

**Diseño e implementación de una metodología estructurada para el análisis de
decisiones aplicada al soporte del diagnóstico de trastorno depresivo en
estudiantes universitarios**

Nicolás Mejía Martínez

Asesor: Mario Castillo Hernández

Maestría en Ingeniería Industrial

Facultad de Ingeniería

Universidad de los Andes

Bogotá D.C, Colombia

2019

Contenido

1	Introducción	4
2	Marco Conceptual	6
2.1	Diagnóstico y Tratamiento de la depresión.....	6
2.2	Escalas.....	7
2.2.1	Escala de Depresión del Centro de Estudios Epidemiológicos (CES-D10)	8
2.3	Modelos de Clasificación.....	9
2.3.1	Regresión Logística	9
2.3.2	Análisis Discriminante	10
2.3.3	K-Nearest Neighbor.....	11
2.3.4	Árboles de Clasificación.....	11
2.3.5	Random Forest.....	12
2.3.6	Redes Bayesianas	13
2.4	Teoría de la Decisión	14
3	Revisión de literatura	16
4	Situación actual de la depresión en Colombia y el mundo.	21
4.1	Prevalencia en la Población general.....	21
4.2	Prevalencia en la población universitaria	23
5	Diseño de la metodología específica	25
6	Aplicación de la metodología específica.....	27
6.1	Selección del grupo de expertos.....	27
6.2	Definición de las variables relevantes	27
6.3	Diseño y formulación del instrumento de medición.	28
6.4	Obtención y toma de datos.....	29
6.5	Análisis estadístico de los resultados de la encuesta.	30
6.5.1	Comparación de la muestra contra la población real de la universidad	30
6.5.2	Análisis Descriptivo de la Muestra.....	31
6.5.3	Pruebas de Hipótesis.....	35
6.6	Definición de los modelos relevantes.....	37
6.7	Construcción de los modelos.....	37
6.7.1	Regresión Logística	38
6.7.2	Discriminante Lineal	38
6.7.3	Classification Tree.....	38
6.7.4	Random Forest.....	39
6.7.5	Support Vector Machine.....	39
6.7.6	AdaBoost	39
6.7.7	Gradient Boosting.....	39
6.7.8	Xtreme Gradient Boosting.....	39
6.7.9	K-Nearest Neighbours	40
6.7.10	Naive Bayes.....	40

7	Análisis de Resultados	41
7.1	Importancia de las Variables.....	41
7.1.1	Regresión Logística	41
7.1.2	Classification Tree	42
7.1.3	Random Forest.....	44
7.1.4	Support Vector Machine Lineal	45
7.1.5	Adaptative Boosting	45
7.1.6	Gradient Boosting.....	46
7.1.7	Xtreme Gradient Boosting.....	46
7.2	Calidad Predictiva de los Modelos	48
7.3	Stacking	51
7.3.1	Clasificador Final Votación.....	53
7.3.2	Stacking mediante XGBoosting	54
7.4	Comparación.....	55
7.5	Identificación de grupos de riesgo.....	56
8	Conclusiones y Recomendaciones	61
9	Bibliografía.....	63

1 Introducción

La depresión es un trastorno psíquico que es entendido como aquella desregulación disruptiva del estado de ánimo que puede verse acompañada por estados de ansiedad, donde interactúan una serie de síntomas afectivos, cognitivos, volitivos o somáticos, como lo son la tristeza patológica, la inhibición, sensación de vacío y desinterés general y/o decaimiento, disminución de la comunicación y del contacto social, alteraciones del apetito y el sueño (insomnio tardío) agitación o enlentecimiento psicomotor, irritabilidad, sentimientos de culpa y de incapacidad, ideas de muerte e incluso intento de suicidio (Garcia, 2009).

Este trastorno llega a afectar un estimado de 1 en 15 adultos (6,7%) en un año; adicionalmente, se estima que una de cada seis personas experimentara depresión en algún momento de su vida. La depresión puede manifestarse en cualquier momento a lo largo de la vida de una persona, pero en promedio aparece durante la adolescencia tardía, hacia los 20 años.

