Proyecto Final: Sistemas Basados en Conocimiento

Aprendizaje Supervisado a 16 personalidades dataset.

Débora Joselyn Tolentino Díaz
UNAM Enes, Morelia
deborajtd.12@gmail.com

Arely Hilda Luis Tiburcio
UNAM Enes, Morelia
arelyluis@comunidad.unam.mx

ABSTRACT

El siguiente documento realiza una implementación de los métodos: Naive Bayes, Regresión Logistica, Máquinas de Soporte Vectorial (SVM), K-Nearest Neighbors y Árboles de Decisión (técnicas pertenecientes al aprendizaje supervisado), a un conjunto de datos que simula las respuestas de 59,999 personas que aplicaron el test de las 16 personalidades.

En dicho test cuyo nombre completo es Myers-Briggs Type Indicator® (MBTI®), mediante la creación de 60 preguntas, se busca que la teoría de los tipos psicológicos descrita por Carl Jung sea comprensible y útil en la vida de las personas.

ACM Reference Format:
Débora Joselyn Tolentino Díaz and Arely Hilda Luis Tiburcio. 2023. Proyecto Final: Sistemas Basados en Conocimiento: Aprendizaje Supervisado a 16 personalidades dataset.. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 DESCRIPCIÓN Y PRE-PROCESAMIENTO DE DATOS

El conjunto de datos a analizar fue creado por Anshul Mehta y fue distribuido en la plataforma Kaggle en el año 2022. Sus datos contenidos son de origen sintético, sin embargo han pasado por diversos cambios en sus versiones con el fin de poder simular lo más parecido posible respuestas dadas por una persona.

El siguiente dataset consta de 61 columnas, las primeras 60 son las preguntas correspondientes al test, la columna 61 es la personalidad resultante al aplicar el cuestionario. Cada pregunta tiene 7 posibles respuestas que van desde el valor -3 que es totalmente de acuerdo al valor 3 que es totalmente en desacuerdo.

Como primer paso se renombraron las columnas, ya que estas contenian la pregunta completa en cada instancia , esto con fines prácticos en las visualizaciones. Al ser un dataset sintético no existen datos faltantes y estos se encuentran bien balanceados, como podemos observar en la figura 2.

Se realizaron gráficos de barras, con el fin de mostrar la distribución de respuestas por cada pregunta.

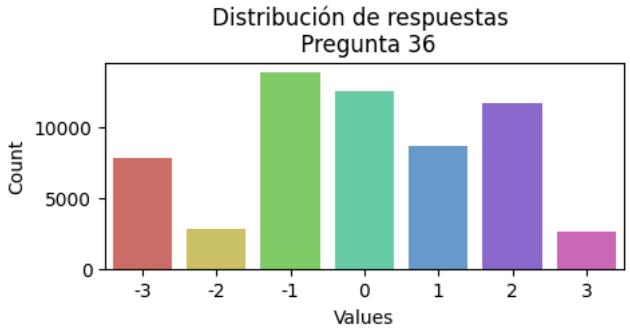

Figure 1: Ejemplo de distribución de preguntas.

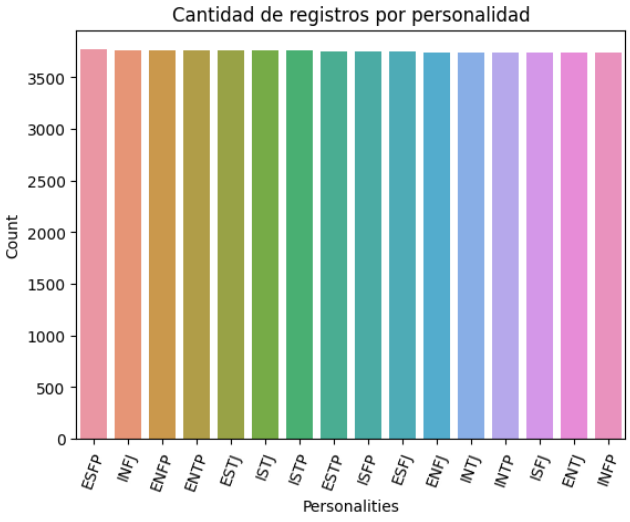

Figure 2: Gráfica que muestra el balance de datos en base a las personalidades.

