

UNIVERSIDAD RICARDO PALMA
ESCUELA DE POSGRADO

MAESTRIA EN INGENIERÍA INFORMÁTICA CON
MENCIÓN EN INGENIERÍA DE SOFTWARE



TESIS para optar el Grado Académico de Maestra en Ingeniería
Informática con mención en Ingeniería de Software

Detección de depresión a través de análisis textual utilizando
aprendizaje automático, 2017

Autor: Bach. Orué Medina, Ariana Maybee

Asesor: Mg. Pérez Godoy Ballón, Luis Alberto

LIMA-PERÚ

2018

AGRADECIMIENTO

El autor expresa su especial agradecimiento a Dios, por la bendición más grande de entregarle el amor de unos padres y hermana que han dado todo para que los acontecimientos más importantes de su vida, sucedan.

Al Mg. Luis Pérez, por su contribución como asesor y su confianza depositada en mí.

A los Mgs. Carlos García Quirós, Jose Espinoza, Olga Suarez, los Doctores Silverio Bustos y Zalatiel Carranza, por sus grandes y excelentes enseñanzas; por ser partícipe de mi desarrollo profesional para obtener el grado de maestro.

Al licenciado e investigador Ian Brossard, por su tiempo invaluable y enorme aporte en este trabajo con su dominio en la especialidad de psicología y conocimientos sobre depresión.

Al docente investigador Jorge Guerra, Responsable del Laboratorio de Robótica e Internet de las cosas de la UNMSM, por brindar su ardua experiencia y conocimientos sobre Aprendizaje Automático.

A la Directora de la Escuela de Ingeniería de Sistemas, de la Universidad César Vallejo, Lily Salazar, por apoyarme y permitirme aplicar los instrumentos para la recolección de datos de la presente investigación. Asimismo, a la Directora Académica de la Universidad César Vallejo - Callao, Pilar Jimenez y al Director de Bienestar Universitario, Edgard Rivero.

A todos los profesores de la Maestría en Ingeniería Informática con mención a ingeniería de software de la Universidad Ricardo Palma, compañeros, quienes de alguna manera me apoyaron y animaron a concluir exitosamente mi proyecto de tesis.

DEDICATORIA

A Dios por ser todopoderoso y dar bendiciones en mi vida.

*A mi madre Norma Medina Nonato, por ser mi compañera, amiga, ejemplo,
confidente y todo en mi vida.*

*A mi padre Julio Orué Solano, por ser mi motivo de admiración y propósito
para lograr lo que anhelo.*

*A mi hermana Shirley Orué Medina por brindarme su amor y apoyo en
cualquier momento.*

A mi hermana Grecia que está en el cielo, por ser mi angel de la guardia.

A Jordans Berenguel por darme la confianza necesaria y su apoyo.

RESUMEN

La presente tesis incluye la creación de un modelo predictivo, el análisis, diseño e implementación de una aplicación web usando análisis textual y aprendizaje automático para la detección precoz de la depresión, la cual es la principal causa de problemas de salud y discapacidad en todo el mundo.

El objetivo de la investigación es optimizar el tiempo de atención de forma oportuna a probables pacientes con depresión, identificar características específicas en individuos considerados con factores de riesgo de depresión y determinar los niveles de depresión a probables pacientes a través del análisis textual utilizando aprendizaje automático.

Se contó con una muestra perteneciente a 596 alumnos de la Universidad César Vallejo; a los cuales se le aplicó dos herramientas estandarizadas a nivel mundial “Inventario de Beck II” y “Test de Frases Incompletas Sacks” para hallar la depresión e identificar a las personas con el trastorno mental mencionado. Es preciso señalar que el tipo de estudio es aplicado y exploratorio.

Cabe resaltar que, para el procesamiento y tratamiento de datos, se utilizó la medida numérica del análisis textual TF-IDF, las técnicas de aprendizaje automático “Máquina de Vectores de Soporte”, el servicio de Microsoft Azure Machine Learning con R.

Finalmente, los resultados indicaron que el modelo predictivo usado para el desarrollo de un aplicativo móvil para detectar depresión tiene una exactitud del 99,2 %.

Palabras claves: Depresión, detección de la depresión, análisis textual, aprendizaje automático, máquina de vectores de soporte, TD-IDF.

ABSTRACT

This thesis includes the creation of a predictive model, the analysis, design and implementation of a web application using textual analysis and automatic learning for the early detection of depression, which is the main cause of health problems and disability throughout the world.

The objective of the research is to optimize the time of attention in a timely manner to probable patients with depression, identify specific characteristics in individuals considered with risk factors for depression and determine the levels of depression to probable patients through textual analysis using machine learning.

There was a sample belonging to 596 students of the César Vallejo University; to which two globally standardized tools "Beck II Inventory" and "Unfinished Sacks Phrases Test" were applied to find depression and identify people with the aforementioned mental disorder. It should be noted that the type of study is applied and exploratory.

It should be noted that, for the processing and processing of data, the numerical measurement of the TF-IDF textual analysis, the machine learning techniques "Support Vectors Machine", the Microsoft Azure Machine Learning service with R.

Finally, the results indicated that the predictive model used for the development of a mobile application to detect depression has an accuracy of 99.2%.

Keywords: Depression, depression detection, textual analysis, machine learning, support vector machine, TD-IDF.

ÍNDICE

AGRADECIMIENTO	2
RESUMEN	4
ABSTRACT	5
CAPÍTULO I: PLANTEAMIENTO DEL ESTUDIO	7
1.1 Introducción	7
1.2 Formulación del problema y justificación del estudio	8
1.3 Antecedentes relacionados con el tema	13
1.4 Objetivos generales y específicos	14
1.5 Limitaciones del estudio	14
CAPÍTULO II: MARCO TEÓRICO	15
2.1 Bases teóricas relacionadas con el tema	15
2.2 Definición de términos usados	31
2.3 Hipótesis	32
2.4 Variables	33
CAPÍTULO III : METODOLOGÍA DE INVESTIGACIÓN	35
3.1 Diseño de investigación	35
3.2 Población y muestra	36
3.3 Técnicas e instrumentos	36
3.4 Recolección de datos	37
CAPÍTULO IV : RESULTADOS Y ANÁLISIS DE RESULTADOS	38
4.1 Desarrollo del Modelo Predictivo	38
4.2 Resultados del análisis textual usando la técnica TF-IDF	39
4.3 Identificación de Patrones	60
4.4 Comparación entre Indicadores de desempeño para clasificadores de dos clases	63
4.5 Tiempo de atención de forma oportuna a probables pacientes con depresión	73
4.6 Eficacia de la detección de probables pacientes con depresión	74
4.7 Web Service para detectar la depresión	77
4.8 Aplicación Web ANTIDEP	77
4.9 Análisis de Resultados	80
CAPÍTULO V : CONCLUSIONES Y RECOMENDACIONES	82
Conclusiones	82
Recomendaciones	83
REFERENCIAS BIBLIOGRÁFICAS	84
ANEXOS	90

CAPÍTULO I: PLANTEAMIENTO DEL ESTUDIO

1.1 Introducción

Ministerio de Salud del Perú (2016) refiere que:

Entre 900 y 1000 muertes por suicidio pueden prevenirse con el apoyo oportuno de familiares o amigos. El primer paso es tener una actitud de ayuda y mostrar acercamiento a la persona que atraviese esta dificultad. Por ello, su detección oportuna será clave para poder evitar un daño autoinfligido y la muerte.

El tema de investigación es la detección de depresión a través del análisis textual utilizando aprendizaje automático, el fundamento para realizar esta investigación fue la problemática que existe en el sector de la salud que aqueja en contra de la vida de las personas y este es el trastorno mental más conocido como “la depresión”.

El objetivo principal de la presente tesis es optimizar el tiempo de atención de forma oportuna a probables pacientes con depresión, identificar características específicas en individuos considerados con factores de riesgo de depresión y determinar los niveles de depresión a probables pacientes a través del análisis textual utilizando aprendizaje automático.

Se contó con una muestra perteneciente a 596 alumnos de la Universidad César Vallejo; a los cuales se le aplicó dos herramientas estandarizadas a nivel mundial “Inventario de Beck II” y “Test de Frases Incompletas Sacks” para hallar la depresión e identificar a las personas con el trastorno mental mencionado. Es preciso señalar que el tipo de estudio es aplicado y exploratorio.

Los resultados indicaron que la exactitud del modelo predictivo que se usará para el desarrollo del web service para detectar la depresión fue del 97,6%; lo cual indica, que el uso del análisis textual y el aprendizaje automático permitirán la detección de la depresión. Igualmente, la precisión del modelo fue de 97,6 % lo que indica que el

modelo creado tiene un margen de error muy bajo a equivocarse. Por último, se indica que el Área Bajo la Curva (en inglés, AUC) el cual determinó la precisión del modelo predictivo y fue de 99,8 %, significando que la probabilidad de detectar a una persona con depresión es muy alta.

La presente investigación está estructurada por cinco capítulos: En el primer capítulo se plantea la problemática actual sobre la depresión, se establece el objetivo general y los objetivos específicos de la presente investigación; así como los antecedentes, la justificación y las limitaciones del estudio. En el segundo capítulo se define el marco teórico, especificando las bases teóricas relacionadas con el tema, la definición de términos usados, el planteamiento de las hipótesis y la identificación de las variables.

En el tercer capítulo se plantea el diseño de la investigación, se define la población y muestra, se menciona y describe las técnicas e instrumentos a usar, como se ha desarrollado la recolección de datos. En el cuarto capítulo se analizan e interpretan los resultados. Por último, en el quinto capítulo se realiza las conclusiones y recomendaciones.

1.2 Formulación del problema y justificación del estudio

Formulación del problema

La Organización Mundial de la Salud (2017) afirma que:

La depresión es la principal causa de problemas de salud y discapacidad en todo el mundo. Según las últimas estimaciones de la Organización Mundial de la Salud, más de 300 millones de personas viven con depresión, un incremento de más del 18% entre 2005 y 2015.

La realidad problemática se encuentra dentro del campo de la salud, donde a través de estudios se demuestra que el 60% y 80% de suicidios son causados por la depresión; ya que no existe una detección temprana o un tratamiento previo.

A continuación, en la Figura 1, se muestra a la depresión como un trastorno mental más común en el Perú.



Figura 1: La depresión es el problema de salud mental más común en el Perú.

Fuente: Instituto Nacional de Salud Mental (2016)

Según el Ministerio de Salud del Perú (2017) “entre 900 y 1000 muertes por suicidio pueden prevenirse con el apoyo oportuno de familiares o amigos. El primer paso es tener una actitud de ayuda y mostrar acercamiento a la persona que atraviese esta dificultad”.

Por ello, su detección oportuna será clave para poder evitar un daño autoinfligido y la muerte.

Según American Psychiatric Association (2017) informa que, para detectar si un paciente es depresivo tiene que pasar un periodo igual o superior a dos semanas hasta que se hagan prominentes las características de una persona depresiva.

La presente investigación aborda y busca apoyar en el campo de salud a detectar la depresión a través de un análisis textual utilizando aprendizaje automático. Es de suma importancia indicar que, a través de este modo de detección, se agilizará los tiempos y optimizarán el tiempo de atención de forma oportuna a probables pacientes con depresión.

La Organización Mundial de la Salud (2017), precisa sobre la depresión que por cuestiones culturales, tabú o falta de conocimientos, no se le suele prestar la atención y el tratamiento temprano para evitar su crecimiento, que hoy afecta al 5 por ciento de la población mundial. Se muestra a continuación el número de pacientes con problemas mentales según grado de instrucción de Lima.

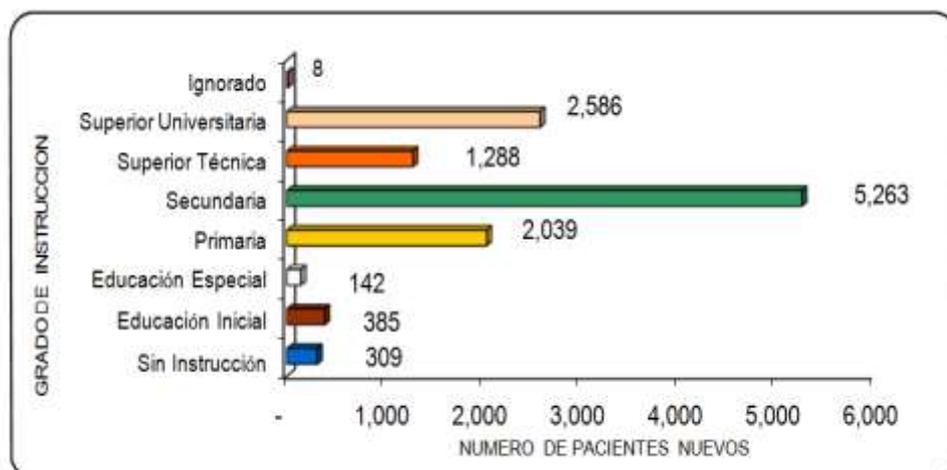


Figura 2: Pacientes nuevos con problemas mentales según grado de instrucción en Lima

Fuente: OEI - Organización de Estados Iberoamericanos (2016)

En la Figura 2, se observa el compendio estadístico anual del año 2016 de la Oficina de Estadística e Informática del Instituto Nacional de Salud Mental “Honorio Delgado – Hideyo Noguchi” donde indica que el mayor número de pacientes se concentra en jóvenes con grado de instrucción de secundaria y superior universitaria.

Por otro lado, la Tabla 1 muestra que los pacientes con problemas mentales más frecuentes de acuerdo al grado de instrucción pertenecen a Secundaria (5 263) de los cuales 2 652 varones y 2 611 mujeres. Le sigue el grado de instrucción Superior universitaria (2 586) con 1 159 varones y 762 mujeres; y en tercer lugar el grado de educación primaria (2 039) con 1 136 varones y 903 mujeres.

Tabla 1

Pacientes nuevos por grupos de edad y género según grado de instrucción en Lima, 2016.

Grado de Instrucción	Total	Género		Grupos de edad													
		M	F	0 - 4		5		6 - 14		15 - 19		20 - 44		45 - 64		65 +	
				M	F	M	F	M	F	M	F	M	F	M	F	M	F
Total	12 020	6 047	5 973	268	67	88	27	1 180	679	1 018	890	2 370	2 679	808	1 005	515	626
Sin Instrucción	309	152	157	80	24	4	-	14	2	6	3	25	16	8	22	15	90
Educación Inicial	385	303	82	183	43	78	26	39	13	-	-	2	-	-	-	1	-
Educación Especial	142	98	44	5	-	5	-	44	9	11	7	29	22	4	6	-	-
Primaria	2 039	1 136	903	-	-	1	1	699	254	30	21	99	145	77	155	221	327
Secundaria	5 263	2 652	2 611	-	-	-	-	383	400	777	641	1 053	980	268	451	171	139
Superior Técnica	1 288	542	746	-	-	-	-	-	-	48	49	387	528	78	155	29	14
Superior Universitaria	2 586	1 159	1 427	-	-	-	-	-	-	136	168	774	987	171	216	77	56
Ignorado	8	5	3	-	-	-	-	-	1	1	1	1	1	2	-	1	-

Fuente: Instituto Nacional de Estadística e Informática (2016).

Nota. El total de pacientes nuevos por grupo de edad y género es de 12 020.

En los últimos años existe mucha bibliografía donde se ha tratado de diagnosticar depresión a través del texto:

Cabrera, Frías, Hernández, Iniestra y Solis (2014) realizan una investigación titulada “Escrutinio de depresión y ansiedad en twitter a través de un programa de análisis de programa” en la Universidad Autónoma Metropolitana, cuyo objetivo era conocer si es posible utilizar estas publicaciones como elementos para detectar depresión o ansiedad, y comparar la congruencia entre la interpretación del software y las pruebas in vivo. Sin embargo, sus resultados indicaron que Twitter no muestra ser indicador de diagnóstico de probables trastornos depresivos o ansiosos.

Volkova, Bachrach y Van Durme (2017) realizan una investigación titulada “Mining User Interests to Predict Perceived Psycho-Demographic Traits on Twitter”, con la finalidad conocer si es posible utilizar estas publicaciones como elementos para detectar depresión o ansiedad, y comparar la congruencia entre la interpretación del software y las pruebas realizadas a los usuarios a través de inventarios de diagnóstico estandarizados. Sin embargo, sus resultados indicaron que Twitter no muestra ser indicador para detectar probables trastornos depresivos o ansiosos.

Los estudios anteriormente mencionados demuestran que los usuarios publican contenidos de manera muy esporádica y eso dificulta la predicción de la depresión a través del análisis textual.

El problema descrito puede formularse de manera general bajo la siguiente interrogante:

Problema General

-----¿Cómo detectar la depresión?

Problemas Específicos

- a) ¿Cómo identificar nuevos patrones en individuos considerados con factores de riesgo de depresión?
- b) ¿De qué manera identificar de forma eficaz a probables pacientes con depresión?
- c) ¿Cómo atender de forma oportuna a probables pacientes con depresión?

Justificación del estudio

Justificación Social

El estudio de este sistema causará un impacto social; ya que está relacionado a la detección oportuna de depresión en las personas. De esta manera apoya al campo de la salud y a través del modelo predictivo y uso del web service tendría más posibilidades de aportar a la disminución de números de suicidios o problemas de trastornos mentales en el Perú.

Justificación Tecnológica

A través de este proyecto se pretende desarrollar una aplicación web basada en un modelo predictivo que concentra la información en un web service usando técnicas de Aprendizaje Automático y Análisis Textual; que permitan la detección de la depresión.

Justificación Económica

La presente investigación servirá para la elaboración de una aplicación web que permita detectar la depresión de forma oportuna generando un gran impacto social y pueda ser comercializado y adquirido por cualquier persona, entidad privada o pública a través de un costo aceptable. Además, puede ser punto de partida para la creación de productos muchos más avanzados que puedan ser utilizados en el campo de salud.

Justificación Académica

La aplicación de las técnicas de aprendizaje automático y el análisis textual permitirán la detección de depresión para posibles pacientes. Además, permitirá comprobar que el uso de herramientas estandarizadas como el Inventario de Depresión de Beck II y el Test de Frases Incompletas Sack; serán gran aporte para la construcción de una aplicación web.

1.3 Antecedentes relacionados con el tema

Cabrera, Frías, Hernández, Iniestra y Solis (2014) realizan una investigación titulada “Escrutinio de depresión y ansiedad en twitter a través de un programa de análisis de programa” en la Universidad Autónoma Metropolitana, cuyo objetivo era conocer si es posible utilizar estas publicaciones como elementos para detectar depresión o ansiedad, y comparar la congruencia entre la interpretación del software y las pruebas in vivo. Sin embargo, sus resultados indicaron que Twitter no muestra ser indicador de diagnóstico de probables trastornos depresivos o ansiosos.

Este artículo publicado me ayudó a determinar los objetivos y la importancia de la aplicación y uso de las tecnologías de minería de datos en la sociedad.

Volkova, Bachrach y Van Durme (2017) realizan una investigación titulada “Mining User Interests to Predict Perceived Psycho-Demographic Traits on Twitter”, con la finalidad conocer si es posible utilizar estas publicaciones como elementos para detectar depresión o ansiedad, y comparar la congruencia entre la interpretación del software y las pruebas en vivo. Sin embargo, sus resultados indicaron que Twitter no muestra ser indicador para detectar probables trastornos depresivos o ansiosos.

A través de esta investigación se determinó en utilizar otras herramientas para poder evaluar depresión a través del análisis textual.

1.4 Objetivos generales y específicos

Objetivo general

Detectar la depresión.

Objetivos específicos

- a) Identificar nuevos patrones en individuos considerados con factores de riesgo de depresión.
- b) Identificar de forma eficaz a probables pacientes con depresión.
- c) Permitir atender de forma oportuna a probables pacientes con depresión.

1.5 Limitaciones del estudio

1. La presente investigación sobre la detección de depresión a través del análisis textual utilizando aprendizaje automático no reemplaza a profesionales de salud.
2. Los estudiantes que participaron como estudio de muestra pertenecen a la Universidad César Vallejo al ser una de las empresas educativas más grandes del Perú. Cabe resaltar que estuvieron de acuerdo en su totalidad en brindarnos la información necesaria.
3. El web service podrá ser utilizado para cualquier persona que desea tener una detección precoz sobre la depresión.
4. El web service será consumido en una aplicación web utilizando el lenguaje en php.

CAPÍTULO II: MARCO TEÓRICO

2.1 Bases teóricas relacionadas con el tema

Estado del Arte

Según Google Trends (2016) afirma:

El interés de búsqueda en línea a lo largo del tiempo en "Ciencia de datos" es similar al interés en "análisis de datos", pero es 50% -100% menos que el interés en "big data". Sin embargo, el interés histórico de búsqueda en la ciencia de la información y la analítica es aproximadamente el doble del interés mostrado en big data hace unos 10 años.

Según Graham (2012), "El arte de la ciencia de datos ha atraído un interés cada vez mayor desde un amplio rango de dominios y disciplinas".

Los estadísticos han tenido mucho que decir sobre la ciencia de datos, ya que son ellos quienes realmente crearon el término "ciencia de datos" y promovió la actualización de las estadísticas a la ciencia de datos como una disciplina más amplia. Esto se refleja en una serie de acciones anteriores, como las siguientes:

- Según Wu (1997) afirmó: "Estadísticas = ¿Ciencia de datos? y sugirió que las estadísticas deberían cambiarse por el nombre de ciencia de datos y que los estadísticos deberían ser conocidos como científicos de datos".

- Cleveland (2001) sugiere que sería apropiado modificar el campo de las estadísticas a la ciencia de datos y ampliar las áreas principales del trabajo técnico del campo de las estadísticas.