Según datos de la OMS, la prevalencia de la depresión alcanza el 8% en las personas jóvenes. Sin embargo, solo un pequeño porcentaje es diagnosticado debido a la falta de capacitación en el personal de salud o al desconocimiento de los síntomas por parte de los pacientes. Así mismo, se menciona que la depresión podría llegar a ser considerada en el año 2020 la segunda causa de discapacidad en países industrializados (Organización Panamericana de la Salud, 1998).

Dentro de la población joven, los estudiantes universitarios han llamado particularmente la atención de los clínicos e investigadores, debido a la cantidad de estímulos y variables a las que están expuestos, que se ven asociadas directamente con un estado de baja salud mental. Del mismo modo, la salud mental del individuo afecta directamente las condiciones con las que este se relaciona con su entorno, determinando, hasta cierto punto su desempeño académico y social.

Las condiciones de salud mental de los estudiantes universitarios pueden verse afectadas en la medida en que las exigencias académicas, las dificultades en el desarrollo del aprendizaje y la adquisición de destrezas plantean situaciones de fracaso o éxito, que acompañadas con

el ritmo de estudio, la intensidad horaria, la exigencia de los docentes, las presiones de grupo, la competitividad entre compañeros, los cambios en horarios de sueño y alimentación pueden alterar esas condiciones de salud, desencadenando en problemas psicológicos que afectan en mayor medida su vida académica y personal (Gutiérrez Rodas, y otros, 2010). En la población universitaria, debido a estos factores mencionados, los principales motivos de consulta en atención psicológica que ofrece las oficinas de bienestar universitario en el país, lo constituyen la depresión y la ansiedad.

Siguiendo esta línea de investigación muchos estudios psicológicos como el de Arco, López, Heilborn, & Fenández (2005) o el de Campo, Díaz, Rueda, & Barros (2005) se han centrado en la población universitaria, encontrando en esta una alta prevalencia de la depresión y el efecto adverso que tiene sobre variables como: el desempeño académico y el bienestar emocional del individuo.

En este sentido, un diagnóstico temprano y a tiempo posee un gran impacto en el manejo apropiado de la enfermedad, junto con las morbilidades que pueden resultar como producto de esta. El tiempo y la cantidad de individuos dificultan a los psicólogos y psiquiatras la identificación de la enfermedad en la totalidad de pacientes. De igual manera, al ser la depresión es una enfermedad coloquial puede llegar a camuflarse ante lo que se entiende como tristeza pese a la presencia de la sintomatología clínica.

Con esto en mente, resulta relevante el desarrollo de una metodología estructurada, soportada en modelos matemáticos robustos y debidamente validados, que apoye el diagnóstico temprano de la depresión, la cual, junto con el concepto de un profesional, permita la identificación temprana y adecuada de un mayor número de pacientes que presenten la enfermedad, la identificación de casos puntuales en necesidad urgente de atención especializada, y que sirva de base en el diseño de planes de acción enfocados a la concientización y mitigación de la prevalencia de la enfermedad sobre la población universitaria.

2 Marco Conceptual

2.1 Diagnóstico y Tratamiento de la depresión

El diagnóstico de la depresión es a menudo llevado a cabo por el médico de cabecera, en el caso de la depresión leve y moderada, o el psicólogo o psiquiatra en el caso de la depresión grave con síntomas psicóticos.

Si bien la depresión es generada por una alteración de los neurotransmisores cerebrales, no existe ninguna prueba biológica contundente que confirme de manera fiable el diagnóstico del trastorno (San Molina & Arranz Martí, 2010). La ausencia de técnicas diagnósticas objetivas hace que el diagnóstico de depresión se realice teniendo en cuenta, exclusivamente, la información que el médico obtiene durante la entrevista con el paciente, y con familiares o personas allegadas (Hospital Clínic de Barcelona, 2007).