2 APLICACIÓN DE MÉTODOS DE APRENDIZAJE SUPERVISADO

En ésta sección se aplican a nuestro conjunto de datos algunos métodos del aprendizaje supervisado, esto con el fin de encontrar el mejor de ellos basándonos en las métricas: precisión, recall y f1-score, además se hace uso de validación cruzada. Se dividió el conjunto de datos en validación y prueba dando un 0.25 a los datos de prueba.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

NAIVE BAYES

Naive Bayes es un algoritmo de aprendizaje automático supervisado simple que utiliza el teorema de Bayes con fuertes suposiciones de independencia entre las características para obtener resultados. Eso significa que el algoritmo simplemente asume que cada variable de entrada es independiente.

Resulta bastante conveniente aplicar Naive Bayes a nuestro conjunto, ya que al tener 59,999 datos etiquetados, es posible entrenar un modelo bastante robusto, que sea capaz de predecir una instancia dadas las respuestas de las preguntas.

Se llevó a cabo con sklearn de Python, utilizando CategoricalNB pues puede manejar características categóricas y discretas. Este calcula las probabilidades que se utilizaran para hacer las clasificaciones. En la tabla 1 podemos ver los resultados obtenidos.

Table 1: Naive Bayes - Métricas

	Precision	recall	f1-score	support
ENFJ	0.91	0.94	0.92	937
ENFP	0.93	0.94	0.94	957
ENTJ	0.95	0.95	0.95	912
ENTP	0.97	0.95	0.96	9558
ESFJ	0.87	0.81	0.84	963
ESFP	0.95	0.94	0.95	962
ESTJ	0.92	0.95	0.93	937
ESTP	0.92	0.95	0.94	917
INFJ	0.92	0.93	0.92	898
INFP	0.90	0.88	0.89	909
INTJ	0.85	0.91	0.88	972
INTP	0.89	0.91	0.90	936
ISFJ	0.89	0.86	0.87	982
ISFP	0.95	0.92	0.93	933
ISTJ	0.91	0.87	0.89	906
ISTP	0.91	0.93	0.92	924
accuracy			0.91	15000
macro avg	0.91	0.91	0.91	15000
weighted avg	0.91	0.91	0.91	15000

REGRESIÓN LOGÍSTICA

La regresión logística es un método conocido en aprendizaje supervisado que se usa para predecir la probabilidad de un resultado y es popular para las tareas de clasificación. El algoritmo predice la probabilidad de aparición de un evento ajustando los datos a una función logística.

Requiere un conjunto de datos con etiquetas. Puede entrenar el modelo proporcionando el modelo y el conjunto de datos etiquetado como entrada para un componente como Entrenar modelo. Después, el modelo entrenado puede usarse para predecir valores para los nuevos ejemplos de entrada.

La regresión logística puede manejar eficientemente conjuntos de datos con un gran número de características. Además, puede lidiar con la multicolinealidad, es decir, la presencia de correlaciones entre las características.
Es así, que para la parte de regresión logística, se hizo uso de la librería sklearn de Python, se construyó el modelo con 700 iteraciones. A continuación en la tabla 2 se muestran los resultados de las métricas al aplicar el modelo.