- Breiman (2001) afirma que era necesario alejarse de las exclusivas dependencias en los modelos de datos (en estadística) y adoptar un conjunto más diverso de herramientas "tales como modelado algorítmico, que trata el mecanismo de datos como desconocido"

- Según Smith (2006) "Se han llevado a cabo intensas discusiones dentro de la comunidad académica y de investigación sobre la creación de la ciencia de datos como disciplina académica".

- “Tratando a la ciencia de datos como una mezcla de estadísticas, matemáticas, informática, gráfica, diseño, minería de datos, interacción hombre-computadora y visualización de información” (Yau, 2009).

- “Una declaración sobre el papel de las estadísticas en la ciencia de los datos fue lanzada por un número de líderes ASA, diciendo que las estadísticas y la máquina de aprendizaje juega un papel central en la ciencia de la información” (Van et al., 2015).

- Muchos estados y ciudades de China han lanzado estrategias nacionales de big data y planes de acción para big data y computación en la nube (CBDIO, 2016).

- Análisis de redes sociales: en Desale [2015], se enumeraron 30 herramientas para el SNA y la visualización, tales como Centrifuge, Commetrix, Sepia, Cytoscape, EgoNet, InFlow, JUNG, Keynetiq, NetMiner, Network Workbench, NodeXL y SocNetV (Redes sociales, Visualizador).

- Otras herramientas: se han desarrollado un número creciente de herramientas y están en desarrollo para la ciencia de datos específica del dominio y el problema, como Alteryx y Tableau para tabletas; SuggestGrid y Mortero Motor de recomendación para recomendar sistemas (Github,2016).

Mediante el estado del arte, se logra apreciar que el análisis textual en la actualidad; se está volviendo una tecnología potencial que permite la identificación y descubrimiento de nuevos patrones que nos puedan aportar información valiosa en la toma de futuras decisiones.

Aprendizaje automático (Machine Learning)

Según Kubat (1998) “La capacidad de que un sistema informático aprenda de manera automática es fundamental en Inteligencia Artificial, pues para que un sistema pueda ser considerado inteligente debe ser al menos capaz de aprender automáticamente”.

Según Murphy (2012) afirma que “El aprendizaje automático es un campo de las Ciencias de la Computación que permite las computadoras aprender, sin tener que ser explícitamente programadas”.

Aprendizaje supervisado

Según Murphy (2012) afirma:

Los sistemas de aprendizaje automático parten de conocimiento previo (conjunto de entrenamiento) para después poder predecir o asignar una etiqueta a casos nuevos. Es decir, el sistema de clasificación aprenderá de experiencias previas para poder después asignar una etiqueta a los sucesivos casos a analizar. Nótese que estos últimos no pueden haber sido analizados previamente por el clasificador. Por ello, en un sistema de clasificación se pueden diferenciar dos fases. (p.40)

La primera de ellas se conoce como la etapa de aprendizaje; esta es la encargada de aprender de los patrones o características de cada instancia.

La segunda fase, conocida como test, es aquella en la cual el clasificador es puesto a prueba haciendo que clasifique otro conjunto de instancias de las cuales se conoce a priori la etiqueta correcta.

De esta forma, se podrá conocer la calidad o el comportamiento del sistema.

Máquinas de Vectores de Soporte (Support Vector Machine)

Cristianini y Taylor (2004) afirman que:

Es una técnica de aprendizaje automático que permite realizar predicciones en base a información suministrada como ejemplos. Este modelo, separa las clases a predecir en dos espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector. La manera más simple de realizar la separación es mediante una línea recta, un plano recto o un hiperplano de n dimensiones, sin embargo, los datos a estudiar no siempre pueden ser separados en la dimensión a la que pertenecen. La técnica de Support Vector Machine ofrece una solución a este problema, proyectando la información a un espacio de características de mayor dimensión, aumentando la capacidad de realizar una separación lineal. (p. 146)

Averaged Perceptron (Perceptrón promediado)

Microsoft (2018) afirma:

Este algoritmo de clasificación es un método de aprendizaje supervisado y requiere un conjunto de datos etiquetados, que incluye una columna de etiqueta. Puede entrenar el modelo proporcionando el modelo y el conjunto de datos etiquetados como una entrada al modelo de tren o a los hiperparámetros del modelo de sintonización.

$$\hat{y} = \text{sign} \left(\left(\sum_{k=1}^K c^{(k)} w^{(k)} \right) \cdot \hat{x} + \sum_{k=1}^K c^{(k)} b^{(k)} \right)$$

Figura 3: Fórmula Averaged Perceptron

Fuente: Murphy (2012)

Bayes Point Machine (Máquina de punto bayes)

Microsoft (2018) afirma:

El algoritmo en este módulo utiliza un enfoque bayesiano para la clasificación lineal llamada "máquina de punto Bayes". Este algoritmo se aproxima eficientemente al promedio bayesiano teóricamente óptimo de clasificadores lineales (en términos de rendimiento de generalización) al elegir un clasificador "promedio", el punto de Bayes. Debido a que Bayes Point Machine es un modelo de clasificación bayesiano, no es propenso a sobreajustar los datos de entrenamiento.

Tree Regression (Arboles de regresión)

Según Microsoft (2018) "Este método de regresión es un método de aprendizaje supervisado y, por lo tanto, requiere un conjunto de datos etiquetados. La columna de etiqueta debe contener valores numéricos".

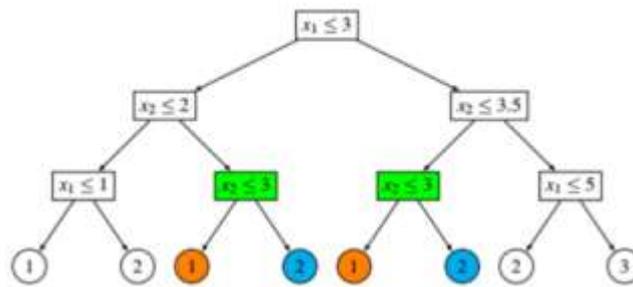


Figura 4: Tree Regression

Fuente: Murphy (2012)

Decision Forest (Bosque de Decisión)

Microsoft (2018) afirma:

Este algoritmo de bosque de decisión es un método de aprendizaje conjunto destinado a tareas de clasificación. Los métodos de Ensemble se basan en el principio general de que, en lugar de confiar en un único modelo, puede obtener mejores resultados y un modelo más general creando múltiples modelos relacionados y combinándolos de alguna manera. En general, los modelos de conjunto proporcionan una mejor cobertura y precisión que los árboles de decisión única. (p.1)

Decision Jungle (Junglas de Decisión)

Microsoft (2018) afirma:

Las junglas de decisión son una extensión reciente de los bosques de decisión. Una jungla de decisión consiste en un conjunto de gráficos acíclicos dirigidos por decisión (DAG). Las junglas de decisión tienen las siguientes ventajas. Al permitir que las ramas de los árboles se fusionen, un DAG de decisión suele tener una huella de memoria menor y un mejor rendimiento de generalización que un árbol de decisiones, aunque a costa de un tiempo de capacitación un poco más largo.

- Las junglas de decisión son modelos no paramétricos que pueden representar límites de decisión no lineales.
- Realizan la selección y clasificación de características integradas y son resistentes ante la presencia de características ruidosas.

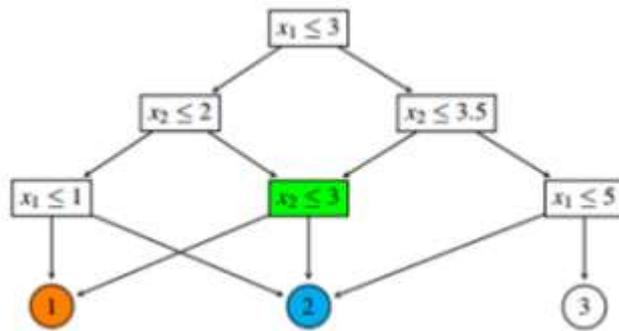


Figura 5: Comportamiento de Junglas de decisión

Fuente: Murphy (2012)

Logistic Regression (Regresión logística)

Microsoft (2018) afirma:

La regresión logística es un método bien conocido en las estadísticas que se utiliza para predecir la probabilidad de un resultado, y es especialmente popular para las tareas de clasificación. El algoritmo predice la probabilidad de ocurrencia de un evento ajustando los datos a una función logística.

Neural Network (Red neuronal)

Microsoft (2018) afirma:

Una red neuronal es un conjunto de capas interconectadas. Las entradas son la primera capa y están conectadas a una capa de salida mediante un gráfico acíclico compuesto por bordes y nodos pesados. Entre las capas de entrada y salida puede insertar múltiples capas ocultas. La mayoría de las tareas predictivas se pueden realizar fácilmente con solo una o algunas capas ocultas. Sin embargo, investigaciones recientes han demostrado que las redes neuronales profundas (DNN) con muchas capas pueden ser muy efectivas en tareas complejas como reconocimiento de imágenes o voz.

Microsoft (2018) afirma:

Las capas sucesivas se usan para modelar niveles crecientes de profundidad semántica. La relación entre las entradas y salidas se aprende del entrenamiento de la red neuronal en los datos de entrada. La dirección del gráfico procede de las entradas a través de la capa oculta y a la capa de salida. Todos los nodos en una capa están conectados por los bordes pesados a los nodos en la siguiente capa.

Análisis Textual

Weiss et al. (2005) afirma:

Corresponde a técnicas automáticas para analizar grandes volúmenes de información textual. Las aproximaciones más básicas, corresponden al conteo absoluto de veces que aparece un término (llamado también token) o a ponderaciones logarítmicas de estas frecuencias (idf). Típicamente, se utiliza el enfoque de bag of words (bolsa de palabras), en el que no se toma en cuenta el orden en el que aparecen las palabras, sino el conjunto de palabras que aparecen). (p.98)

Según Según Vuotto, Bogetti, Fernández (2015) afirmaron que:

Las principales actividades que se realizan en la etapa de tratamiento previo para el análisis textual de los documentos son:

Análisis léxico

Consiste en la transformación de un documento en palabras o autoridades para describir el contenido del texto.

El trabajo léxico plantea variadas disyuntivas sobre cuáles términos pueden ser descriptores o no del documento, para lo cual deben tomarse decisiones prestando principal atención en la frecuencia de los términos, los pesos de estos en el conjunto del texto y de la colección, y finalmente cómo se resuelve el tratamiento de los números, y caracteres especiales. Ya que en muchos casos pueden ser de gran utilidad para describir un contenido y en otros pueden presentar mucha vaguedad al respecto.

Eliminación de palabras vacías

Las palabras vacías son inútiles en el momento de la recuperación y no describen semánticamente un texto; usualmente conformadas por el grupo de los artículos, preposiciones, y conjunciones; las cuales por lo general mantienen una presencia de un 80% en el total de palabras de la colección. Eventualmente también pueden incluirse otros términos. Su eliminación reduce considerablemente el peso de los textos a almacenar y agiliza el trabajo de los motores de almacenamiento y de búsqueda.

La eliminación de palabras vacías, al igual que el trabajo léxico, debe realizarse con cuidado para evitar minimizar la posibilidad de acierto discriminando términos significativos.

Procesos de stemming (reconocimiento de stems)

Consiste en el proceso de eliminación automática de partes no esenciales de los términos (sufijos, prefijos) para reducirlos a su parte esencial (lema) y ayudar a la correcta indización y recuperación. Existen variadas teorías y algoritmos, las técnicas más identificativas son la lematización flexiva (que elimina plurales, género y terminaciones verbales) frente a lematización derivativa (que elimina además sufijos derivativos).

La aplicación de técnicas de stemming puede resultar contradictoria a los beneficios de la recuperación cuando no se realiza correctamente, ya que debe ir acompañada con un trabajo terminológico que ofrezca al sistema el listado de lemas y de implementación de algoritmos efectivos, como es el caso del aportado por Porter para eliminación de sufijos que con simplicidad arroja resultados comparables a aquellos algoritmos de mayor complejidad.

Detección de grupos nominales

Son conjuntos de términos que se agrupan en función de un sustantivo que cumple el rol de núcleo en el documento. Reducir esta tarea sólo a la selección de sustantivos puede ser fundamental para la elección de los términos de indexación y simplificar los índices del sistema, principalmente cuando se trabaja a texto completo.

Construcción de tesauros

El tesoro refiere a un conjunto seleccionado de términos de peso semántico en un área específica del saber, estableciendo relaciones terminológicas entre ellos. De esta forma al buscar un término en el tesoro se puede conocer de forma inmediata qué relaciones mantiene este con otros términos pertenecientes al mismo tópico.

Esta herramienta no sólo normaliza el vocabulario al tratar su sinonimia, sino que permite identificar niveles jerárquicos dentro de un mismo tema y moverse en esa matriz en el proceso de descripción semántica y de búsqueda de información.

TF-IDF (Term frequency – Inverse document frequency)

Robertson (2014) afirma que:

Es una medida numérica que expresa cuán relevante es una palabra con respecto a un grupo de palabras dentro de una colección. Esta medida se utiliza como factor de ponderación para castigar a las palabras que son escritas por más de un segmento (dificultan la separación de estos) y premia a aquellas palabras que son seguidas en un segmento y no en los demás (esto facilita la separación de segmentos). El valor TF-IDF aumenta proporcionalmente al número de veces que una palabra es mencionada, pero es compensada por la frecuencia con la que esta palabra es escrita en el total de usuarios. (p.95)

Depresión

La depresión es una de las alteraciones del estado de ánimo más frecuentes (Cabrera et al., 2014). Consiste en la reducción de la capacidad de disfrutar, sentir interés, concentrar la atención, y en un aumento del cansancio, algunas veces es acompañada de sentimientos de culpabilidad o desesperanza y de pensamientos de muerte. Según Holleran (2010) afirma que “la depresión tiene alta prevalencia en adultos jóvenes”. Su correcto diagnóstico y tratamiento es un reto para el sistema de salud (Mitchell AJ, 2009).

Detección de la Depresión

Según el Instituto Nacional de Salud Mental (2001), “La depresión es difícil de detectar, por lo que es conocida como la enfermedad invisible, y muchas veces no es detectada a tiempo”. Estudios anteriores sugieren que la depresión puede ser detectada a través de cuestionarios (Houston et al, 2001), el texto que escriben las personas (Neuman et al., 2010), tales como los textos disponibles en redes sociales, fotos de ayuda, entre otros.

Según American Psychiatric Association (2017), para cumplir los criterios de depresión, el paciente debe tener, como mínimo, cinco de los siguientes síntomas durante un periodo igual o superior a dos semanas. Al menos uno de dichos síntomas debe consistir en estado de ánimo deprimido la mayor parte del día casi todos los días. Los síntomas a vigilar son:

- Estado de ánimo deprimido
- Disminución notable del interés o del placer por las actividades
- Pérdida importante de peso y/o disminución del apetito
- Insomnio o hipersomnia
- Retraso psicomotor o agitación
- Sentimiento de falta de valor o culpa
- Disminución de la capacidad para concentrarse
- Pensamientos recurrentes sobre la muerte o el suicidio.

Inventario de Depresión de Beck II

Según Otiniano e Irarrázabal (2012) afirman que:

El constructor de Depresión puede ser medido a través del Inventario de Depresión de BDI-II el cual consiste en 21 ítems puntuados en una escala Likert de 4 puntos donde mayor será la severidad de la depresión cuando el puntaje sea más alto. Este instrumento ha sido validado en diversos países de habla hispana, tales como España, Argentina, Estados Unidos con población latina, Costa Rica, Chile, entre otros. (Se muestra en el Anexo 1)

Frases Incompletas Sacks

El test de Frases Incompletas de Sacks es una prueba proyectiva verbal, de diagnóstico en diversas áreas de relaciones sociales (Maglio, et al, 2010). Consiste en una hoja de 60 frases incompletas distribuidas en 15 que aluden a diversos ámbitos de la vida de una persona. Las personas que llenan este test, deben hacerlo completando las frases con lo primero que se les ocurra. Se busca que, de esta manera, proyecten sus ideas, valores, creencias, anhelos, fantasías, temores, etc. En el anexo 2 se observa el test de frases incompletas que se construyó.

Librería tm-r

Según Feinerer (2008) afirma:

La estructura principal para la gestión de documentos en tm es una llamada Cuerpo, que representa una colección de documentos de texto. Un corpus es un concepto abstracto, y no pueden existir varias implementaciones en paralelo. La implementación por defecto es el llamado VCorpus (corto para Volátil Corpus) que da cuenta de una semántica como se conoce de la mayoría R Objetos: corpus son R objetos contenidos totalmente en la memoria.

Denotamos esta tan volátil ya que una vez el R objeto es destruido, todo el corpus se ha ido. Tal corpus volátil puede ser creado a través del constructor VCorpus (x, readerControl). Otra aplicación es la PCorpus que implementa una Corpus permanente la semántica, es decir, los documentos se almacenan físicamente fuera de R (por ejemplo, en una base de datos), que corresponde R los objetos son básicamente sólo los punteros a estructuras externas, y los cambios en los corpus subyacentes se reflejan a todos R objetos asociados con ella.

En comparación con el corpus volátil el corpus encapsulado por un objeto corpus permanente no se destruye si el correspondiente R objeto se libera. Es por eso, que para el análisis de los datos obtenidos en las encuestas trabajamos con la librería tm-r. (p.102)

Python

Van (2009) afirma:

Python es un lenguaje interpretado, lo cual puede ahorrarte mucho tiempo durante el desarrollo ya que no es necesario compilar ni enlazar. El intérprete puede usarse interactivamente, lo que facilita experimentar con características del lenguaje, escribir programas descartables, o probar funciones cuando se hace desarrollo de programas de abajo hacia arriba. Es también una calculadora de escritorio práctica. Python permite escribir programas compactos y legibles. Los programas en Python son típicamente más cortos que sus programas equivalentes en C, C++ o Java por varios motivos

- Los tipos de datos de alto nivel permiten expresar operaciones complejas en una sola instrucción.
- La agrupación de instrucciones se hace por sangría en vez de llaves de apertura y cierre.

Curva ROC (Receiver-operating characteristics)

Ruiz de Adana (2009) afirma:

Sus siglas en español significan la curva Característica Operativa del Receptor. Las curvas ROC son índices de la exactitud diagnóstica y proporcionan un criterio unificador en el proceso de evaluación de una prueba, debido a sus diversas aplicaciones. El uso de las curvas ROC en la evaluación de pruebas diagnósticas presenta las siguientes ventajas.

- Son una representación fácilmente comprensible de la capacidad de discriminación de la prueba en todo el rango de puntos de corte.
- Son simples, gráficas y fáciles de interpretar visualmente.
- No requieren un nivel de decisión particular porque está incluido todo el espectro de puntos de corte.
- Son independientes de la prevalencia, ya que la sensibilidad y la especificidad se obtienen en distintos subgrupos. (p.199)

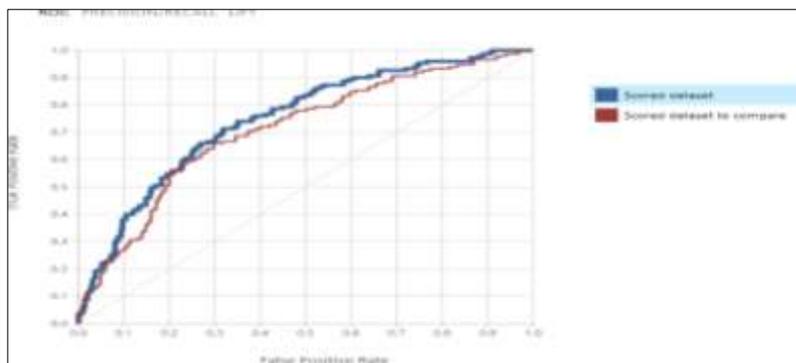


Figura 6: Curva ROC en Azure Machine Learning Studio

Fuente: Microsoft (2018)

Azure Machine Learning Studio

Microsoft (2018) afirma:

Es una herramienta colaborativa de arrastrar y soltar que puede usar para crear, probar y desplegar soluciones de análisis predictivo en sus datos. Machine Learning Studio publica modelos como servicios web que pueden ser consumidos fácilmente por aplicaciones personalizadas o herramientas de BI como Excel.

Machine Learning Studio es donde se reúnen la ciencia de datos, el análisis predictivo, los recursos de la nube y sus datos, incluye cientos de paquetes integrados y soporte para código personalizado.

Las etiquetas de clase en el conjunto de entrenamiento pueden tomar solo 2 valores posibles, a los que generalmente nos referimos como positivos o negativos. Las instancias positivas y negativas que un clasificador predice correctamente se llaman verdaderos positivos (TP) y verdaderos negativos (TN), respectivamente. Del mismo modo, las instancias incorrectamente clasificadas se llaman falsos positivos (FP) y falsos negativos (FN).

La matriz de confusión es simplemente una tabla que muestra el número de instancias que corresponden a cada una de estas 4 categorías. Azure Machine Learning decide automáticamente cuál de las dos clases en el conjunto de datos es la clase positiva.

Si las etiquetas de clase son booleanas o enteros, las instancias etiquetadas como "verdadero" o "1" tienen asignada la clase positiva. Si las etiquetas son cadenas, como en el caso del conjunto de datos de ingresos, las etiquetas se ordenan alfabéticamente y el primer nivel se elige como la clase negativa, mientras que el segundo nivel es la clase positiva. (p. 1)

Tabla 2

Matriz de confusión de clasificación binaria

	Predicted	
	Positive	Negative
Actual True	TP	FN
Actual False	FP	TN

Fuente: Microsoft (2018)

Nota. Las etiquetas de la matriz de confusión son booleanas.

