/Extro	Sens/Intuit	Pensar/Sentir	Juzg/Perc	Promedio
Precisión	82.00%	60.00%	65.00%	54.00%	65.25%
Ind. Recuperación	77.00%	77.00%	65.00%	57.00%	69.00%
F1	68.00%	67.00%	65.00%	45.00%	61.25%

Tabla 5.4.3 Precisión, índice de recuperación y f1 para Bayesiano

Bayesiano	Intro/Extro	Sens/Intuit	Pensar/Sentir	Juzg/Perc	Promedio
Precisión	58.00%	60.00%	65.00%	33.00%	54.00%
Ind. Recuperación	76.00%	77.00%	64.00%	57.00%	68.50%
F1	66.00%	67.00%	64.00%	42.00%	59.75%

Los mejores resultados obtenidos para la prueba B en cuanto a clasificaciones correctas (exactitud) fueron de 77.23% para Introversión/Extroversión, Tabla 5.3.1, 77.23% para Sensorial/Intuitivo, Tabla 5.3.2, 64.73% para Pensar/Sentir, Tabla 5.3.3 y 57.59% para Juzgar/Percibir, Tabla 5.3.4. Dos de las cuatro mejores calificaciones en exactitud se obtuvieron por medio del algoritmo de bosque aleatorio, que en promedio tuvo una exactitud de 69.08%. Además de obtener el promedio más alto de exactitud, el clasificador de bosque aleatorio también fue el mejor en precisión e índice de recuperación.

Se puede notar como los resultados del experimento B son mejores que los del experimento A. Sin embargo, en los dos experimentos el clasificador con mejor exactitud y métricas resultó ser el del Bosque Aleatorio con un porcentaje de 55.11 en el experimento A y de 69.08 en el experimento B. Estos resultados pueden ser debido a que el algoritmo del bosque aleatorio hace conjuntos y subconjuntos de los datos de entrenamiento lo que reduce el efecto que pueden tener los datos más alejados de una clase, sin embargo, como se afirma en el estudio (Wolpert, 1995) siempre van a haber conjuntos de datos en los que un clasificador es mejor que otro. En este estudio se podrían

pulir todavía más los datos al igual que recolectar una muestra mayor de las clases con menos ejemplos, al igual que, también se podría intentar reducir en mayor grado el vector de vocabulario para mejorar los resultados de los tres algoritmos. Otra manera de mejorar los resultados sería utilizando técnicas lingüísticas como la asociación de las palabras con sinónimos y antónimos para hacer el vocabulario más robusto con respecto a las variaciones de vocabulario y parafraseo. Además, se podría probar una combinación de clasificadores como en (Segrera et al., 2006).

6 CONCLUSIONES

En esta investigación se revisó la capacidad de tres algoritmos famosos por sus altos niveles de exactitud y bajo nivel de complejidad para clasificar de manera automática la personalidad de usuarios de habla hispana. En dos experimentos comparamos los resultados de clasificación utilizando los clasificadores de Maquinas de Soporte Vectorial, Bosque Aleatorio, y Bayesiano Ingenuo. Los mejores resultados se obtuvieron usando el Bosque Aleatorio. Con el tratamiento que se le dio al conjunto de datos pudo clasificar correctamente el 69.08% de los casos. Por lo tanto, los resultados son positivos y con algo de mejoras en la cantidad de datos para la muestra y mejor preprocesamiento, este clasificador podría mejorar y obtener un mejor porcentaje para clasificar la personalidad de los usuarios. La aplicación que puede tener esta herramienta en el campo de reclutamiento es grande, ya que las empresas hoy más que nunca buscan una fuerza laboral que se amolde a su filosofía, por lo tanto, un clasificador de personalidad que funcione con tal rapidez y exactitud supondría una buena referencia para encontrar los rasgos que se buscan en una nueva contratación. Para continuar con esta investigación se plantea un nuevo experimento donde, después de incrementar la exactitud del clasificador, este pueda ser utilizado en un proceso de reclutamiento en un escenario concreto y delimitado donde los candidatos obtengan sus resultados y retroalimentación del sistema, así como una entrevista personal para cotejar los resultados que se obtengan mediante estas dos técnicas y así ver su grado de exactitud también.

REFERENCIAS

Barrik, M. R., & Mount, M. K. (2006). The Big Five Personality Dimensions and Job Performance: a Meta-analysis. *Freshwater Biology*.

Cortes, C. &. (1995). Support-Vector Networks. *Machine Learning*, 273-297.

Tomás, J. L. (2005). Una aproximación multilingüe a la clasificación de preguntas. *Procesamiento del Lenguaje Natural*, 391-398.

Faliagka, E., Tsakalidis, A., & Tzimas, G. (2012). An Integrated E-Recruitment System for Automated Personality Mining and Applicant Ranking. *Internet Research*, 551-568.

Gill, A. J. (2009). What are they blogging about?Personality, topic and motivation in blogs. . *InProceedings of the 3rd internationalAAAI conference on weblogs and social media*, 18-25.

Hurtz, G., & Donovan, J. (2001). Personality and Job Performance: The Big Five Revisited. *The Journal of applied psychology*, 79-869.

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of research and development*.

Mitchell, T. (1997). *Machine Learning.* McGraw-Hill.

Oberlander, J. (2006). Whose thumb is it anyway?: classifying author personality from weblog text. . *En Proceedings of the COLING/ACL on Main conference poster sessions*, pp. 627-634.

Ortigosa, A., Carro, R., & Quiroga, J. (2014). Predicting user personality by mining social interactions in Facebook. *Journal of Computer and System Sciences*, 57-71.

Patel, S. (2017). *Medium.* Retrieved from https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

Pennebaker, J. W. (2007). The development and psychometric properties of LIWC2007. *LIWC.net*.

Saraswat, M. (n.d.). *Hacker Earth*. Retrieved from https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/tutorial-random-forest-parameter-tuning-r/tutorial/

Segrera, S., & Moreno, M. N. (2006). An experimental comparative study of web mining methods for recommender systems. *Proceedings of the Sixth WSEAS International Conference on Distance Learning and WebEngineering*, 56-61.

Jones, K. (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, 11-21.

Tandera, T., Suhartono, H., Suhartono, D., Wongso, R., & Prasetio, Y. (2017). Personality Prediction System from Facebook Users. *Procedia Computer Science*, 604 - 611.

The Myers & Briggs Foundation. (2018). Retrieved from https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/home.htm?bhcp=1

Velasco, A. (2017). *El Financiero*. Retrieved from http://www.elfinanciero.com.mx/tech/mexicanos-pasan-horas-trabajando-no-en-internet

Wolpert, D. M. (1995). No Free Lunch Theorems for Search. *Technical Report SFI-TR-95-02-010*.