Table 2: Regresión Logística - Métricas

	Precision	recall	f1-score	support
ENFJ	0.92	0.92	0.92	912
ENFP	0.91	0.93	0.92	913
ENTJ	0.94	0.95	0.95	996
ENTP	0.96	0.95	0.95	1008
ESFJ	0.86	0.87	0.86	893
ESFP	0.93	0.94	0.93	960
ESTJ	0.94	0.94	0.94	957
ESTP	0.93	0.95	0.94	967
INFJ	0.93	0.92	0.92	942
INFP	0.91	0.89	0.90	967
INTJ	0.89	0.90	0.89	951
INTP	0.91	0.90	0.91	883
ISFJ	0.89	0.90	0.90	903
ISFP	0.92	0.91	0.91	949
ISTJ	0.91	0.88	0.90	925
ISTP	0.90	0.91	0.90	874
accuracy			0.92	15000
macro avg	0.92	0.92	0.92	15000
weighted avg	0.92	0.92	0.92	15000

MÁQUINAS DE SOPORTE VECTORIAL

Las máquinas de vector soporte o Support Vector Machines (SVM) son otro tipo de algoritmo de machine learning supervisado aplicable a problemas de regresión y clasificación, aunque se usa más comúnmente como modelo de clasificación.

Se implementó utilizando, de igual manera, la librería sklearn de Python, se aplicaron 3 modelos en los que a cada uno se le modificó el parámetro 'kernel', el cuál es la función matemática que nos permitirá calcular la similitud entre los puntos de datos en el espacio de características, se emplearon las opciones 'linear', 'poly' y 'rbf'. Ésto con el fin de obtener el que se ajustara mejor a nuestros datos y nos proporcionara mejores resultados.

El que se encontró más optimo fue 'rbf', ya que nos permite modelar relaciones no lineales y es bastante útil cuando los datos no son linealmente separables en el espacio de características original, ya que los transforma y mapea en un espacio de dimensionalidad superior donde se vuelven linealmente separables. Resultados el tabla 3.

Table 3: SVC: rbf

	precision	recall	f1-score	support
ENFJ	0.99	0.99	0.99	912
ENFP	0.98	0.99	0.99	913
ENTJ	0.98	0.99	0.99	996
ENTP	0.99	0.98	0.99	1008
ESFJ	0.99	0.99	0.99	893
ESFP	0.99	0.98	0.99	960
ESTJ	0.99	0.99	0.99	957
ESTP	0.99	0.99	0.99	967
INFJ	0.99	0.99	0.99	942
INFP	0.98	0.99	0.98	967
INTJ	0.99	0.99	0.99	951
INTP	0.99	0.98	0.98	883
ISFJ	0.99	0.99	0.99	903
ISFP	0.99	1.00	0.99	949
ISTJ	0.99	0.99	0.99	925
ISTP	0.99	0.98	0.99	874
accuracy			0.99	15000
macro avg	0.99	0.99	0.99	15000
weighted avg	0.99	0.99	0.99	15000

K-NEAREST NEIGHBORS

Al aplicar el modelo KNN al conjunto de datos, un paso necesario es definir la k con la cual se implementará el modelo, se calcularon las méticas f1 y accuracy, con k's del 1 al 18, posterior a ello se eligió la mejor de ellas en base a sus metricas y esa k fue la que se eligió para entrenar el modelo. Se eligió una distancia Euclidiana, después de haber probado con otros tipos como Manhattan y Minkowski. En la tabla 4 vemos los resultados de éste modelo.

Table 4: KNN Métricas

	precision	recall	f1-score	support
ENFJ	0.99	0.99	0.99	937
ENFP	0.99	0.98	0.99	938
ENTJ	0.99	0.99	0.99	915
ENTP	0.99	0.99	0.99	1906
ESFJ	0.99	0.99	0.99	980
ESFP	0.99	0.98	0.98	920
ESTJ	0.99	0.99	0.99	949
ESTP	0.99	0.99	0.99	916
INFJ	0.99	0.99	0.99	911
INFP	0.99	0.99	0.99	967
INTJ	0.99	0.99	0.99	934
INTP	0.98	0.99	0.98	959
ISFJ	0.98	0.99	0.98	950
ISFP	0.99	0.99	0.99	894
ISTJ	0.99	0.99	0.99	973
ISTP	0.99	0.99	0.99	951
accuracy			0.99	15000
macro avg	0.99	0.99	0.99	15000
weighted avg	0.99	0.99	0.99	15000