El proceso de aprendizaje automático

Ya sea que una organización use Azure ML u otro enfoque, el proceso básico de aprendizaje automático es muy similar. La Figura 7 muestra cómo se ve normalmente.

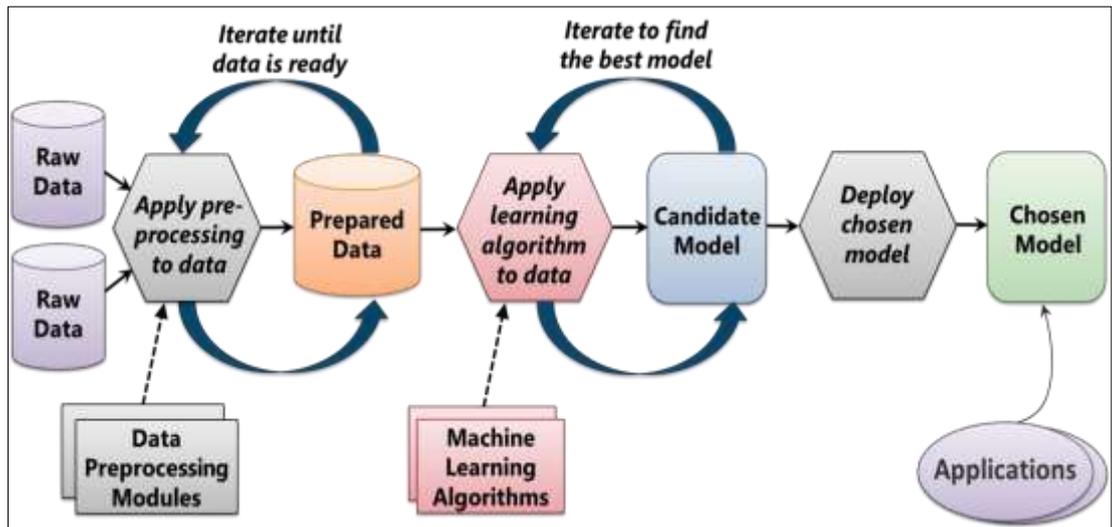


Figura 7: El proceso de aprendizaje automático comienza con datos brutos y termina con un modelo derivado de esos datos.

Fuente: Microsoft (2018)

Como se muestra en la figura 7, el aprendizaje automático comienza con los datos: cuanto más tenga, mejores serán sus resultados. Debido a que vivimos en la era de los grandes datos, el aprendizaje automático se ha vuelto mucho más popular en los últimos años. Tener muchos datos para trabajar en muchas áreas diferentes permite que las técnicas de aprendizaje automático se apliquen a un conjunto más amplio de problemas (Microsoft, 2018).

Dado que el proceso comienza con datos, la elección de los datos correctos para trabajar es de suma importancia.

Microsoft (2018) afirma:

El objetivo del preprocesamiento de datos es crear lo que se llama datos preparados. Pero llegar aquí generalmente no es simple. Como se muestra en la Figura 7, es un proceso iterativo, con varios módulos de preprocesamiento de datos diferentes aplicados a los datos brutos. Una vez que un equipo de aprendizaje automático tiene los datos correctos preparados, puede pasar al siguiente paso.

Para hacer esto, el equipo usa algoritmos de aprendizaje automático para trabajar con los datos preparados. El científico de datos del equipo elige un algoritmo de aprendizaje automático, decide qué aspectos de los datos preparados se deben usar y luego examina el resultado. El objetivo es determinar qué combinación de algoritmo de aprendizaje automático y datos preparados genera los resultados más útiles.

Modelo Predictivo para consumir datos en Web

Azure ML proporciona varios componentes diferentes para realizar un modelo predictivo y poder consumir sus datos mediante un web service.

La Figura 8 muestra cuáles son esos componentes y dónde encajan en el proceso de aprendizaje automático.

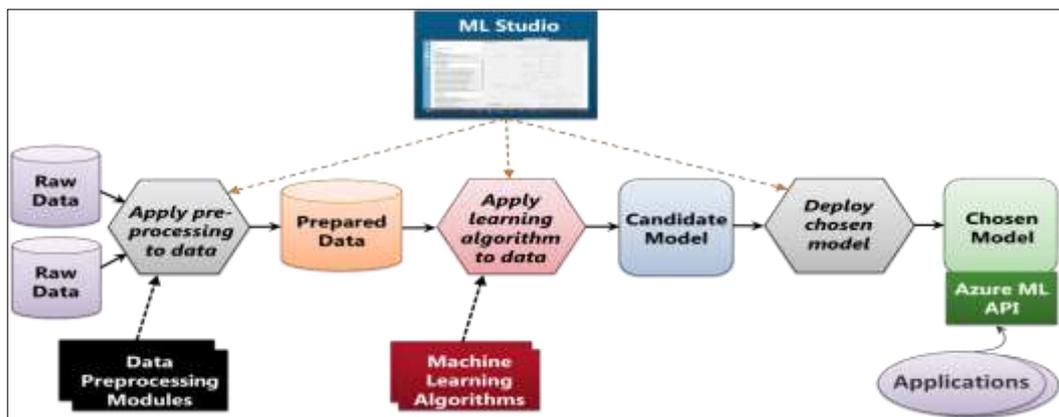


Figura 8: Azure ML proporciona una herramienta gráfica para administrar el proceso de aprendizaje automático, un conjunto de módulos de preprocesamiento de datos, un conjunto de algoritmos de aprendizaje automático y una API para exponer un modelo a las aplicaciones.

Fuente: Microsoft (2018)

Microsoft (2018) afirma que los componentes que proporciona Azure ML son los siguientes:

Machine Learning Studio, una herramienta gráfica que se puede usar para controlar el proceso de principio a fin. Con esta herramienta, las personas en el equipo de aprendizaje automático pueden aplicar módulos de preprocesamiento de datos a datos brutos, ejecutar experimentos sobre los datos preparados mediante un algoritmo de aprendizaje automático y probar el modelo resultante. Una vez que se encuentra un modelo efectivo, ML Studio también ayuda a sus usuarios a implementar ese modelo en Microsoft Azure.

- ✓ Un conjunto de módulos de preprocesamiento de datos.
- ✓ Un conjunto de algoritmos de aprendizaje automático.
- ✓ Una API de Azure ML que permite a las aplicaciones acceder al modelo elegido.

Aplicación Web

Adobe Systems Incorporated (2018) afirma:

Una aplicación Web es un sitio Web que contiene páginas con contenido sin determinar, parcialmente o en su totalidad. El contenido final de una página se determina sólo cuando el usuario solicita una página del servidor Web. Dado que el contenido final de la página varía de una petición a otra en función de las acciones del visitante, este tipo de página se denomina página dinámica.

Las aplicaciones Web se crean en respuesta a diversas necesidades o problemas. En esta sección se describen los usos más habituales de las aplicaciones Web y se proporciona un ejemplo sencillo.

2.2 Definición de términos usados

Análisis textual: Es un análisis matemático para deducir patrones y tendencias que existen en los datos, patrones que no pueden detectarse mediante una exploración tradicional de los datos porque las relaciones son demasiado complejas o por el volumen de datos que se maneja.

Aprendizaje automático: Es un tipo de inteligencia artificial (AI) que proporciona a las computadoras la capacidad de aprender, sin ser programadas explícitamente

Depresión: La depresión es un trastorno afectivo con síntomas psíquicos y somáticos, y para su diagnóstico y tratamiento es clave valorar el entorno del paciente.

Detección de la Depresión: Consiste en localizar o descubrir posibles características que determinen un grado depresivo en personas; por lo que nos ayudará a reconocer fácilmente esta enfermedad para ponerle remedio a tiempo y comenzar cuanto antes a combatirla.

2.3 Hipótesis

General

El uso del análisis textual y el aprendizaje automático en una aplicación web ayudará a detectar la depresión.

Específicas

- a) El uso del análisis textual y el aprendizaje automático en un modelo predictivo permitirá identificar nuevos patrones en individuos considerados con factores de riesgo de depresión.
- b) El uso del análisis textual y el aprendizaje automático en una aplicación web identificará de forma eficaz a probables pacientes con depresión.
- c) El uso del análisis textual y el aprendizaje automático en una aplicación web permitirá atender de forma oportuna a probables pacientes con depresión.

2.4 Variables

La variable independiente tiene dos valores el uso y no uso del análisis textual y el aprendizaje automático en una aplicación web. Mientras la variable dependiente es la detección de la depresión.

Para la detección precoz de la depresión se encontraron tres indicadores que determinarán la mejora de este proceso con el uso del análisis textual y el aprendizaje automático:

- Tiempo de atención de forma oportuna a probables pacientes con depresión
- Número de patrones identificados.
- Número de pacientes detectados

Definición conceptual

Análisis textual: Es un análisis matemático para deducir patrones y tendencias que existen en los datos.

Aprendizaje automático: Es un tipo de inteligencia artificial (AI) que proporciona a las computadoras la capacidad de aprender, sin ser programadas explícitamente.

Detección de la depresión. Consiste en localizar o descubrir posibles características que determinen un grado depresivo en personas.

Depresión: La depresión es un trastorno afectivo con síntomas psíquicos y somáticos, y para su diagnóstico y tratamiento es clave valorar el entorno del paciente.

Operacionalización de las variables

Variables	Indicadores
X= Análisis textual y aprendizaje automático en una aplicación web	
Y= Detección de la depresión	<ul style="list-style-type: none">- Tiempo de atención de forma oportuna a probables pacientes con depresión- Número de patrones identificados.- Número de pacientes detectados

CAPÍTULO III : METODOLOGÍA DE INVESTIGACIÓN

3.1 Diseño de investigación

Para Eyssautier (2006), “el método de investigación aplicada es aquel que utiliza los conocimientos obtenidos en las investigaciones en la práctica, y con ello traer beneficios a la sociedad” (p. 225); y mediante esta investigación se busca aplicar conocimientos de aprendizaje automático y análisis textual de tal manera que traiga consigo grandes beneficios para el campo de la salud. Según el objeto de estudio, la presente investigación es aplicada.

Por otro lado, la investigación es de tipo tecnológico porque se busca aplicar un modelo predictivo para centralizar los datos en un aplicativo y de esa forma se pueda corregir la situación actual que ha dado origen al estudio de la investigación. Carmona, Y. (2011), “expresa que una investigación tecnológica es una actividad que, a través de la aplicación del método científico, está encaminada a descubrir nuevos conocimientos (investigación básica), a la que posteriormente se le buscan aplicaciones prácticas (investigación aplicada) para el diseño o mejoramiento de un producto, proceso industrial o maquinaria y equipo” (p.78).

Según Hernández et al. (2006) la investigación es exploratoria ya que “el objetivo es examinar un tema o problema de investigación poco estudiado del cual se tienen muchas dudas o ha abordado antes. Es decir, cuando existen solo ideas vagamente relacionadas con el tema” (p.100).

Así mismo el diseño de la investigación es exploratorio, ya que se busca examinar y explorar los problemas conllevados a la depresión para proporcionar conocimientos que permitan la reducción de esta problemática.

3.2 Población y muestra

Población

Unidad de análisis: Alumnos de nivel universitario de la Universidad César Vallejo en el año 2017.

La población está conformada por alumnos de nivel universitario de la Universidad César Vallejo, al ser una de las más grandes empresas educativas del país; según el ranking de las 500 mayores empresas de América Economía.

Muestra

Según Torres, Bardales (2006) la muestra es un subconjunto de la población en que se llevará a cabo la investigación la muestra es una parte representativa de la población.” (p.165)

La muestra será de tipo dirigida porque la elección de sujetos u objetos de estudio depende del criterio del investigador. Se seleccionan sujetos o casos “típicos”.

En este caso la muestra fue elegida por **conveniencia** y los datos utilizados corresponden a 596 alumnos que pertenecen a la Universidad César Vallejo.

3.3 Técnicas e instrumentos

Según Hernández, Fernández y Baptista (2006), se dispone de diversos tipos de instrumentos para medir las variables de interés y en algunos casos llegan a combinarse varias técnicas de recolección de datos (p.23). Entre ellas se tiene:

- **Técnicas**

- a) Análisis documental

- **Instrumentos**

- a) Test de Frases incompletas de Sacks

- b) Inventario de Depresión de Beck II

3.4 Recolección de datos

Para la recolección de datos se procedió obteniendo los datos a partir del test de frases incompletas de Sacks y del Inventario de Depresión de Beck II que se aplicaron en las aulas del turno mañana, tarde y noche de la Universidad César Vallejo.

Además, se obtuvieron datos de las historias clínicas del Departamento de Psicología de la Universidad César Vallejo; para comprobar si los resultados obtenidos a través del modelo predictivo permiten la detección de personas con depresión.

Para realizar el procesamiento de información de la presente investigación, se empleará el TF-IDF que es una medida numérica que expresa cuán relevante es una palabra para un documento en una colección. De tal manera que se realizará el siguiente procedimiento que nos servirá como un procesamiento de texto previo para la aplicación de la técnica TF-IDF:

- Análisis léxico
- Eliminación de palabras vacías
- Procesos de stemming (reconocimiento de stems)
- Detección de grupos nominales

Luego realizaremos un análisis de los datos a través de los datos recolectados del Inventario de Depresión de Beck II; ya que estos permitirán identificar el tipo de clase de personas (con o sin depresión).

Por otro lado, para la minería de datos usaremos la librería tm de R.

Por último, se aplicará una de las técnicas de Aprendizaje Automático (Machine Learning) llamada Support Vector Machine.

CAPÍTULO IV : RESULTADOS Y ANÁLISIS DE RESULTADOS

4.1 Desarrollo del Modelo Predictivo

El diseño del modelo predictivo se realizó con una base de datos extraída a través de encuestas realizadas (ver anexo 03 y anexo 04) en la universidad César Vallejo. Este modelo fue diseñado en Microsoft Azure Machine Learning Studio.

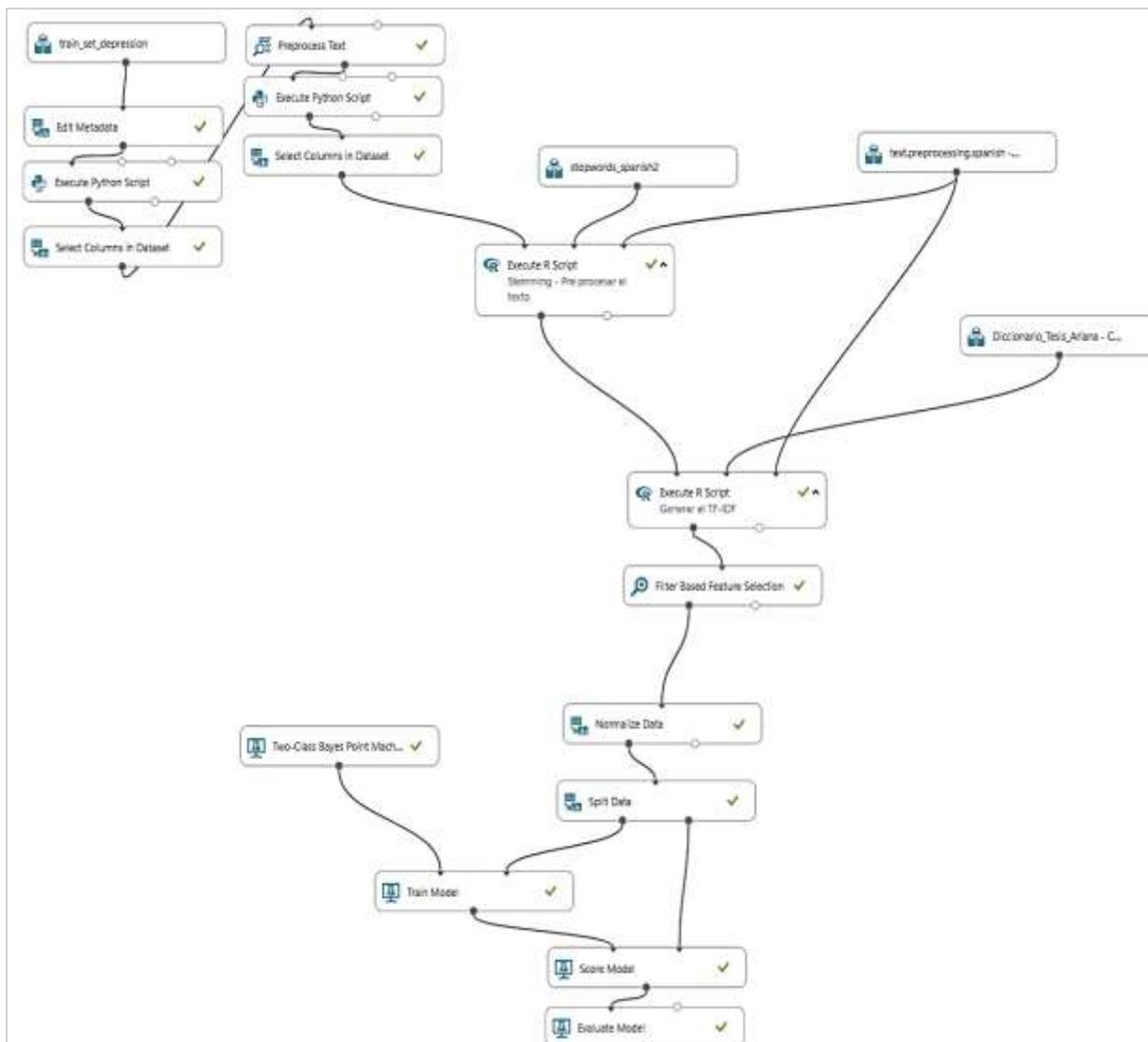


Figura 9: Modelo predictivo para detectar depresión
Fuente: Elaboración Propia

La base de datos utilizada que se observa en la figura consiste en 596 respuestas reales de estudiantes. A través del modelo, se ha implementado módulos como “Execute Python Script”:

- ✓ El primer módulo Python Script nos sirve para devolver los 60 resultados de las frases incompletas realizadas y transformarlos en una cadena.
- ✓ El segundo módulo Python Script nos permite quitar las tildes a las vocales de nuestros datos obtenidos.
- ✓ El tercer módulo Execute R Script determina los pasos para el preprocesamiento de datos: reemplazar los caracteres especiales (¡,¡,?,etc), eliminar los caracteres duplicados, reemplaza los números, convertir a minúsculas el texto, eliminar palabras vacías (de, el, ella, etc.).
- ✓ Por último, otro de los módulos con código Phyton permite procesar el texto excluyendo palabras con número de caracteres menores a 2 y mayores a 25. Además, asigna puerto de entradas opcionales a variables y conjuntos de datos.

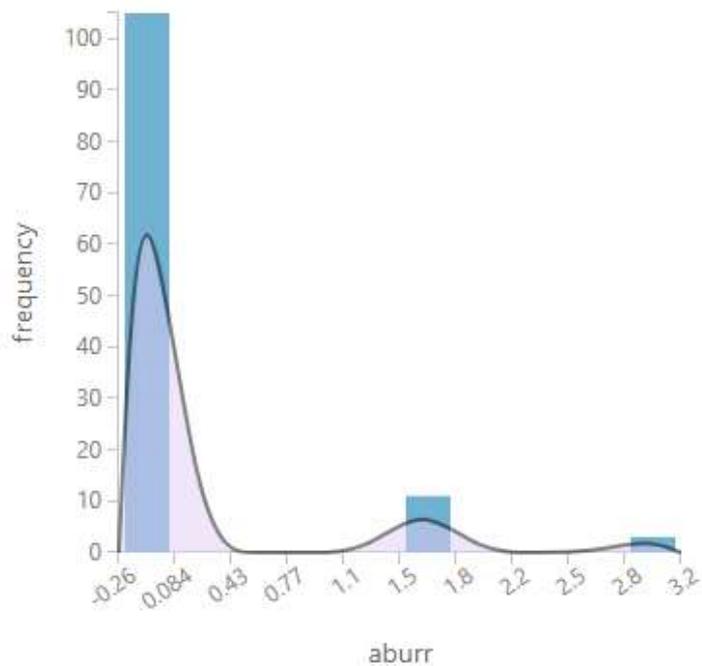
4.2 Resultados del análisis textual usando la técnica TF-IDF

Se ha aplicado la técnica del TF-IDF a la base de datos. Mediante esta técnica se ha obtenido las siguientes 100 palabras con sus respectivas ponderaciones numéricas, que van a permitir hallar el nivel de presencia que contienen estas palabras en una persona con depresión.

Además, en los siguientes gráficos se observarán los datos estadísticos de cada una de ellas, que corresponden al producto de ambos factores. Los resultados son una representación de la importancia del término en cada documento.

Para la demostración de la aplicación de la técnica del TF-IDF se ha tomado los 20 patrones más relevantes; donde se visualiza la frecuencia de cada palabra obtenida que ya ha sido normalizada.

De la palabra “aburr” se obtiene la siguiente figura 10, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

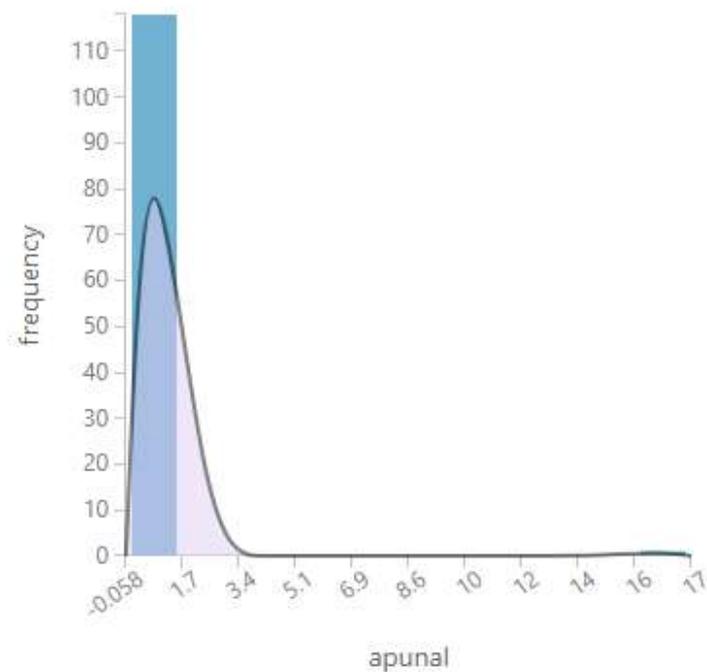


Mean	-0.0141
Median	-0.2604
Min	-0.2604
Max	3.1886
Standard Deviation	0.72

Figura 10: Histograma del patrón “aburr”

Fuente: Elaboración propia

De la palabra “apunal” se obtiene la siguiente figura 11, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

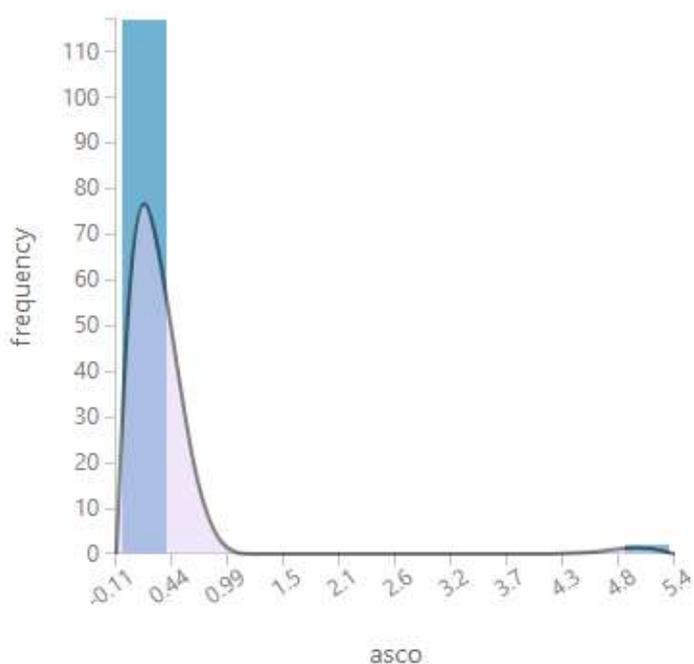


Mean	0.0873
Median	-0.058
Min	-0.058
Max	17.2337
Standard Deviation	1.5851

Figura 11: Histograma del patrón “apunal”

Fuente: Elaboración propia

De la palabra “asco” se obtiene la siguiente figura 12, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

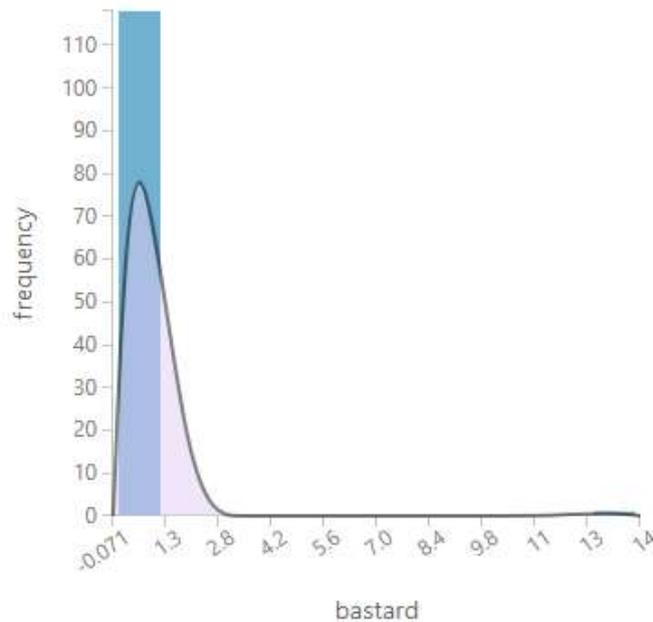


Mean	-0.0183
Median	-0.1106
Min	-0.1106
Max	5.3816
Standard Deviation	0.709

Figura 12: Histograma del patrón “asco”

Fuente: Elaboración propia

De la palabra “bastard” se obtiene de la figura 13, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

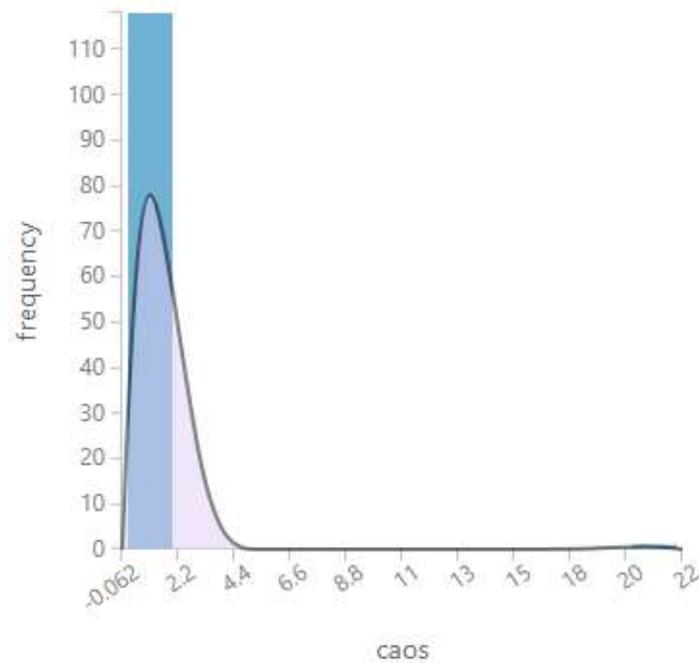


Mean	0.0476
Median	-0.0711
Min	-0.0711
Max	14.0594
Standard Deviation	1.2953

Figura 13: Histograma del patrón “bastard”

Fuente: Elaboración propia

De la palabra “caos” se obtiene la siguiente figura 14, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

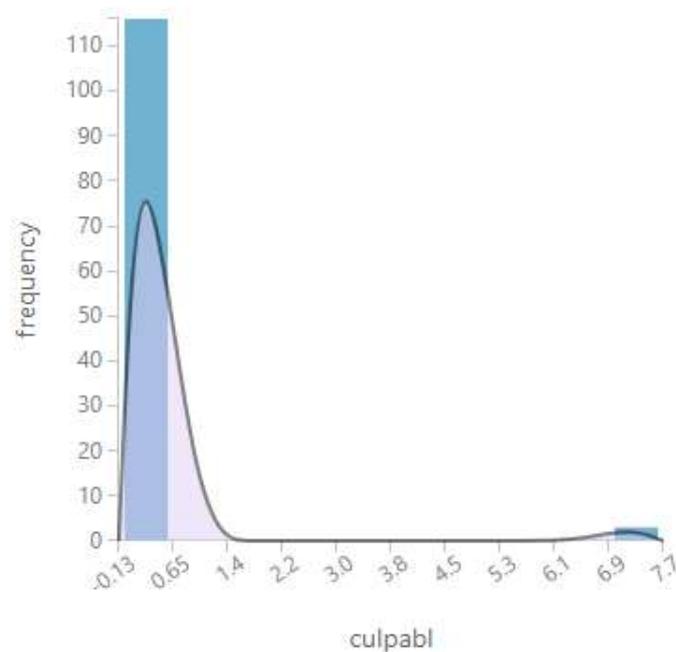


Mean	0.1241
Median	-0.0619
Min	-0.0619
Max	22.0628
Standard Deviation	2.0282

Figura 14: Histograma del patrón “caos”

Fuente: Elaboración propia

De la palabra “culpabl” se obtiene la siguiente figura 15, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

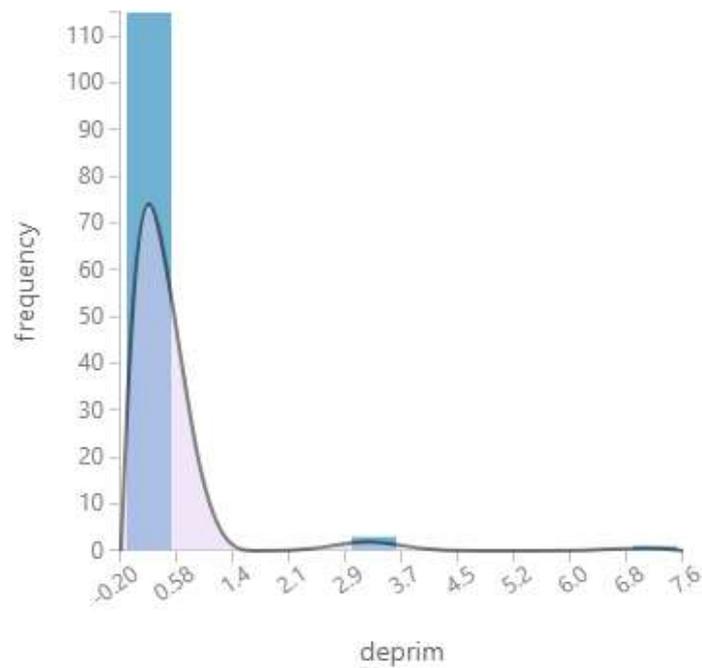


Mean	0.0656
Median	-0.1306
Min	-0.1306
Max	7.6551
Standard Deviation	1.2257

Figura 15: Histograma del patrón “culpabl”

Fuente: Elaboración propia

De la palabra “deprim” se obtiene la siguiente figura 16, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

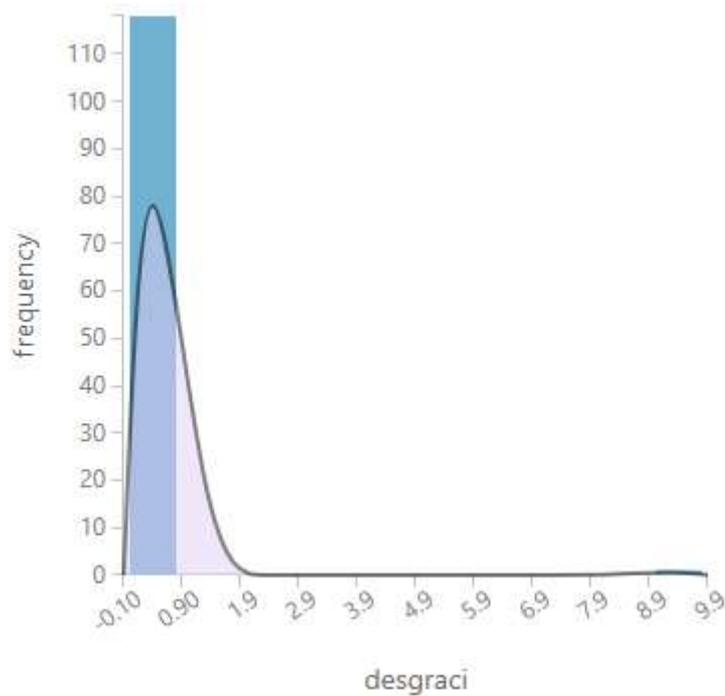


Mean	-0.0389
Median	-0.2023
Min	-0.2023
Max	7.5776
Standard Deviation	0.9331

Figura 16: Histograma del patrón “deprim”

Fuente: Elaboración propia

De la palabra “desgraci” se obtiene la siguiente figura 17, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

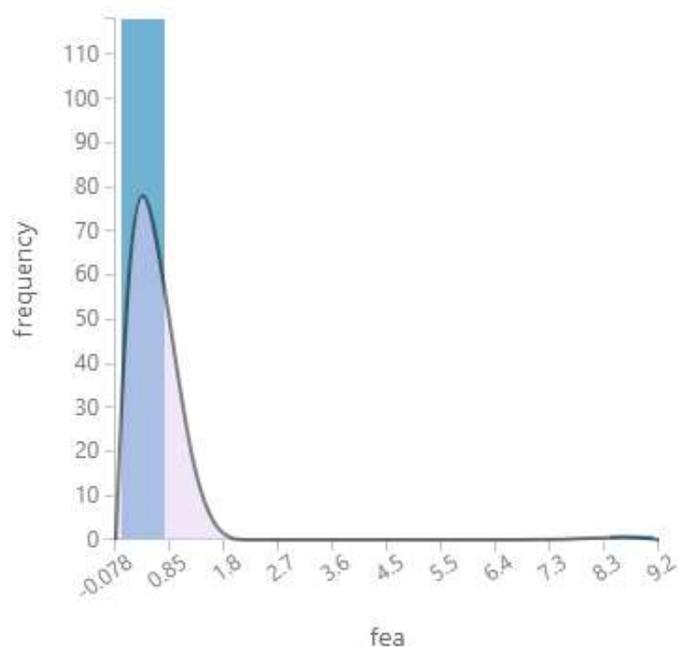


Mean	-0.0167
Median	-0.1008
Min	-0.1008
Max	9.9163
Standard Deviation	0.9183

Figura 17: Histograma del patrón “desgraci”

Fuente: Elaboración propia

De la palabra “fea” se obtiene la siguiente figura 18, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

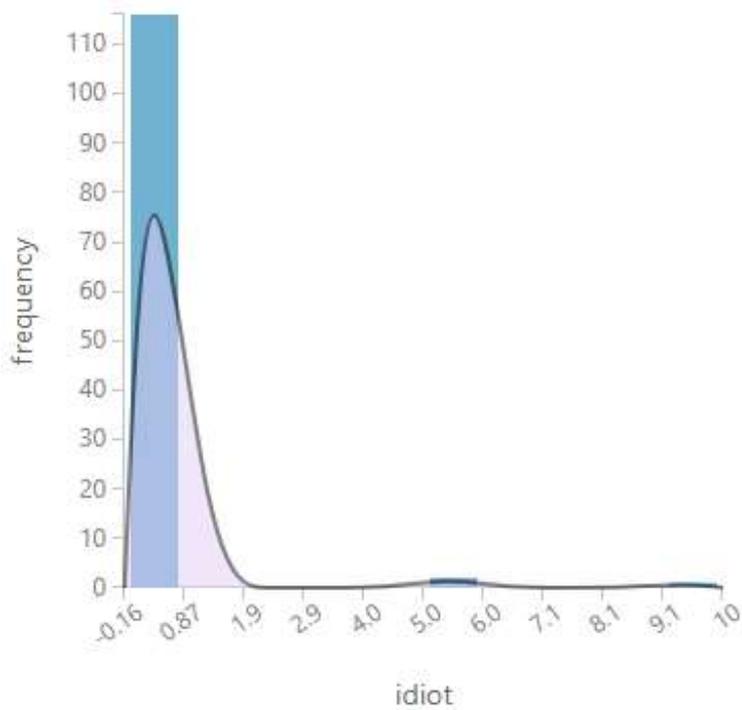


Mean	0.0001
Median	-0.0776
Min	-0.0776
Max	9.1774
Standard Deviation	0.8484

Figura 18: Histograma del patrón “fea”

Fuente: Elaboración propia

De la palabra “idiot” se obtiene la siguiente figura 19, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

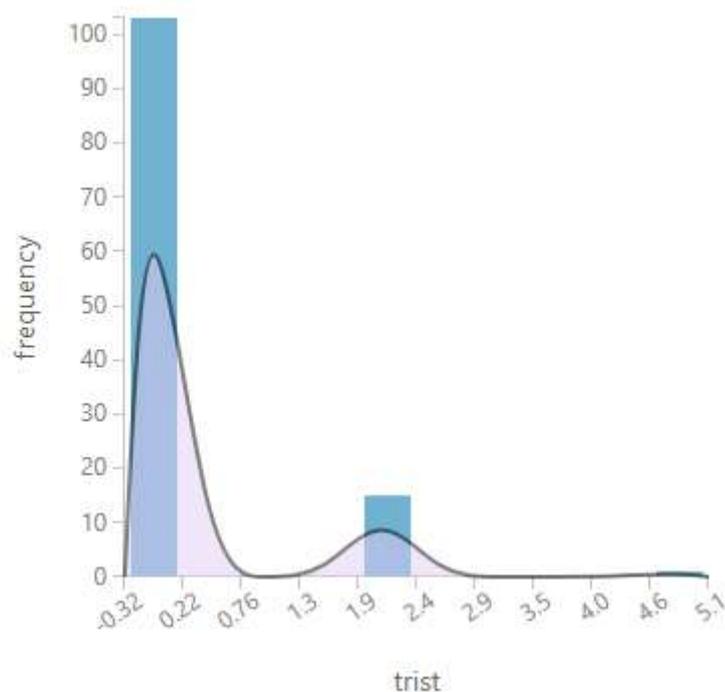


Mean	0.0089
Median	-0.1645
Min	-0.1645
Max	10.1533
Standard Deviation	1.1502

Figura 19: Histograma del patrón “idiot”

Fuente: Elaboración propia

De la palabra “trist” se obtiene la siguiente figura 20, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

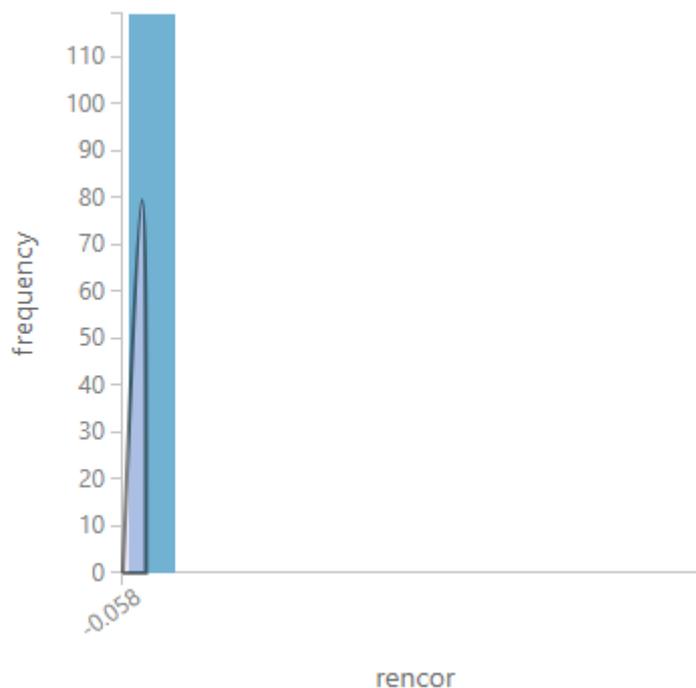


Mean	0.0646
Median	-0.3241
Min	-0.3241
Max	5.1165
Standard Deviation	1.0194

Figura 20: Histograma del patrón “trist”

Fuente: Elaboración propia

De la palabra “rencor” se obtiene la siguiente figura 21, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

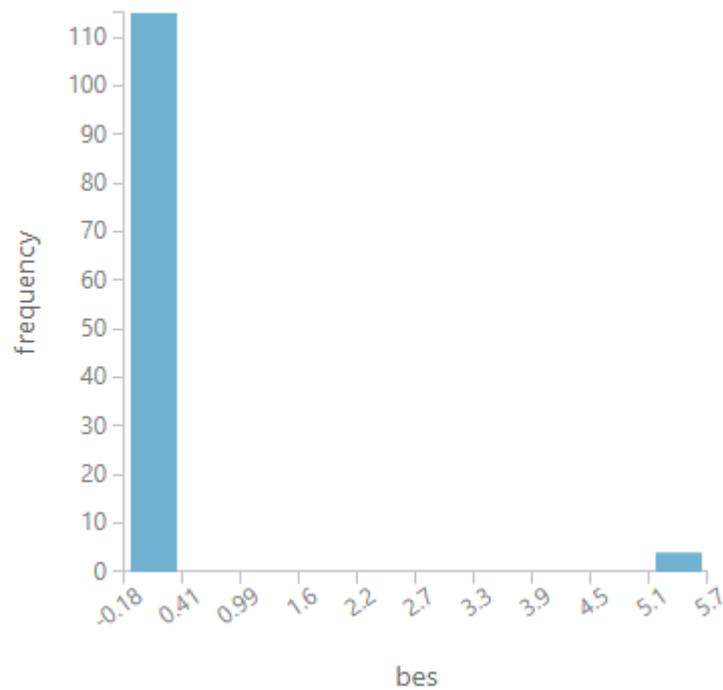


Mean	-0.058
Median	-0.058
Min	-0.058
Max	-0.058
Standard Deviation	0

Figura 21: Histograma del patrón “rencor”

Fuente: Elaboración propia

De la palabra “bes” se obtiene la siguiente figura 22, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