ÁRBOLES DE DECISIÓN

Árboles de decisión es un tipo de algoritmo de aprendizaje automático supervisado, clasifica y lleva a cabo la regresión de los datos utilizando respuestas verdaderas o falsas a determinadas preguntas

Se aplicó usando la librería sklearn de Python, se experimentó con 2 modelos, modificandoles el criterio cuyas opciones son 'Gini' o 'Entropia', con éste se mide la calidad de una división en el árbol. Mientras Gini mide la homogeneidad de las clases, la Entropia mide la incertidumbre. Al no obtener una diferencia sustancial en los resultados, se continuó la exploración utilizando Gini, ya que es más eficiente computacionalmente. Observamos los resultados en la tabla 5.

Table 5: Árboles Métricas

	precision	recall	f1-score	support
ENFJ	0.64	0.66	0.65	937
ENFP	0.64	0.65	0.65	938
ENTJ	0.68	0.67	0.67	915
ENTP	0.62	0.65	0.63	906
ESFJ	0.60	0.59	0.59	980
ESFP	0.64	0.65	0.65	920
ESTJ	0.64	0.67	0.65	949
ESTP	0.66	0.66	0.66	916
INFJ	0.62	0.64	0.63	911
INFP	0.68	0.62	0.65	967
INTJ	0.62	0.62	0.62	934
INTP	0.65	0.60	0.62	959
ISFJ	0.62	0.62	0.62	950
ISFP	0.61	0.64	0.62	894
ISTJ	0.62	0.61	0.62	973
ISTP	0.63	0.64	0.64	951
accuracy			0.64	15000
macro avg	0.64	0.64	0.64	15000
weighted avg	0.64	0.64	0.64	15000

3 CONCLUSIONES

Los resultados obtenidos mostraron que la mayoría de los métodos tuvieron un desempeño notable, con excelentes valores de precisión, recall y f1-score. Las Máquinas de Soporte Vectorial con kernel 'rbf' demostraron ser la mejor opción hablando de precisión logrando en promedio el 99%, vemos también que Naive Bayes alcanza del 91%, Regresión Logística del 92%. Esto demuestra que los modelos son capaces de predecir correctamente las personalidades en función de las respuestas proporcionadas en el cuestionario. En resumen, este trabajo proporciona la implementación y evaluación de varios métodos de aprendizaje supervisado aplicados al conjunto de datos del test de las 16 personalidades. Los resultados obtenidos son prometedores y demuestran la utilidad de estos métodos en la predicción de las personalidades de las personas en base a sus respuestas.