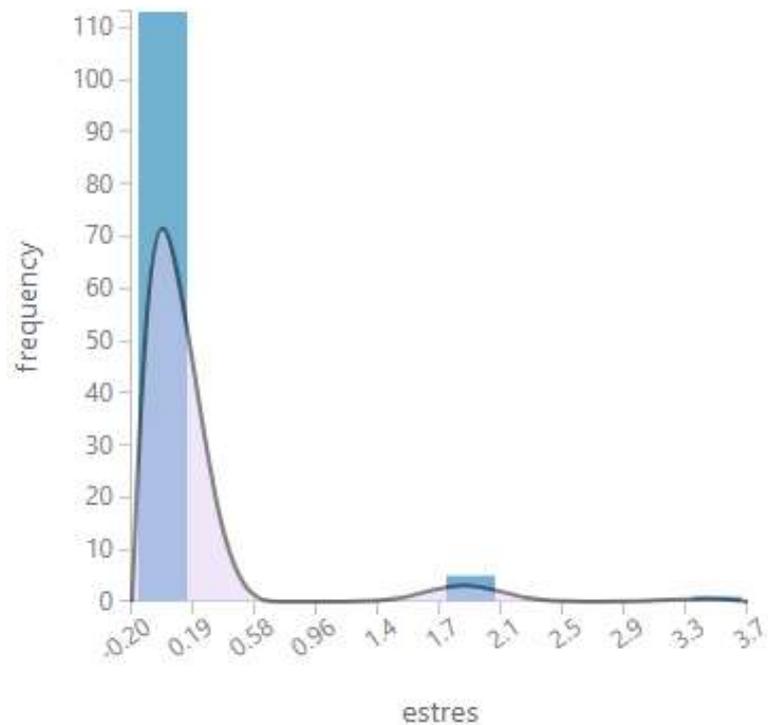


Mean	0.0199
Median	-0.1765
Min	-0.1765
Max	5.6667
Standard Deviation	1.0576

Figura 22: Histograma del patrón “bes”

Fuente: Elaboración propia

De la palabra “estres” se obtiene la siguiente figura 23, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

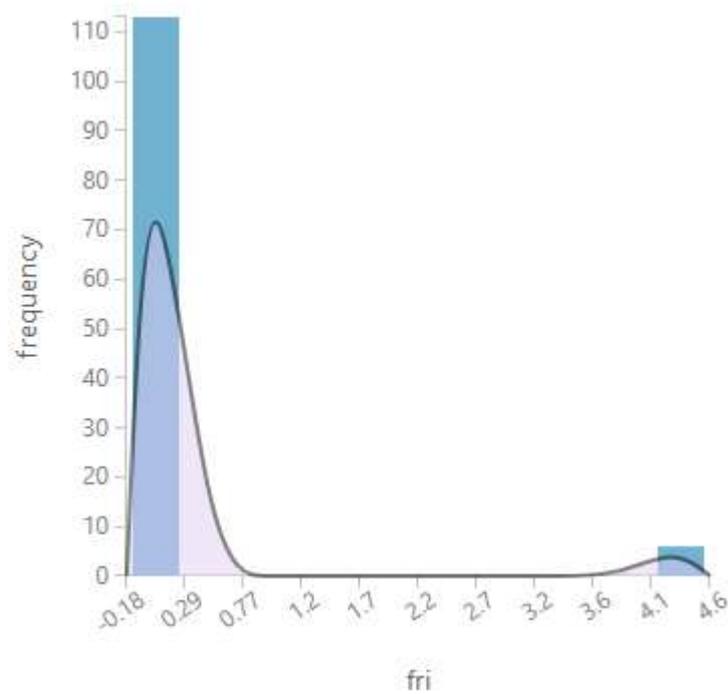


Mean	-0.0843
Median	-0.1983
Min	-0.1983
Max	3.6768
Standard Deviation	0.5227

Figura 23: Histograma del patrón “estres”

Fuente: Elaboración propia

De la palabra “fri” se obtiene la siguiente figura 24, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

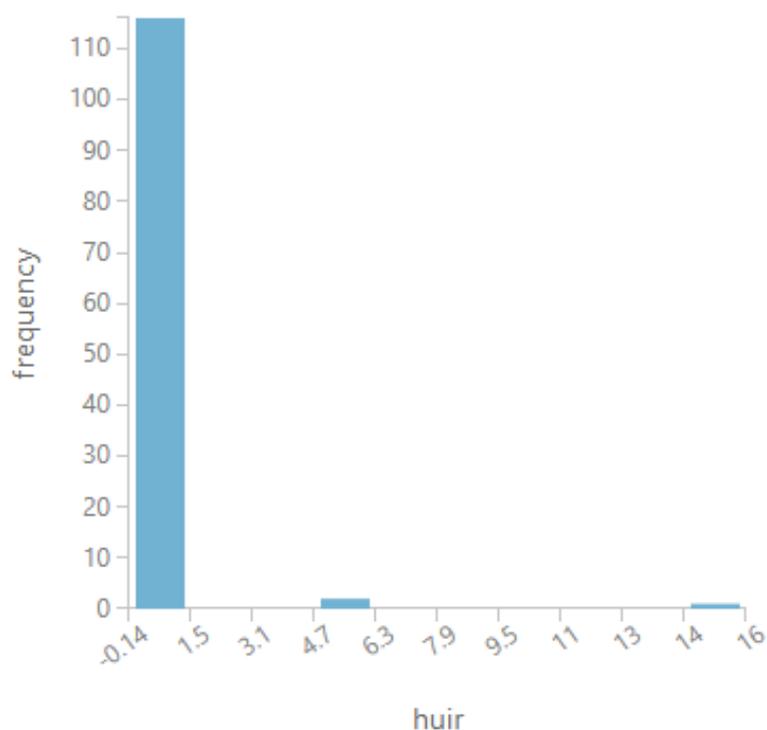


Mean	0.0565
Median	-0.1844
Min	-0.1844
Max	4.5931
Standard Deviation	1.0498

Figura 24: Histograma del patrón “fri”

Fuente: Elaboración propia

De la palabra “huir” se obtiene la siguiente figura 25, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

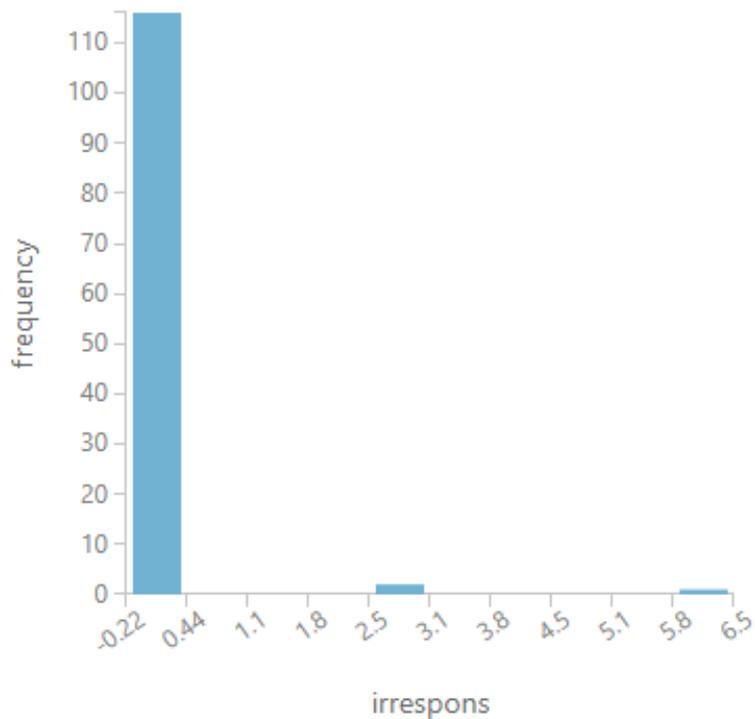


Mean	0.0906
Median	-0.1353
Min	-0.1353
Max	15.9925
Standard Deviation	1.6256

Figura 25: Histograma del patrón “huir”

Fuente: Elaboración propia

De la palabra “irrespons” se obtiene la siguiente figura 26, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

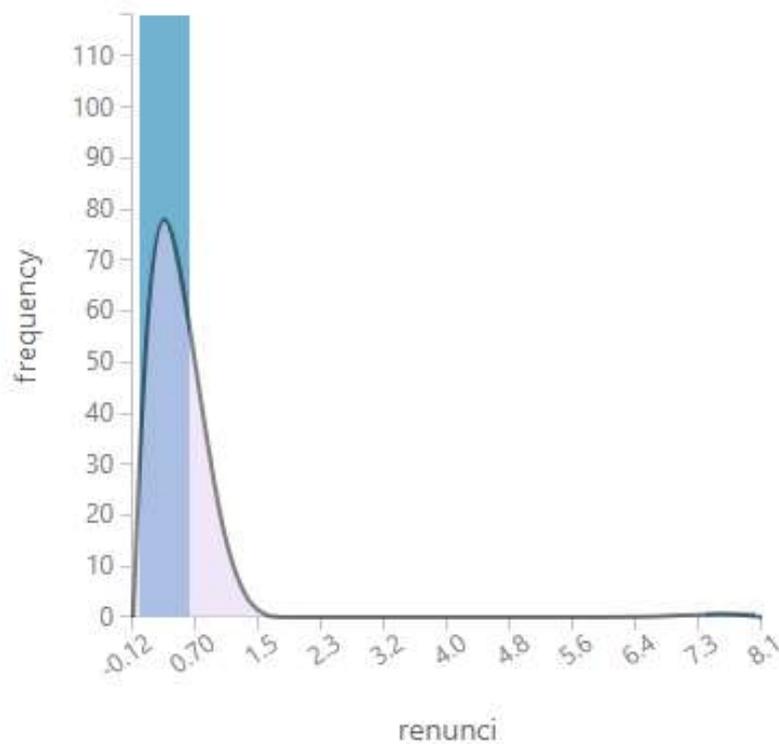


Mean	-0.112
Median	-0.2244
Min	-0.2244
Max	6.4625
Standard Deviation	0.7454

Figura 26: Histograma del patrón “irrespons”

Fuente: Elaboración propia

De la palabra “renunci” se obtiene la siguiente figura 27, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

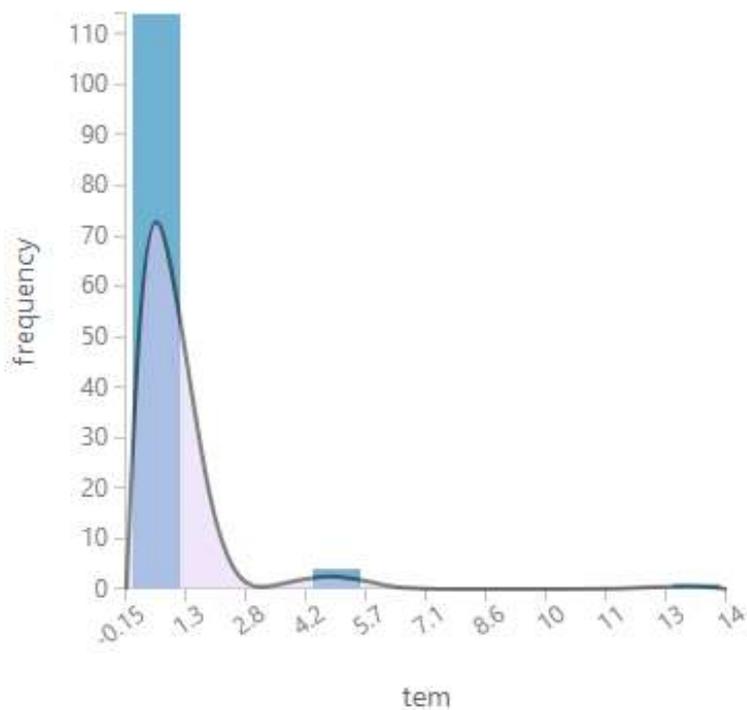


Mean	-0.0549
Median	-0.1238
Min	-0.1238
Max	8.076
Standard Deviation	0.7517

Figura 27: Histograma del patrón “renunci”

Fuente: Elaboración propia

De la palabra “renunci” se obtiene la siguiente figura 28, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.

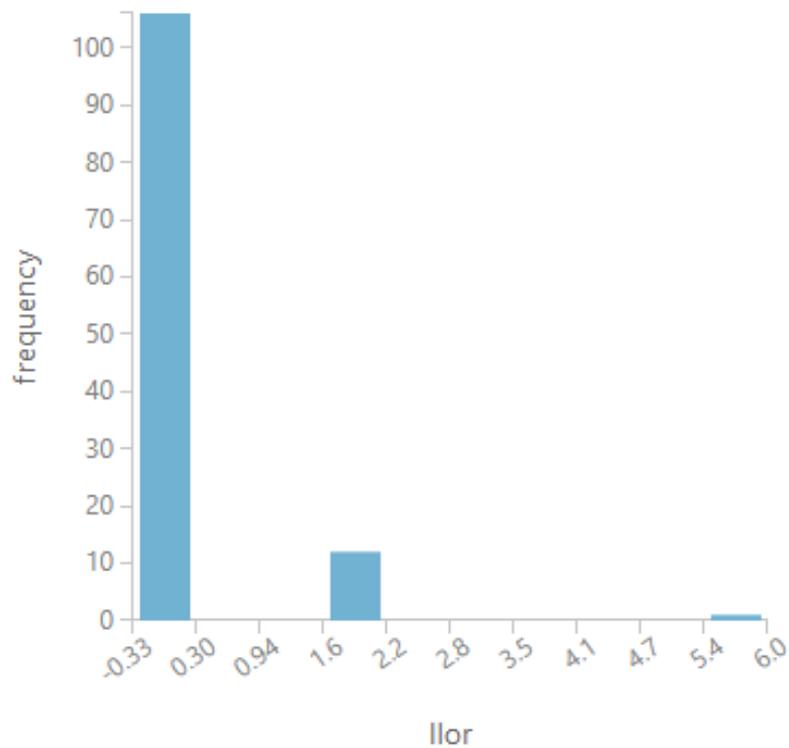


Mean	0.1385
Median	-0.1461
Min	-0.1461
Max	14.3698
Standard Deviation	1.5804

Figura 28: Histograma del patrón “tem”

Fuente: Elaboración propia

De la palabra “renunci” se obtiene la siguiente figura 29, donde se muestra sus respectivos cálculos de la ponderación que corresponde al producto de ambos factores (TF-IDF). Cabe resaltar, que el gráfico determina el valor medio (mean), el valor mediano (median), el menor valor (min) y el máximo valor (max) de las palabras identificadas por la técnica TF-IDF obtenidas de la colección total de encuestas realizadas; de esta forma se determina que más del 50 % usa la palabra mostrada en el gráfico.



Mean	-0.067
Median	-0.3337
Min	-0.3337
Max	6.0128
Standard Deviation	0.8513

Figura 29: Histograma del patrón “llor”

Fuente: Elaboración propia

4.3 Identificación de Patrones

Los patrones que se van a identificar se representarán a través de palabras más consistentes y éstas tendrán como variables un peso promedio y una desviación estándar del peso; debido al experimento realizado en Microsoft Azure Machine Learning Studio.

Para la identificación de patrones se ha realizado un tratamiento a la base de datos que consiste en aplicar la técnica del TF-IDF donde nos brinda una tabla con los patrones identificados que permiten detectar a posibles pacientes con depresión.

La Tabla 3, muestra las palabras generadas después de aplicar la técnica de TF-IDF, además se muestran los siguientes campos: característica (feature), peso medio (Weight Mean) y desviación estándar de pesos; los cuales nos sirven para cuantificar la cantidad de variación o dispersión de un conjunto de valores de datos.

Tabla 3

Identificación de patrones

Feature	Weight Mean	Weight Standard Deviation
especi	8.4028	1.54438
Bias	-0.489561	0.0896497
desgraci	0.321952	0.12119
duel	0.281051	0.156443
asco	0.260786	0.122112
inusual	0.254837	0.319503
equilibri	0.241441	0.095252
nub	0.241025	0.139792
spiderm	0.240436	0.165116
parezc	0.234533	0.13647
inmortal	0.221626	0.122147
enfurec	0.216656	0.113036
conviv	0.213272	0.110485
nacimient	0.21238	0.109561
concurs	0.208198	0.112419

trist	0.207807	0.077567
ajen	0.207622	0.122744
caos	0.204726	0.124708
normalaz	0.202548	0.115006
recre	0.196728	0.108354
Feature	Weight Mean	Weight Standard Deviation
buen	-0.196653	0.102206
ausenci	0.196323	0.119042
bonit	-0.192684	0.110183
malet	0.191533	0.123899
rocker	0.188959	0.115118
pinch	0.186102	0.116654
reyn	0.184366	0.123617
rencor	0.177159	0.12343
remedi	0.177152	0.125728
Rar	0.169464	0.0781001
quis	0.162062	0.0929607
amas	0.154722	0.124026
recuperar	0.147515	0.120757
irrit	0.147448	0.071377
enfad	0.147184	0.0824387
Est	0.146665	0.126882
avergonz	0.143575	0.203006
Diri	0.14044	0.0734278
apunal	0.136012	0.218102
estres	0.132888	0.0727659
pisot	0.131266	0.21847
Not	0.130548	0.0801739
agust	0.12635	0.093555
conversari	0.123576	0.144726
mias	0.123311	0.0636779
Llor	0.122626	0.0850941
espiritual	0.119845	0.0743855
huir	0.113371	0.0982295
Vivi	0.10724	0.0744849
echo	0.106612	0.15843
guitarr	0.0987051	0.0848446
Mis	-0.0985051	0.0963467

qued	0.0980657	0.0792465
Nac	0.0975609	0.0820405
habri	0.0964927	0.0950548
inici	0.0958277	0.0806729
tont	0.0952868	0.0749133
irrespons	0.0951953	0.0806395
detr	0.0903771	0.151259
expong	0.0870618	0.243132
Gfa	0.0870618	0.243132
Dig	0.0860019	0.0846848
Feature	Weight Mean	Weight Standard Deviation
horribl	0.0855819	0.104116
adel	-0.0827869	0.0977738
Tal	0.0826379	0.0813482
manzan	0.0819316	0.162671
escondi	0.0801728	0.161613
despediri	0.0763363	0.155329
deprim	0.0725374	0.0847027
york	0.0715868	0.165145
empres	-0.0712501	0.102243
nadi	0.06881	0.0824035
caraj	0.0685019	0.164073
religi	0.0678651	0.0815996
empez	0.0654479	0.077875
superficial	0.0617506	0.0757242
Sol	0.0599539	0.0842724
Fri	0.0591089	0.0950422
seas	0.0585217	0.169416
Dic	0.0565407	0.0941463
renunci	0.0556266	0.088763
culpabl	-0.0534836	0.113036
escuch	0.051522	0.0837303
Sid	-0.0402124	0.0880172
Toc	0.035103	0.0930206
Nul	0.0339815	0.0921508
disfuncional	-0.0327072	0.086205
atras	0.0317727	0.105484
unos	0.0289633	0.120259

aburr	0.0272154	0.0870934
primer	0.0259657	0.0857047
Bes	0.0234228	0.0881648
mat	0.0186485	0.101113
demasi	-0.0184956	0.088261
tonteri	-0.0160847	0.128622
Fea	0.0141466	0.101197
Tem	-0.006621	0.119696
idiot	0.00324802	0.0947958
volv	-0.0030739	0.100373
Era	0.00055888	0.0871707

Fuente: Elaboración propia

Nota. Muestra los 100 patrones representados en texto que se han encontrado después de aplicar la técnica de frecuencia de términos y frecuencia inversa de documentos (TF – IDF).

Por último, se implementó uno de los algoritmos clasificadores para evaluar el desempeño del modelo predictivo.

4.4 Comparación entre Indicadores de desempeño para clasificadores de dos clases

Se realizará una comparación entre los algoritmos clasificadores de dos clases con la finalidad de saber cuál de ellos tiene un mejor resultado en su probabilidad y de esa forma determinar su uso en nuestro modelo predictivo. Para este caso, nuestras dos clases son las siguientes:

Clase 0 = Sin depresión

Clase 1 = Con Depresión

Para dicha comparación, interpretaremos los siguientes indicadores:

TP: True Positive (Verdadero positivo), significa que el diagnóstico fue positivo y realmente si tienen depresión.

TN: True Negative (Verdadero negativo), significa que el diagnóstico fue negativo y realmente no tienen depresión.

FP: False Positive (Falso positivo), significa que el diagnóstico fue positivo y realmente no tienen depresión.

FN: False Negative (Falso Negativo), significa que el diagnóstico fue negativo y realmente si tienen depresión.

Accuracy: Es la proporción del total de predicciones correctas, obteniéndose de la siguiente manera

$$\frac{TP+TN}{TP+FP+FN+TN}$$

Precision: Es la proporción de precisión de nuestro modelo, obteniéndose de la siguiente manera:

$$\frac{TP}{TP+FP}$$

Recall: Es la proporción de los casos reales que se identifican como positivos.

$$\frac{TP}{TP+FN}$$

F1 score: Es el promedio ponderado de Precisión y Recuperación.

$$2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

AUC: Es una abreviatura para el área bajo la curva. Se usa en el análisis de clasificación para determinar cuál de los modelos usados predice mejor las clases.

A través de estos indicadores se muestra la relevancia de este modelo. Por esto, es conveniente elegir un modelo que evitase los falsos positivos; es decir, mayor exactitud, precisión y recuperación de la información. Cabe resaltar, que para entrenar el siguiente modelo se usará 90% para entrenar y 10% para validar de nuestra muestra de 596 encuestas.

En la figura 30, se muestra el desempeño del algoritmo Averaged Perceptron a través de una matriz de confusión empleada en el aprendizaje supervisado. Por otro lado, se muestra la exactitud (accuracy), precisión (precision), la recuperación (recall) y la medida útil de la precisión de una prueba (F1 Score). Además, se aprecia la curva ROC en el gráfico, representada por una línea azul que permite determinar el rendimiento del modelo de clasificación del algoritmo mencionado.

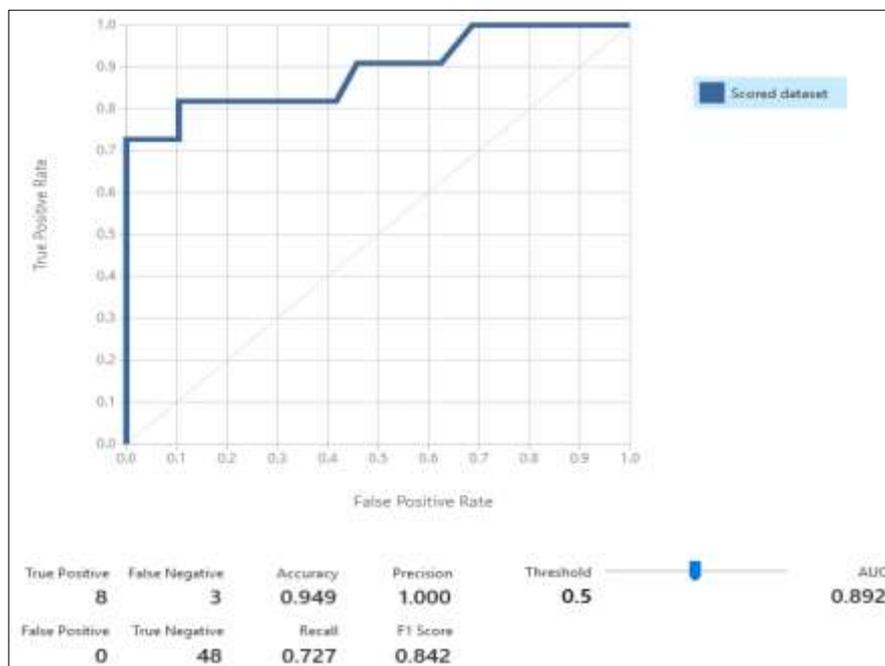


Figura 30: Indicador de desempeño del clasificador Averaged Perceptron

Fuente: Elaboración Propia

En la figura 31, se muestra el desempeño del algoritmo Bayes Point Machine a través de una matriz de confusión empleada en el aprendizaje supervisado. Por otro lado, se muestra la exactitud (accuracy), precisión (precision), la recuperación (recall) y la medida útil de la precisión de una prueba (F1 Score). Además, se aprecia la curva ROC en el gráfico, representada por una línea azul que permite determinar el rendimiento del modelo de clasificación del algoritmo mencionado.

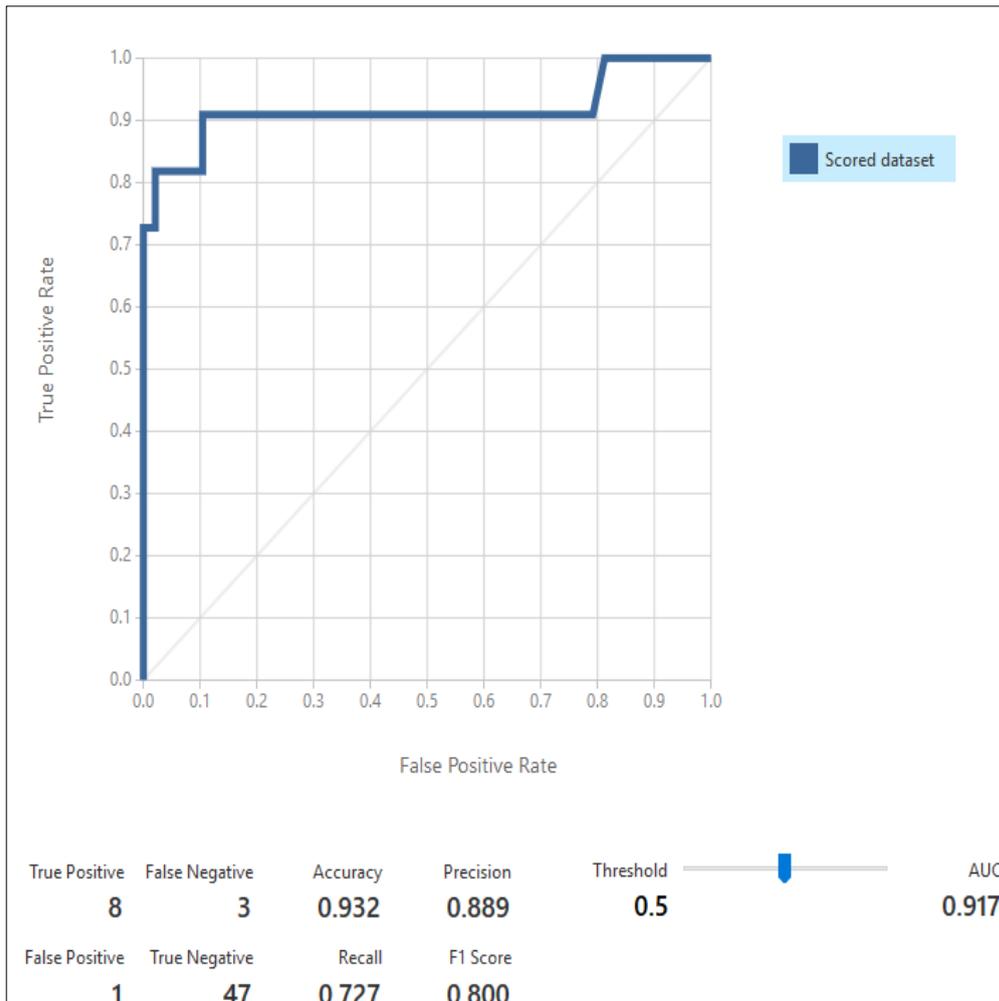


Figura 31: Indicador de desempeño del clasificador Bayes Point Machine

Fuente: Elaboración Propia

En la figura 32, se muestra el desempeño del algoritmo Boosted Decision Tree a través de una matriz de confusión empleada en el aprendizaje supervisado. Por otro lado, se muestra la exactitud (accuracy), precisión (precision), la recuperación (recall) y la medida útil de la precisión de una prueba (F1 Score). Además, se aprecia la curva ROC en el gráfico, representada por una línea azul que permite determinar el rendimiento del modelo de clasificación del algoritmo mencionado.

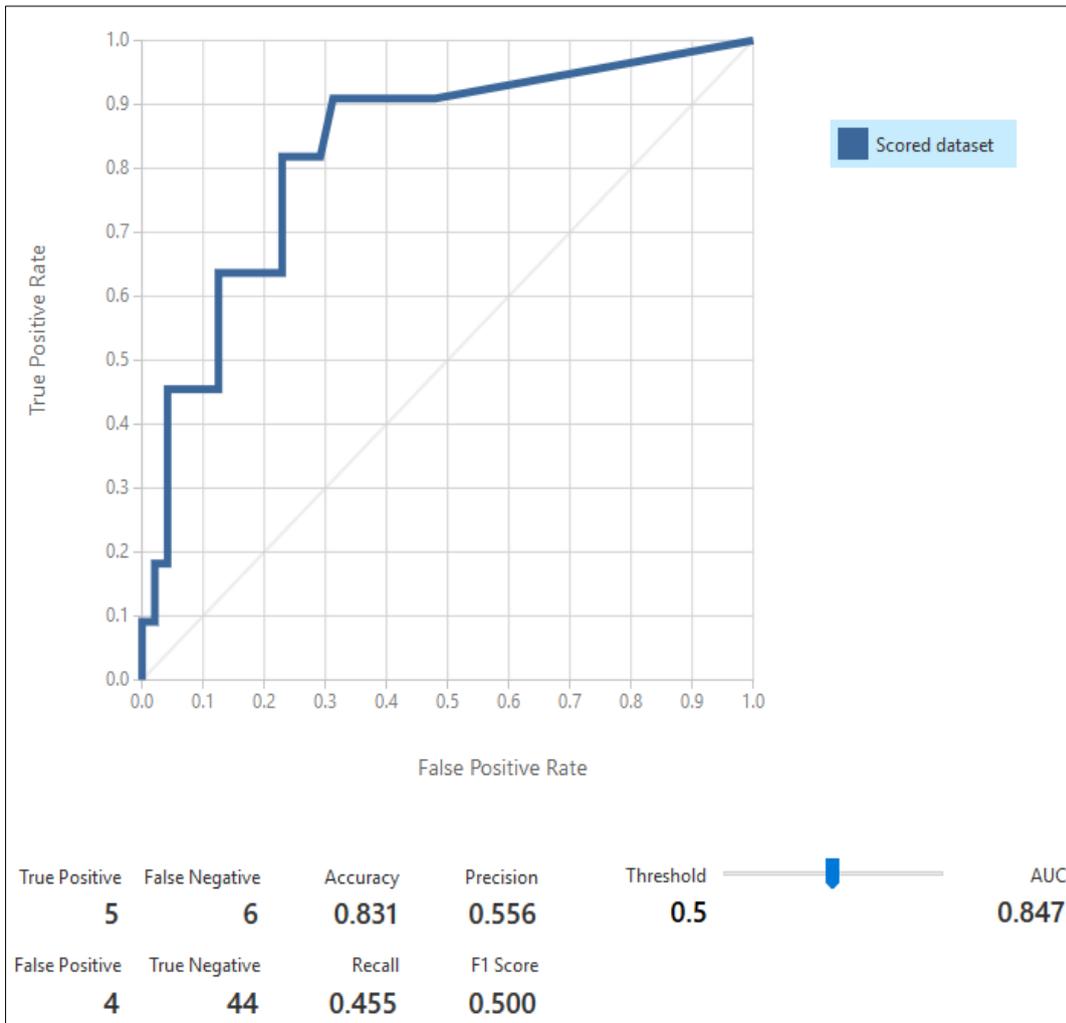


Figura 32: Indicador de desempeño del clasificador Boosted Decision Tree

Fuente: Elaboración Propia

En la figura 33, se muestra el desempeño del algoritmo Decision Forest a través de una matriz de confusión empleada en el aprendizaje supervisado. Por otro lado, se muestra la exactitud (accuracy), precisión (precision), la recuperación (recall) y la medida útil de la precisión de una prueba (F1 Score). Además, se aprecia la curva ROC en el gráfico, representada por una línea azul que permite determinar el rendimiento del modelo de clasificación del algoritmo mencionado.

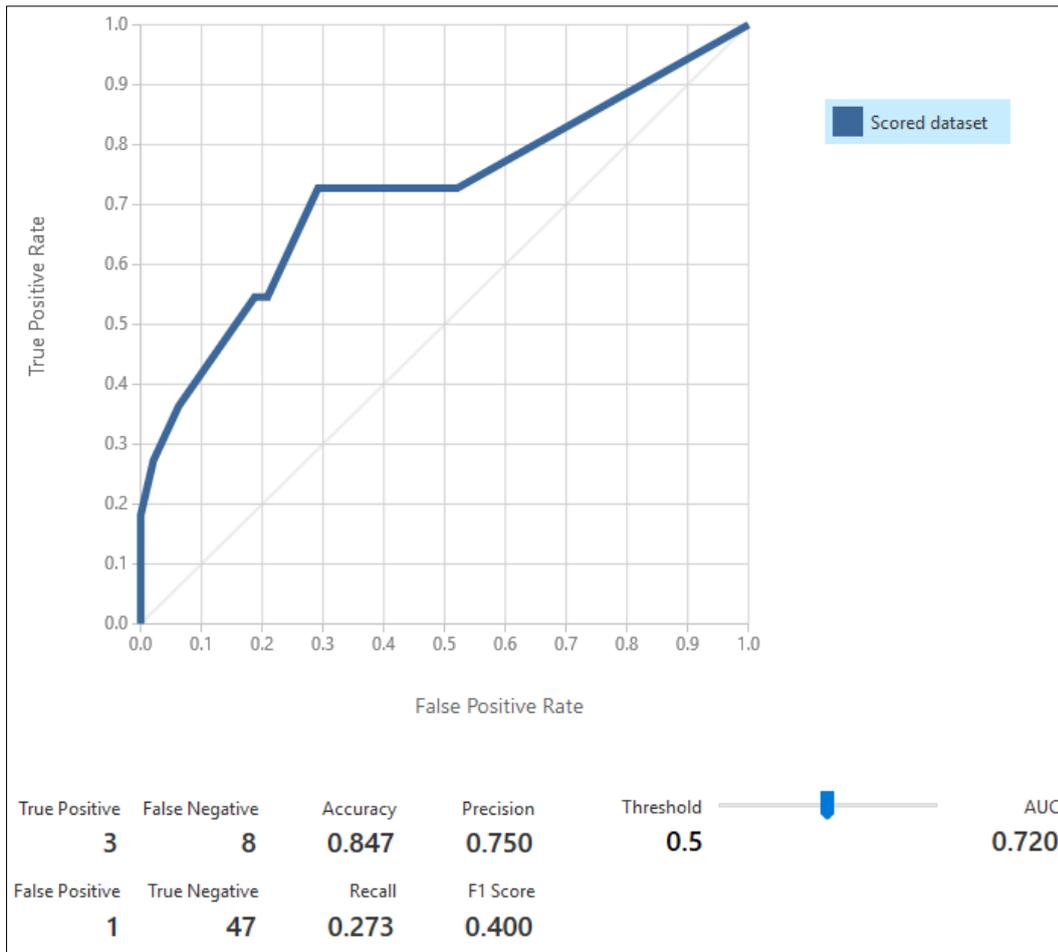


Figura 33: Indicador de desempeño del clasificador Decision Forest

Fuente: Elaboración Propia

En la figura 34, se muestra el desempeño del algoritmo Decision Jungle a través de una matriz de confusión empleada en el aprendizaje supervisado. Por otro lado, se muestra la exactitud (accuracy), precisión (precision), la recuperación (recall) y la medida útil de la precisión de una prueba (F1 Score). Además, se aprecia la curva ROC en el gráfico, representada por una línea azul que permite determinar el rendimiento del modelo de clasificación del algoritmo mencionado.

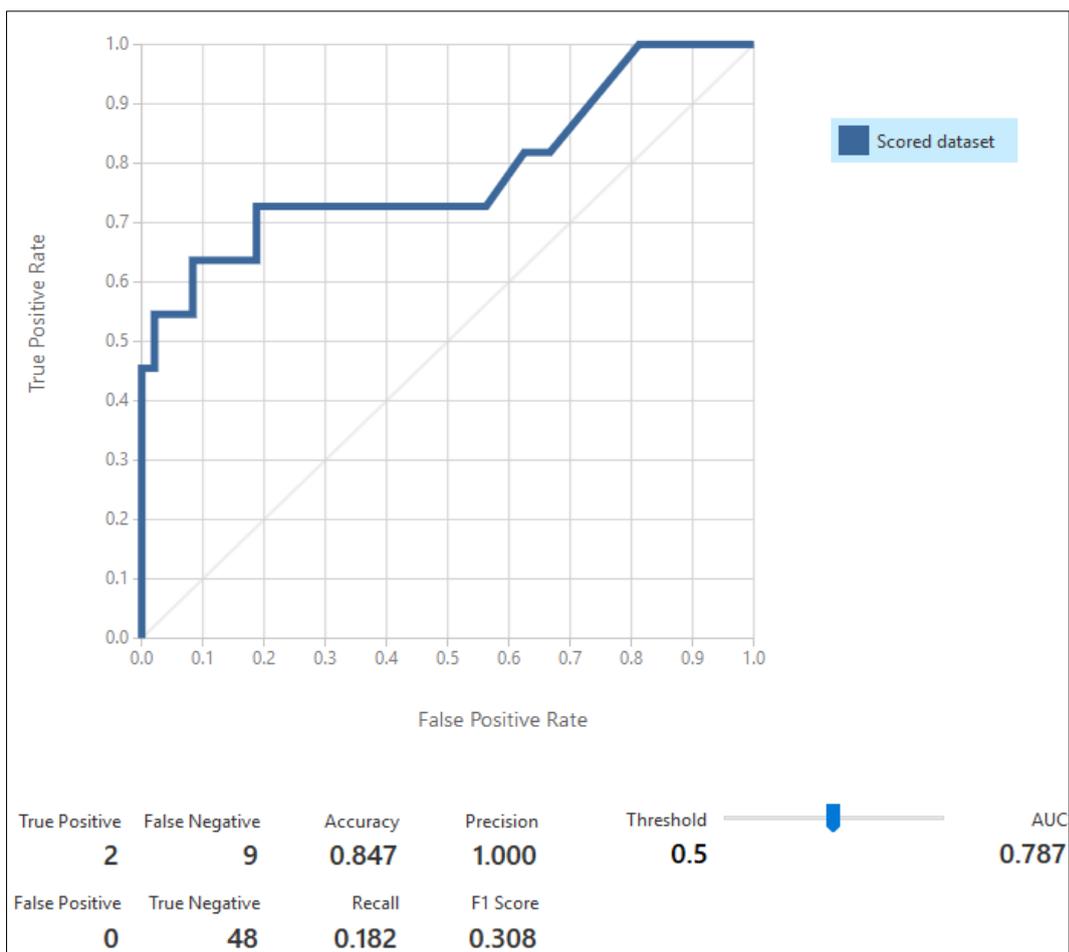


Figura 34: Indicador de desempeño del clasificador Decision Jungle

Fuente: Elaboración Propia

En la figura 35, se muestra el desempeño del algoritmo Support Vector Machine a través de una matriz de confusión empleada en el aprendizaje supervisado. Por otro lado, se muestra la exactitud (accuracy), precisión (precision), la recuperación (recall) y la medida útil de la precisión de una prueba (F1 Score). Además, se aprecia la curva ROC en el gráfico, representada por una línea azul que permite determinar el rendimiento del modelo de clasificación del algoritmo mencionado.

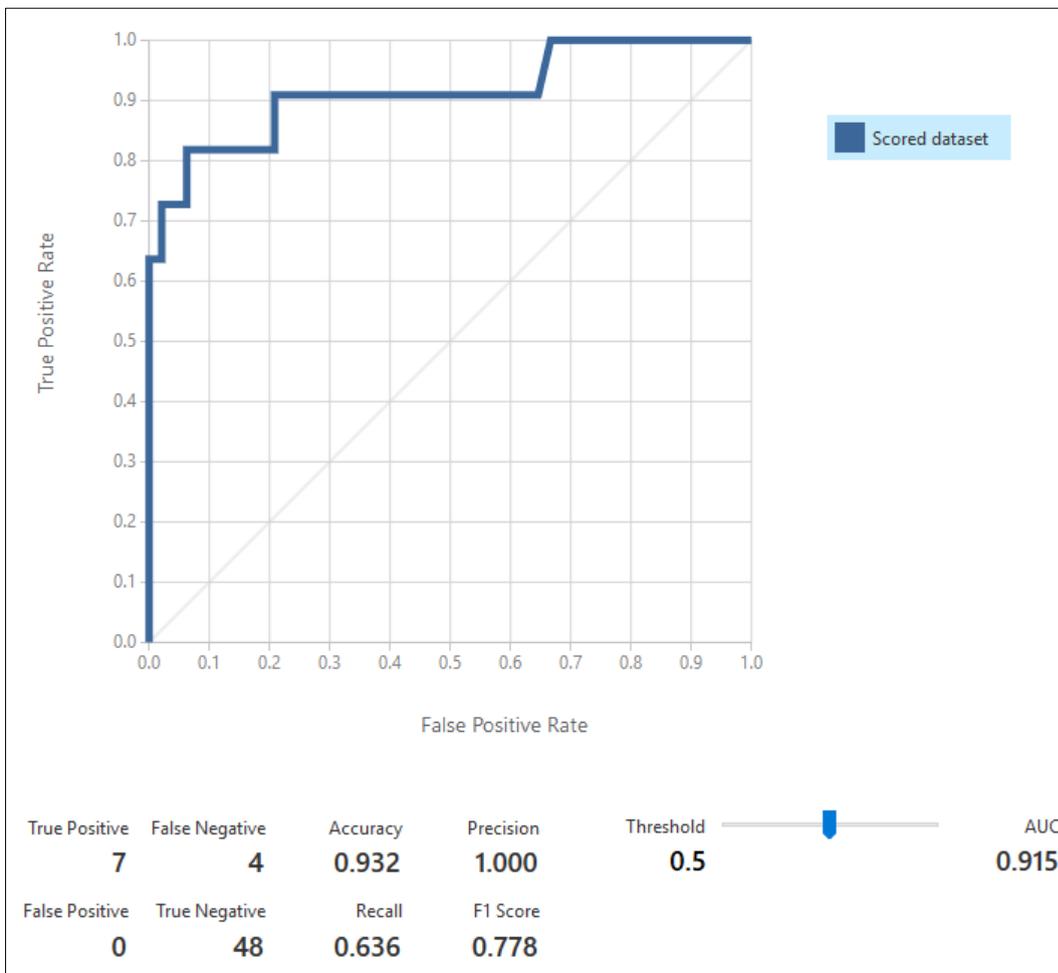


Figura 35: Indicador de desempeño del clasificador Support Vector Machine

Fuente: Elaboración Propia

En la figura 36, se muestra el desempeño del algoritmo Logistic Regression a través de una matriz de confusión empleada en el aprendizaje supervisado. Por otro lado, se muestra la exactitud (accuracy), precisión (precision), la recuperación (recall) y la medida útil de la precisión de una prueba (F1 Score). Además, se aprecia la curva ROC en el gráfico, representada por una línea azul que permite determinar el rendimiento del modelo de clasificación del algoritmo mencionado.