4 BIBLIOGRAFÍA

1.4. Support Vector Machines. (s. f.). scikit-learn. https://scikit-learn.org/stable/modules/svm.html

APÉNDICE A. PREGUNTAS DEL TEST

(1) 'You regularly make new friends.'
(2) 'You spend a lot of your free time exploring various random topics that pique your interest'
(3) 'Seeing other people cry can easily make you feel like you want to cry too'
(4) 'You often make a backup plan for a backup plan.'
(5) 'You usually stay calm, even under a lot of pressure'
(6) 'At social events, you rarely try to introduce yourself to new people and mostly talk to the ones you already know'
(7) 'You are very sentimental.'),
(8) 'You like to use organizing tools like schedules and lists.'
(9) 'Even a small mistake can cause you to doubt your overall abilities and knowledge.'
(10) 'You feel comfortable just walking up to someone you find interesting and striking up a conversation.'
(11) 'You are not too interested in discussing various interpretations and analyses of creative works.'
(12) 'You are more inclined to follow your head than your heart.'
(13) 'You usually prefer just doing what you feel like at any given moment instead of planning a particular daily routine.'
(14) 'You rarely worry about whether you make a good impression on people you meet.'
(15) 'You enjoy participating in group activities.'
(16) 'You like books and movies that make you come up with your own interpretation of the ending.'
(17) 'Your happiness comes more from helping others accomplish things than your own accomplishments.'
(18) 'You are interested in so many things that you find it difficult to choose what to try next.'
(19) 'You avoid leadership roles in group settings.'
(20) 'You are definitely not an artistic type of person.'
(21) 'You think the world would be a better place if people relied more on rationality and less on their feelings.'
(22) 'You enjoy watching people argue.'
(23) 'You tend to avoid drawing attention to yourself.'
(24) 'Your mood can change very quickly.'
(25) 'You lose patience with people who are not as efficient as you.'
(26) 'You often end up doing things at the last possible moment.'
(27) 'You have always been fascinated by the question of what, if anything, happens after death.'
(28) 'You become bored or lose interest when the discussion gets highly theoretical.'
(29) 'You find it easy to empathize with a person whose experiences are very different from yours.'
(30) 'You usually postpone finalizing decisions for as long as possible.'
(31) 'You rarely second-guess the choices that you have made.'
(32) 'After a long and exhausting week, a lively social event is just what you need.'
(33) 'You enjoy going to art museums.'
(34) 'You often have a hard time understanding other people's feelings.'
(35) 'You rarely feel insecure.'
(36) 'You avoid making phone calls.'
(37) 'You often spend a lot of time trying to understand views that are very different from your own.'
(38) 'In your social circle, you are often the one who contacts your friends and initiates activities.'
(39) 'If your plans are interrupted, your top priority is to get back on track as soon as possible.'
(40) 'You are still bothered by mistakes that you made a long time ago.'
(41) 'You rarely contemplate the reasons for human existence or the meaning of life.'
(42) 'Your emotions control you more than you control them.'
(43) 'You take great care not to make people look bad, even when it is completely their fault.'
(44) 'Your personal work style is closer to spontaneous bursts of energy than organized and consistent efforts.'
(45) 'When someone thinks highly of you, you wonder how long it will take them to feel disappointed in you.'
(46) 'You would love a job that requires you to work alone most of the time.'
(47) 'You believe that pondering abstract philosophical questions is a waste of time.'
(48) 'You feel more drawn to places with busy, bustling atmospheres than quiet, intimate places.'
(49) 'You know at first glance how someone is feeling.'
(50) 'You often feel overwhelmed.'
(51) 'You complete things methodically without skipping over any steps.'
(52) 'You are very intrigued by things labeled as controversial.'
(53) 'You would pass along a good opportunity if you thought someone else needed it more.'
(54) 'You struggle with deadlines.'
(55) 'You feel confident that things will work out for you.'

APÉNDICE B. TIPOS DE PERSONALIDADES

(1) ISTJ - Introverted sensor thinker judger.
(2) ISTP - Introverted sensor thinker perceiver.
(3) ISFP - Introverted sensor feeler perceiver.
(4) ISFJ - Introverted sensor feeler judger.
(5) INFP - Introverted intuitor feeler perceiver.
(6) INFJ - Introverted intuitor feeler judgger
(7) INTP - Introverted intuitor thinker perceiver.
(8) INTJ - Introverted intuitor thinker judger.
(9) ESFP - Extroverted sensor feeler perceiver.
(10) ESFJ - Extroverted sensor feeler judgger
(11) ESTP - Extroverted sensor thinker perceiver.
(12) ESTJ - Extroverted sensor thinker judger.
(13) ENFP - Extroverted intuitor feeler perceiver.
(14) ENFJ - Extroverted intuitor feeler judgger
(15) ENTP - Extroverted intuitor thinker perceiver.
(16) ENTJ - Extroverted intuitor thinker judger.