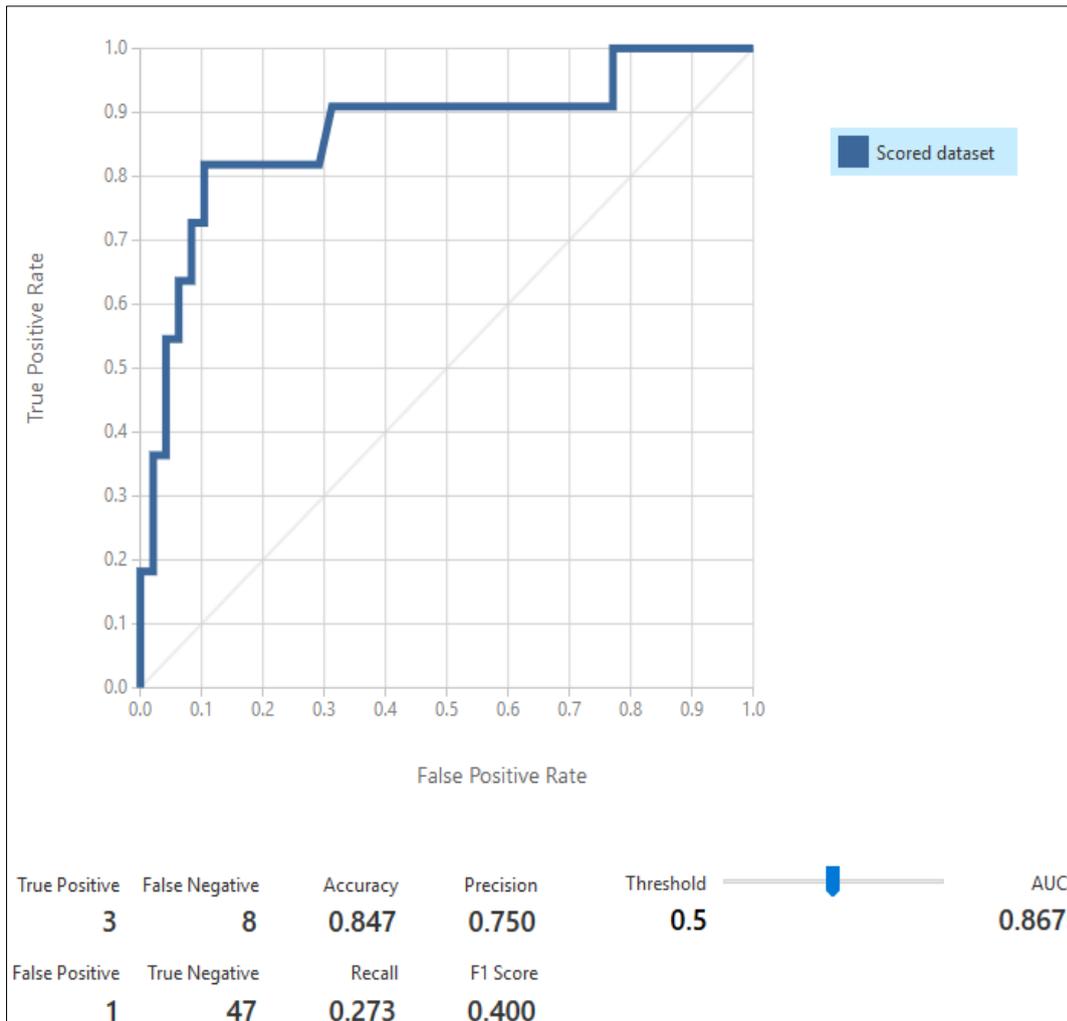


Figura 36: Indicador de desempeño del clasificador Logistic Regression

Fuente: Elaboración Propia

En la figura 37, se muestra el desempeño del algoritmo Neural Network a través de una matriz de confusión empleada en el aprendizaje supervisado. Por otro lado, se muestra la exactitud (accuracy), precisión (precision), la recuperación (recall) y la medida útil de la precisión de una prueba (F1 Score). Además, se aprecia la curva ROC en el gráfico, representada por una línea azul que permite determinar el rendimiento del modelo de clasificación del algoritmo mencionado.

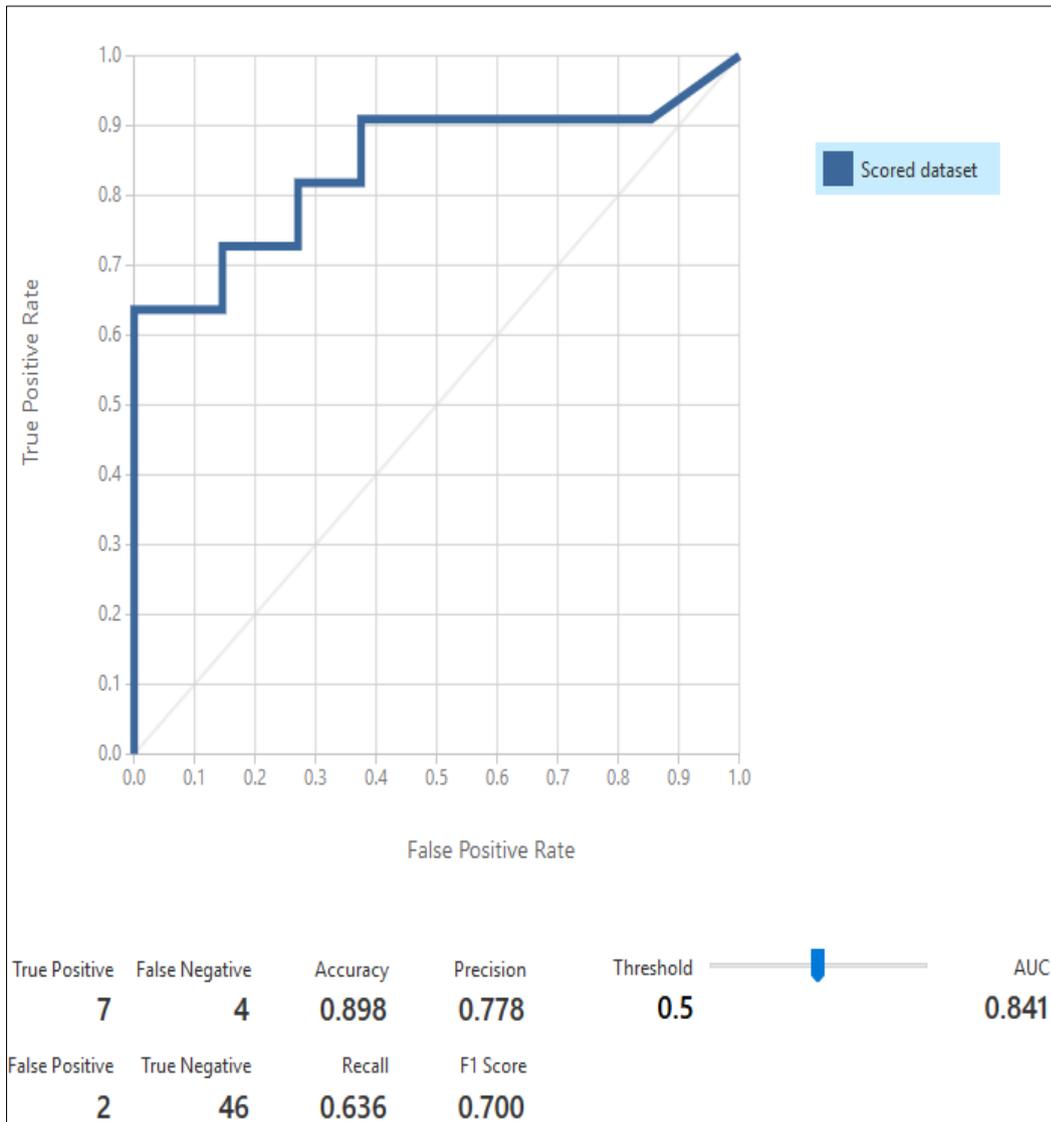


Figura 37: Indicador de desempeño del clasificador Neural Network

Fuente: Elaboración Propia

A continuación, la tabla 4 muestra la comparación de los algoritmos con sus respectivas propiedades: Accuracy, precisión, recall, F1 score y AUC.

Tabla 4

Comparación de los algoritmos clasificadores de dos clases

Depresión	Algoritmo	Accuracy	Precision	Recall	F1 Score	AUC
	Averaged Perceptron	0.949	1.000	0.727	0.842	0.892
	Bayes Point Machine	0.932	0.889	0.727	0.800	0.917
	Boosted Decision Tree	0.831	0.556	0.455	0.500	0.847
	Decision Forest	0.847	0.750	0.273	0.400	0.720
	Decision Jungle	0.847	1.000	0.182	0.308	0.787
	SVM	0.932	1.000	0.636	0.778	0.915
	Logistic Regression	0.847	0.750	0.273	0.400	0.867
	Neural Network	0.898	0.778	0.636	0.700	0.841

Fuente: Elaboración propia.

Nota. El algoritmo con mayor puntaje numérico elegido fue Bayes Point Machine.

A través de la Tabla 4, se aprecia que el clasificador de dos clases que se debe elegir para la evaluación del modelo predictivo es Bayes Point Machine; debido a que, sus valores que se muestran en sus indicadores de desempeño son más alto que el resto de los algoritmos.

4.5 Tiempo de atención de forma oportuna a probables pacientes con depresión

Según (2017) para cumplir los criterios de depresión, el paciente debe tener como mínimo cinco de los siguientes síntomas presentados anteriormente, durante un periodo igual o superior a dos semanas.

Además, en la problemática se indica que existe una cantidad considerable de casos sobre suicidios a causa de la depresión; ya que los afectados no son atendidos de forma oportuna porque no saben que padecen de dicha enfermedad.

A través de esta aplicación web basada en el análisis textual utilizando aprendizaje automático se logra identificar patrones de depresión. Debido a la disponibilidad y fácil acceso de la aplicación “ANTIDEP” los posibles pacientes podrán obtener un resultado inmediato, menor a 2 horas; esto les permite optimizar el tiempo de atención para acudir con ayuda profesional de forma oportuna.

4.6 Eficacia de la detección de probables pacientes con depresión

Uno de los parámetros que más destacan para evaluar la eficacia de una prueba diagnóstica son las curvas ROC, por el hecho de que son índices de la exactitud diagnóstica y proporcionan un criterio unificador en el proceso de evaluación de una prueba (Ruiz de Adana, 2009).

Para hallar la eficacia de este modelo predictivo se ha usado el clasificador “Bayes Point Machine” ya que, de acuerdo a la comparación con los demás algoritmos, este presenta un mejor desempeño con respecto a sus indicadores (accuracy, precisión, recall, F1 Score, AUC).

La figura 38 muestra un problema de predicción de clases binario, en la que los resultados se etiquetan positivos (p) o negativos (n). Hay cuatro posibles resultados a partir de un clasificador binario como el propuesto.

Se consideró una prueba diagnóstica que persiga determinar si una persona tiene depresión o no. Para este modelo se entrenó con el 90% de datos y el 10 % de validación de una muestra de 596 encuestas. Los resultados fueron los siguientes:

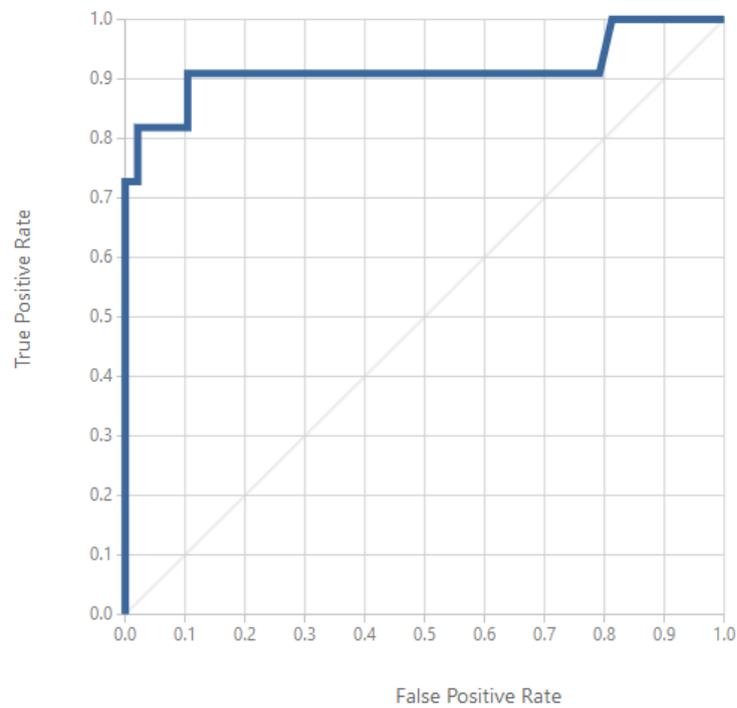


Figura 38: Curva de ROC

Fuente: Elaboración Propia

A partir de las P instancias positivas y N negativas se construyó la Matriz de Confusión que se muestra en la tabla 5.

Tabla 5

Matriz de confusión

True Positive	False Negative
8	3
False Positive	True Negative
1	47

Fuente: Elaboración Propia.

Nota. Los resultados mostrados pertenecen al 10% de los datos evaluados de la base de dato de 596 encuestas.

En la Tabla 5, se muestra la matriz de confusión donde se aprecia que los Verdaderos Positivos (personas con diagnóstico positivo que realmente sufren de depresión) son de 8 instancias, los Falsos Negativos (personas con diagnóstico negativo que realmente sufren de depresión) de 3 instancias, los Falsos Positivos (personas con diagnóstico positivo que realmente no sufren de depresión) de 1 instancia y, por último, los Verdaderos Negativos (personas con diagnóstico negativo que realmente no sufren de depresión) de 47 instancias.

Además, en la herramienta de Microsoft Azure usada conjuntamente con la Librería R dieron los siguientes resultados, indicando que:

La exactitud del modelo predictivo y el tratamiento de datos realizado fue de 0.932.

Accuracy
0.932

De la misma forma, la precisión del tratamiento de datos fue de 0.889.

Precision
0.889

Por último, el área bajo la curva ROC (AUC, area under the curve) para esta prueba de diagnóstico fue de 0.917. Este permite mostrar la capacidad del test para discriminar posibles pacientes con y sin depresión de forma eficaz.

AUC
0.917

4.7 Web Service para detectar la depresión

Esta tecnología se ha realizado a través de un modelo predictivo de Aprendizaje automático de Azure como un servicio web, puede usar una API REST para enviar datos y obtener predicciones.

Además, al generar el Web Service te brinda en su resultado un API Key.

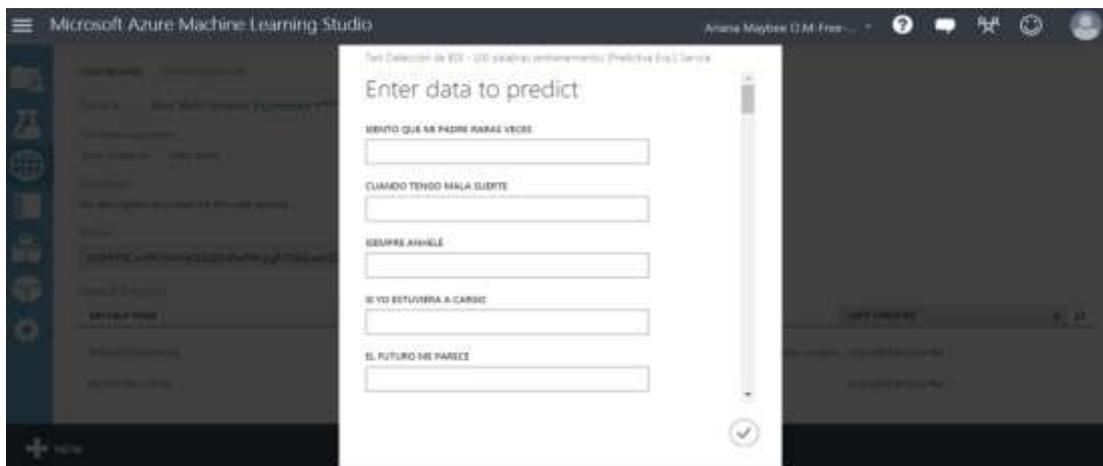


Figura 39: Web service para detectar depresión

Fuente: Elaboración Propia

4.8 Aplicación Web ANTIDEP

La aplicación web ANTIDEP permite detectar la depresión en posibles pacientes de depresión con un nivel de exactitud de 93% y con la capacidad de discriminar entre personas con depresión y sin depresión a un 92 %. ANTIDEP está orientada al patrón de arquitectura Modelo Vista Controlador. El lenguaje de programación utilizado para el desarrollo es PHP. Esta aplicación tiene dos módulos

“¿Para qué sirve?” y “¿Depresión o no? Haz clic”.



Figura 40: Módulo principal de la aplicación web “ANTIDEP”

Fuente: Elaboración propia

El link para obtener el acceso a la aplicación web es el siguiente:
<http://predictivo.byethost7.com/>

El primer módulo consiste en brindar información sobre la depresión. Además, contiene un botón “IR A CUESTIONARIO” que te direcciona al siguiente módulo.

El segundo módulo está compuesto por 60 preguntas que pertenecen al cuestionario de “Frases Incompletas de Sacks”. A través de este módulo, se consumen los datos del Web Service que se ha creado en la herramienta de Microsoft Azure Machine Learning.

Figura 41: Segundo módulo de la aplicación web “ANTIDEP”

Fuente: Elaboración propia

Además, dependiendo de la respuesta si tiene o no depresión esta aplicación brindará cualquier de los dos resultados: si tiene o no depresión. En caso, de tener depresión, aparece una ventana indicando una lista de consejos.



Figura 42: Ventana de respuesta positiva indicando que no existe depresión

Fuente: Elaboración propia



Figura 43: Módulo de consejos para personas que tienen depresión

Fuente: Elaboración propia

4.9 Análisis de Resultados

El modelo predictivo fue entrenado con 537 encuestas y fue validado con 59. Se utilizaron 596 encuestas en total.

En la matriz de confusión se mostró que: los Verdaderos Positivos (personas con diagnóstico positivo que realmente sufren de depresión) fueron de 8 instancias, los Falsos Negativos (personas con diagnóstico negativo que realmente sufren de depresión) de 3 instancias, los Falsos Positivos (personas con diagnóstico positivo que realmente no sufren de depresión) de 1 instancia y, por último, los Verdaderos Negativos (personas con diagnóstico negativo que realmente no sufren de depresión) de 47 instancias. Esto quiere decir que, de 59 instancias, en 55 los resultados fueron correctos y en 4 instancias los resultados fueron erróneos.

Los resultados indican que la exactitud del modelo predictivo fue del 93,2%; lo cual indica, que el uso del análisis textual y el aprendizaje automático permitirán la

detección de la depresión de forma eficaz. Igualmente, la precisión del modelo fue de 88,9 % lo que indica que el modelo creado tiene un margen de error muy bajo a equivocarse.

Cabe resaltar, que para la validación de dicho resultado se han utilizado herramientas reconocidas y validadas a nivel mundial como el “Inventario de Beck II” y “Frasas incompletas Sacks” que apoyaron a la identificación de los alumnos con depresión. La implementación de estas herramientas, fueron autorizadas por la Dra. Jimenez Berrú Pilar, Directora Académica; y Dr. Edgard Rivero, Director del Bienestar Universitario de la de la Universidad César Vallejo – Callao. (Ver anexo 04)

Por último, se indica que AUC (area under the curve) es de 91,7%, significando que la probabilidad de detectar a una persona con depresión es muy alta.

CAPÍTULO V : CONCLUSIONES Y RECOMENDACIONES

Conclusiones

Se concluye que el uso del análisis textual y el aprendizaje automático en un modelo predictivo permitió identificar patrones específicos en individuos con depresión a través de las técnicas de TF-IDF. Los patrones identificados se representaron en palabras que pasaron por un proceso de normalización, a las cuales se les agregó una ponderación numérica para saber su frecuencia en los documentos obtenidos en la base de datos. Cabe resaltar que, mediante los patrones identificados, se ayudó a detectar a personas con depresión.

Se concluye que el uso del análisis textual y el aprendizaje automático en una aplicación web permitió detectar de forma eficaz la depresión; debido que esta aplicación llamada “ANTIDEP” consumió los datos del modelo predictivo que se generó en Microsoft Azure Machine Learning; el cual tiene una probabilidad muy alta de 91,7% para detectar de forma correcta la depresión.

Se concluye que el uso del análisis textual y el aprendizaje automático en una aplicación web optimizó el tiempo de atención de forma oportuna a probables pacientes con depresión de una forma considerable; debido a que el tiempo promedio de detección de depresión es de dos semanas y bajo el uso de esta aplicación web se puede detectar en menos de dos horas.

Finalmente, se puede concluir que la presente investigación permitió la creación de una aplicación web utilizando técnicas de análisis textual y aprendizaje automático para poder detectar la depresión; teniendo como propósito apoyar a la disminución de casos de suicidio en el país y en el mundo.

Recomendaciones

- a. Se recomienda utilizar la aplicación web como herramienta de detección precoz de la depresión.
- b. Se recomienda la actualización de la base de datos en el modelo predictivo de forma periódica; para obtener datos de relevancia que se enfrenten a problemáticas actuales y se logre una mejor toma de decisiones.
- c. Se recomienda el apoyo incondicional del profesional del campo que se esté abarcando, para obtener conocimientos más precisos y claros respecto a la investigación que se está realizando.
- d. Se recomienda plantear posteriores investigaciones, aplicando técnicas y conocimientos de inteligencia artificial, para que permitan aumentar medidas correctivas ante cualquier suceso o problemática de gran relevancia en el sector de la salud en Perú.

REFERENCIAS BIBLIOGRÁFICAS

Adobe Systems Incorporated (2018). Aspectos básicos de las aplicaciones web. Recuperada de: <https://helpx.adobe.com/es/dreamweaver/using/web-applications.html>

Alcalde, R., Valdiviezo, V. (2013). *Desarrollo de un método para la clasificación automática de imágenes faciales basada en el atractivo aplicando máquinas de aprendizaje*. (Tesis para optar por el título profesional de Ingeniero Informático). Recuperada de: <http://dspace.unitru.edu.pe/handle/UNITRU/8318>

Alvira, M. (2002). *Perspectiva cualitativa / perspectiva cuantitativa en la metodología sociológica*. México. México: Mc Graw Hill.

American Psychiatric Association. *Diagnostic and statistical manual of mental disorder: DSM-5*. Arlington (VA): American Psychiatric Association; 2017.

Beck A. T., Steer R.A., Brown G.K. (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

Bishop, C (2006). *Pattern Recognition and Machine Learning*. Edición: 1st ed. 2006. Corr. 2nd printing 2011 (15 de febrero de 2010). **ISBN-10: 0387310738**

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers*. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92* (pp. 144-152). New York, NY, USA: ACM

- Breiman (2001). *Statistical modeling: The two cultures*. *Statistical Science* 16, 3 (2001), 199–231.
- Cabrera, Frías, Hernández, Iniestra y Solís (2014). *Scrutiny of depression and anxiety on Twitter through a word analysis program* (pg.16-21). Recuperado de: [https://doi.org/10.1016/S2007-5057\(15\)72164-2](https://doi.org/10.1016/S2007-5057(15)72164-2).
- Cleveland (2001). *Data science: An action plan for expanding the technical areas of the field of statistics*. Internatio
- Cristianini, N. y Shawe-Taylor, J. (2004). *An introduction to support vector machines and other kernel-based learning methods kernel methods for pattern analysis*. University of London. ISBN: 9780521780193
- Cortes, C. & Vapnik, V. (1995). *Support-vector networks*. *Machine Learning*, 20(3), 273–297.
- Eyssautier, M. (2006). *Metodología de la Investigación*. ISBN: 9789706863843. Editorial: Cengage Learning. Edición: 5.
- Feldman, R. and James Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.
- Google. 2016d. Google Trends. (2016). <https://www.google.com.au/trends/explore#q=datalyticsz=Etc> Retrieved on 14 November 2016.
- Graham, M (2012). *The art of data science*. In *Astrostatistics and Data Mining, Springer Series in Astrostatistics*, Vol. 2. 47–59.
- Hernández; R., Fernández C., & Baptista P. (2006). *Metodología de la investigación*. (5ta. Ed.). Mexico: D.F.

Holleran, S. (2010). *The Early Detection of Depression from Social Networking Sites*. Recuperado de: <http://hdl.handle.net/10150/196085>.

Houston TK, Cooper LA, Thi Vu H, Kahn J, Toser J, Ford DE. *Screening the public for depression through the internet*. *Psychiatric Services* 2001; 52:362–7

Kubat, M. (1998). *Machine Learning and Data Mining: Methods and Applications*. ISBN: 978-0-471-97199-3

Longbing Cao. 2017. *Data science: A comprehensive overview*. *ACM Comput. Surv.* 50, 3, Article 43 (June 2017), 42 pages. DOI: <http://dx.doi.org/10.1145/3076253>

Maglio, Norma, Fatelevich, Marisa, Luque, Adriana Edith, Biasella, Rogelio y Melillo, Oscar (2010). *Frases incompletas de J. Sacks y S. Levy: Relación materna*. II Congreso Internacional de Investigación y Práctica Profesional en Psicología XVII Jornadas de Investigación Sexto Encuentro de Investigadores en Psicología del MERCOSUR. Facultad de Psicología - Universidad de Buenos Aires, Buenos Aires.

Murphy, K (2012). *Machine Learning: A Probabilistic Perspective*. *Adaptive Computation and Machine Learning Series* (24 de agosto de 2012). ISBN-10: 0262018020

Microsoft (2018). *Azure Machine Learning Studio*. Recuperado de: <https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-ml-studio>

Ministerio de Salud (2016). *Minsa: Cerca de 1000 muertes por suicidio pueden evitarse por año si se detecta oportunamente*. Recuperado de:

<https://www.gob.pe/institucion/minsa/noticias/14854-minsa-cerca-de-1000-muertes-por-suicidio-pueden-evitarse-por-ano-si-se-detecta-oportunamente>

Neumann N, Lundin D, Poole AM (2010) *Comparative Genomic Evidence for a Complete Nuclear Pore Complex in the Last Eukaryotic Common Ancestor*. PLoS ONE5(10): e13241. <https://doi.org/10.1371/journal.pone.0013241>

Oficina de Estadística e Informática del Instituto Nacional de Salud Mental (2016). *Compendio Estadístico*. Recuperado de: <http://www.insm.gob.pe/oficinas/estadistica/compendio.html>

Organización Mundial de la salud (2017). *Día Mundial de la Salud: 6 señales para detectar si una persona padece depresión*. Recuperado de: <http://www.who.int/campaigns/world-health-day/2017/event/es/>

Organización Mundial de la salud (2017). *Hablemos de depresión*. Recuperado de: <http://http://www.who.int/campaigns/world-health-day/2017/event/es/>

Otiniano Campos, Fiorella (2012), *Validez de constructo y eficacia diagnóstica de las escalas depresión mayor y trastorno de ansiedad del inventario clínico multiaxial de Millon III (MCMII-III)*. (Para optar por el título de licenciado en psicología, Pontificia Universidad Católica del Perú). (Acceso 2 de Octubre del 2017).

Radio Programas del Perú (11 de Setiembre del 2016). *Cada año pueden evitarse mil suicidios con la ayuda de familiares o amigos*. Recuperado de: <http://vital.rpp.pe/expertos/la-depresion-es-el-trastorno-mas-comun-en-el-peru-noticia-971106>

Radio Programas del Perú (14 de Junio del 2016). *Depresión: alrededor de 1 millón 700 peruanos la sufren*. Recuperado de: <http://http://vital.rpp.pe/expertos/la-depresion-es-el-trastorno-mas-comun-en-el-peru-noticia-971106>

Ruiz de Adana (1 de Mayo del 2009). *Eficacia de una prueba diagnóstica: parámetros utilizados en el estudio de un test*. Recuperado de: http://www.jano.es/ficheros/sumarios/1/0/1736/30/00300032_LR.pdf

Smith (2006). *Data science as an academic discipline*. *Data Science Journal* 5 (2006), 163–164.

Srivastava, A.N. and Sahami, M. (2009). *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC. ISBN:1420059408 9781420059403

Stephen Robertson, (2004). "Understanding inverse document frequency: on theoretical arguments for IDF", *Journal of Documentation*, Vol. 60 Issue: 5, pp.503-520, <https://doi.org/10.1108/00220410410560582>

Torres, B. (2006). *Población Muestra y Muestreo*. (1^a. ed.). España. ÍNDIGO

Van, Fuentes, Jordan, Newton, Ray, Temple y Wickham (2015). *ASA Statement on the Role of Statistics in Data Science*. Recuperado de: <http://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/>.

Van, G. (2009). *El tutorial de python*. Copyright © Python Software Foundation. Recuperado de: <http://python.org.ar/pyar/Tutorial>

Volkova, Bachrach y Van Durme (2017). *Mining User Interests to Predict Perceived Psycho-Demographic Traits on Twitter*. IEEE Second International Conference on. Oxford, UK. Electronic ISBN: 978-1-5090-2251-9 Recuperado de: <http://doi.org/10.1109/BigDataService.2016.28>.

Vuotto, Bogetti y Fernández (2015). *Aplicación del factor TF-IDF en el análisis semántico de una colección documental*. Revista de Bibliotecología y Ciencias de la información. ISSN 1562-4730. Recuperado de:

<https://doi.org/10.5195/biblios.2015.227>

Weiss, S.M., Indurkha, N., Zhang, T. and Damerou, F. (2005). *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer.

Wu, J. (1997). *Statistics = Data Science?* Recuperado de:
<http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf>.

Yau, N. (2009). *Rise of the Data Scientist*. Recuperado de:
<http://flowingdata.com/2009/06/04/rise-of-thedata-scientist/>.

ANEXOS

Anexo 01: Matriz de consistencia

DETECCIÓN DE DEPRESIÓN A TRAVÉS DEL ANÁLISIS TEXTUAL UTILIZANDO APRENDIZAJE AUTOMÁTICO, 2017				
PLANTEAMIENTO DEL PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLES	INDICADORES
Pregunta general	Objetivo general	Hipótesis general	Independiente	
¿Cómo detectar la depresión?	Detectar la depresión.	El uso del análisis textual y el aprendizaje automático en una aplicación web ayudará a detectar la depresión.	<p>X1: Análisis textual y aprendizaje automático en una aplicación web</p> <p>Uso del análisis textual y aprendizaje automático en una aplicación web=1</p> <p>No uso del análisis textual y aprendizaje automático en una aplicación web=0</p>	
Preguntas específicas	Objetivos específicos	Hipótesis específicas	Dependiente	I1: Tiempo de Atención de forma
1. ¿Cómo identificar nuevos	1. Identificar nuevos patrones	1. El uso del análisis textual y el		

patrones en individuos considerados con factores de riesgo de depresión?	en individuos considerados con factores de riesgo de depresión.	aprendizaje automático en un modelo predictivo permitirá identificar nuevos patrones en individuos considerados con factores de riesgo de depresión.	Y1: Detección de la depresión	oportuna a probables pacientes con depresión.
2. ¿De qué manera identificar de forma eficaz a probables pacientes con depresión?	2. Identificar de forma eficaz a probables pacientes con depresión.	2. El uso del análisis textual y el aprendizaje automático en una aplicación web identificar de forma eficaz a probables pacientes con depresión.		I2: Número de patrones identificados.
3. ¿Cómo atender de forma oportuna a probables pacientes con depresión?	3. Permitir atender de forma oportuna a probables pacientes con depresión.	3. El uso del análisis textual y el aprendizaje automático en una aplicación web permitirá atender de forma oportuna a probables pacientes con depresión.		I3: Número de pacientes detectados.

Fuente: Elaboración Propia

Anexo 02: Inventario de Depresión Beck II

BDI II		
<u>NOMBRE</u>	<u>EDAD</u>	<u>CARRERA</u>
<p>INSTRUCCIONES: En este cuestionario aparecen varios grupos de afirmaciones. Por favor, lea con atención cada una. A continuación, señale cuál de las afirmaciones de cada grupo describe mejor cómo se ha sentido durante esta última semana, incluido en el día de hoy. Si dentro de un mismo grupo, hay más de una afirmación que considere aplicable a su caso, márquela también. Asegúrese de leer todas las afirmaciones dentro de cada grupo antes de efectuar la elección</p>		
1)	5)	
<input type="checkbox"/> No me siento triste	<input type="checkbox"/> No me siento especialmente culpable.	
<input type="checkbox"/> Me siento triste.	<input type="checkbox"/> Me siento culpable en bastantes ocasiones.	
<input type="checkbox"/> Me siento triste continuamente y no puedo dejar de estarlo.	<input type="checkbox"/> Me siento culpable en la mayoría de las ocasiones.	
<input type="checkbox"/> Me siento tan triste o tan desgraciado que no puedo soportarlo.	<input type="checkbox"/> Me siento culpable constantemente.	
2)	6)	
<input type="checkbox"/> No me siento especialmente desanimado respecto al futuro.	<input type="checkbox"/> No creo que esté siendo castigado.	
<input type="checkbox"/> Me siento desanimado respecto al futuro.	<input type="checkbox"/> Me siento como si fuese a ser castigado.	
<input type="checkbox"/> Siento que no tengo que esperar nada.	<input type="checkbox"/> Espero ser castigado.	
<input type="checkbox"/> Siento que el futuro es desesperanzador y las cosas no mejorarán.	<input type="checkbox"/> Siento que estoy siendo castigado.	
3)	7)	
<input type="checkbox"/> No me siento fracasado.	<input type="checkbox"/> No estoy decepcionado de mí mismo.	
<input type="checkbox"/> Creo que he fracasado más que la mayoría de las personas.	<input type="checkbox"/> Estoy decepcionado de mí mismo.	
<input type="checkbox"/> Cuando miro hacia atrás, sólo veo fracaso tras fracaso.	<input type="checkbox"/> Siento vergüenza de mí mismo.	
<input type="checkbox"/> Me siento una persona totalmente fracasada.	<input type="checkbox"/> Me detesto.	
4)	8)	
<input type="checkbox"/> Las cosas me satisfacen tanto como antes.	<input type="checkbox"/> No me considero peor que cualquier otro.	
<input type="checkbox"/> No disfruto de las cosas tanto como antes.	<input type="checkbox"/> Me autocrítico por mis debilidades o por mis errores.	
<input type="checkbox"/> Ya no obtengo una satisfacción auténtica de las cosas.	<input type="checkbox"/> Continuamente me culpo por mis faltas.	
<input type="checkbox"/> Estoy insatisfecho o aburrido de todo.	<input type="checkbox"/> Me culpo por todo lo malo que sucede.	

9)

- No tengo ningún pensamiento de suicidio.
- A veces pienso en suicidarme, pero no lo cometería.
- Desearía suicidarme.
- Me suicidaría si tuviese la oportunidad.

10)

- No lloro más de lo que solía llorar.
- Ahora lloro más que antes.
- Llora continuamente.
- Antes era capaz de llorar, pero ahora no puedo, incluso aunque quiera.

11)

- No estoy más irritado de lo normal en mí.
- Me molesto o irrito más fácilmente que antes.
- Me siento irritado continuamente.
- No me irrito absolutamente nada por las cosas que antes solían irritarme.

12)

- No he perdido el interés por los demás.
- Estoy menos interesado en los demás que antes.
- He perdido la mayor parte de mi interés por los demás.
- He perdido todo el interés por los demás.

13)

- Tomo decisiones más o menos como siempre he hecho.
- Evito tomar decisiones más que antes.
- Tomar decisiones me resulta mucho más difícil que antes.
- Ya me es imposible tomar decisiones.

14)

- No creo tener peor aspecto que antes.
- Me temo que ahora parezco más viejo o poco atractivo.
- Creo que se han producido cambios permanentes en mi aspecto que me hacen parecer poco atractivo.
- Creo que tengo un aspecto horrible.

15)

- Trabajo igual que antes.
- Me cuesta un esfuerzo extra comenzar a hacer algo.
- Tengo que obligarme mucho para hacer algo.
- No puedo hacer nada en absoluto.

16)

- Duermo tan bien como siempre.
- No duermo tan bien como antes.
- Me despierto una o dos horas antes de lo habitual y me resulta difícil volver a dormir.
- Me despierto varias horas antes de lo habitual y no puedo volverme a dormir.

17)

- No me siento más cansado de lo normal.
- Me canso más fácilmente que antes.
- Me canso en cuanto hago cualquier cosa.
- Estoy demasiado cansado para hacer nada.

18)

- Mi apetito no ha disminuido.
- No tengo tan buen apetito como antes.
- Ahora tengo mucho menos apetito.
- He perdido completamente el apetito.

19)

- Últimamente he perdido poco peso o no he perdido nada.
- He perdido más de 2 kilos y medio.
- He perdido más de 4 kilos.
- He perdido más de 7 kilos.
- Estoy a dieta para adelgazar SI / NO.

20)

- No estoy preocupado por mi salud más de lo normal.
- Estoy preocupado por problemas físicos como dolores, molestias, malestar de estómago o estreñimiento.
- Estoy preocupado por mis problemas físicos y me resulta difícil pensar algo más.
- Estoy tan preocupado por mis problemas físicos que soy incapaz de pensar en cualquier cosa.

21)

- No he observado ningún cambio reciente en mi interés.
- Estoy menos interesado por el sexo que antes.
- Estoy mucho menos interesado por el sexo.
- He perdido totalmente mi interés por el sexo

Anexo 03: Frases Incompletas de Sacks

TEST DE FRASES INCOMPLETAS DE SACKS

NOMBRE _____ EDAD _____ CARRERA _____

INSTRUCCIONES: A continuación hay sesenta frases incompletas. Lea cada una y complétela con lo primero que venga a su mente. Trabaje tan aprisa como le sea posible. En caso de que no pueda completar una, encierre el número correspondiente en un círculo y terminela después.

- 1.- Siento que mi padre raras veces _____
- 2.- Cuando tengo mala suerte _____
- 3.- Siempre anhelé _____
- 4.- Si yo estuviera a cargo _____
- 5.- El futuro me parece _____
- 6.- Las personas que están sobre mí _____
- 7.- Sé que es tonto pero tengo miedo de _____
- 8.- Creo que es un verdadero amigo _____
- 9.- Cuando era niño (a) _____
- 10.- Mi idea de mujer (hombre) perfecta (o) _____
- 11.- Cuando veo una hombre y a una mujer juntos _____
- 12.- Comparada con las demás familias, la mía _____
- 13.- En las labores me llevo mejor con _____
- 14.- Mi madre _____
- 15.- Haría cualquier cosa por olvidar la vez que _____
- 16.- Si mi padre tan sólo _____
- 17.- Siento que tengo la habilidad para _____
- 18.- Sería perfectamente feliz si _____
- 19.- Si la gente trabaja para mí _____
- 20.- Yo espero _____
- 21.- En la escuela, mis maestros _____
- 22.- La mayoría de mis amistades no saben que tengo miedo de _____
- 23.- No me gusta la gente con _____
- 24.- Antes _____
- 25.- Pienso que la mayoría de las muchachas (os) _____
- 26.- Yo creo que la vida matrimonial _____
- 27.- Mi familia me trata como _____

- 28.- Aquellos con los que trabajo _____
- 29.- Mi madre y yo _____
- 30.- Mi más grande error fue _____
- 31.- Desearía que mi padre _____
- 32.- Mi mayor debilidad es _____
- 33.- Mi ambición secreta en la vida _____
- 34.- La gente que trabaja para mí _____
- 35.- Algún día yo _____
- 36.- Cuando veo el jefe venir _____
- 37.- Quisiera perder el miedo de _____
- 38.- La gente que más me agrada _____
- 39.- Si fuera joven otra vez _____
- 40.- Creo que la mayoría de las mujeres (hombre) _____
- 41.- Si tuviera relaciones sexuales _____
- 42.- la mayoría de las familias que conozco _____
- 43.- Me gusta trabajar con la gente que _____
- 44.- Creo que la mayoría de las madres _____
- 45.- Cuando era más joven me sentía culpable de _____
- 46.- Siento que mi padre es _____
- 47.- Cuando la suerte se vuelve en contra mía _____
- 48.- Cuando doy órdenes, yo _____
- 49.- Lo que más deseo en la vida _____
- 50.- Dentro de algún tiempo _____
- 51.- La gente a quien yo considero mis superiores _____
- 52.- Mis temores en ocasiones me obligan a _____
- 53.- Cuando no estoy, mis amigos _____
- 54.- Mi más vivido recuerdo de la infancia _____
- 55.- Lo que menos me gusta de las mujeres (hombres) _____
- 56.- Mi vida sexual _____
- 57.- Cuando era niño (a), mi familia _____
- 58.- La gente que trabaja conmigo, generalmente _____
- 59.- Me agrada mi madre pero _____
- 60.- La peor cosa que he hecho _____

Anexo 04: Acuerdo de Confidencialidad

ACUERDO DE CONFIDENCIALIDAD

Al objeto de garantizar la confidencialidad de la presente investigación, se hace necesario la firma de un acuerdo que garantice unos niveles de confianza entre las partes. Para lo cual se elaborará un documento que consta de los siguientes acuerdos:

DE UNA PARTE: Dirección Académica, en su nombre y representación **Dra. Pilar Jiménez Berrú**, Dirección de Bienestar Universitario, en su nombre y representación **Dr. Edgard Rivero Contreras** en calidad de directores de las presentes áreas mencionadas; aprueban que se realicen las pruebas de: "**Método BDI II**" y "**Test de Frases Incompletas Sacks**" el día **Viernes 06 de Octubre del 2017** a los alumnos de la **Universidad César Vallejo** situada en **Av. Argentina 1795, Callao**. Además indican que la toma de encuestas se encuentra a cargo de la **Ing. Ariana Orué Medina**, a la cual se asigna un equipo de apoyo y soporte conformado por la **Lic. Karen Flores Rivera** con su equipo de trabajo respectivo.

DE OTRA PARTE: **Ing. Orué Medina Ariana Maybee**, en calidad de investigadora y autora de la presente investigación referente a la "**Detección de Depresión aplicando aprendizaje automático**" y Jefe de Practicas de la **Universidad César Vallejo** situada en **Av. Argentina 1795, Callao**; indica que se hace responsable de la información física y digital de los encuestados, el día **Viernes 06 de Octubre del 2017**. Dicha información física fue recibida personalmente y proporcionada por el equipo de trabajo perteneciente a Dirección de Bienestar Universitario de la misma universidad.

Reunidos en la **Universidad César Vallejo** situada en **Av. Argentina 1795**, a **Viernes 06 de Octubre de 2017**.

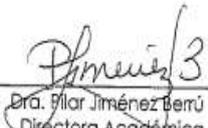
EXPONEN

I - Que las partes, anteriormente citadas, están interesadas en el desarrollo de la investigación y disponen que la confidencialidad de los datos que han sido recabados estén ante el manejo para fines de investigación de la Ing. Ariana Maybee Orué Medina.

II - Que las partes, en virtud de lo anteriormente expuesto, convinieron que el presente Acuerdo de Confidencialidad se rija por la normativa aplicable al efecto y, en especial que la investigación realizada; se presente la muestra de encuestas con una codificación en los nombres completos de los encuestados.

Se firma el presente acuerdo en señal de conformidad.

Callao, 05 de Octubre del 2017.


Dra. Pilar Jiménez Berrú
Directora Académica
Universidad César Vallejo-Callao




Ing. Ariana Orué Medina
Jefe de Practicas e Investigadora
Universidad César Vallejo-Callao


Dr. Edgard Rivero Contreras
Director de Bienestar Universitario
Universidad César Vallejo-Callao