Estar deprimido puede entenderse en tres niveles diferentes, cada uno de los cuales posee su propia aproximación a la evaluación y diagnóstico. Existe el nivel sintomatológico (referente únicamente al estado de ánimo decaído, el nivel sindrómico (el cual involucra una serie de síntomas que acompañan el ánimo decaído), y el nivel nosológico (en el cual los síntomas se deben a otro trastorno o condición). Cada nivel, posee su lógica de evaluación diferente. Cuando se busca medir el nivel de tristeza, con la presencia de otros síntomas, normalmente se utilizan cuestionarios y escalas que permiten cuantificar la gravedad de los síntomas durante un continuo. Cuando se trata de efectuar un diagnóstico, la aproximación adecuada es, como se mencionó anteriormente, comprobar si se cumplen una serie de requisitos diagnósticos, mediante la entrevista con el paciente (Vázquez, Hernangómez, Hervás, & Nieto, 2005).

En este sentido, la línea clásica de diagnóstico psicológico se basa principalmente en la entrevista llevada a cabo por el profesional donde se evalúa en el paciente la presencia de síntomas reportados por los manuales de diagnóstico. El Manual de diagnóstico y estadístico de los trastornos mentales (DSM) presenta una serie ordenada de síntomas y criterios que sirven de guía durante las entrevistas psicológicas. Para el trastorno depresivo mayor, se enuncian como variables relevantes: estado de ánimo depresivo la mayor parte del día, disminución del interés o de la capacidad para el placer en todas o casi todas las actividades, cambios importantes de peso sin hacer régimen, insomnio o hipersomnia,

sentimientos de inutilidad o de culpa, disminución de la capacidad para pensar o concentrarse y pensamientos recurrentes de muerte (Hospital Clínic de Barcelona, 2007). Esta sintomatología es ampliamente aceptada en la literatura psicológica y médica, y como se verá más adelante, resulta ser la base de los modelos estadísticos que buscan generar un diagnóstico de la enfermedad.

Hay diversos tipos de depresión, y el tratamiento variará de un tipo a otro y, más importante, de un paciente a otro, por lo cual no es posible generalizar las formas de tratamiento en su totalidad. Sin embargo, parece existir un consenso médico en el tratamiento general de tres tipos fundamentales del trastorno depresivo (Hospital Clínic de Barcelona, 2007).

Por una parte, se encuentra el trastorno depresivo mayor, que es una enfermedad con un componente neurobiológico muy importante, lo cual justifica la prelación del tratamiento farmacológico sobre el psicológico. En segundo lugar, está el trastorno depresivo adaptativo, en el cual existe una dificultad para afrontar una situación estresante real, por lo cual el tratamiento psicológico es especialmente relevante. Por último, en el trastorno depresivo persistente (distimia), existiría una causa biológica junto a aspectos de carácter que limitan la capacidad del paciente para afrontar los diversos conflictos del día a día. Por este motivo, el tratamiento combinado farmacológico y psicológico suele ser el indicado.

2.2 Escalas

Los instrumentos de medición, como las escalas, se utilizan para cuantificar atributos, cualidades o propiedades, constructos o conceptos completamente teóricos, que es imposible medir o cuantificar de otra forma (Bland, 2002).

Las escalas, como todos los instrumentos de medición, deben mostrar altos valores de validez y de confiabilidad, para poder ser utilizados y que los resultados obtenidos sean acordes a la realidad. La validez alude la capacidad del instrumento de medir el constructo que pretende cuantificar y la confiabilidad, a la propiedad de mostrar resultados similares, libre de error, en repetidas mediciones (Campo-Arias & Oviedo, 2008).

2.2.1 Escala de Depresión del Centro de Estudios Epidemiológicos (CES-D10)

El CES-D es una escala corta diseñada para medir sintomatología depresiva en la población general. Las preguntas de la escala revisan síntomas asociados con la depresión que han sido ampliamente utilizadas en escalas más largas y previamente validadas. La escala tiene una alta consistencia interna y una adecuada confiabilidad (Sawer Radloff, 1977).

Esta escala fue desarrollada para su uso en estudios de epidemiología de la sintomatología depresiva, en la población general. Su propósito difiere de otras escalas de depresión, las cuales son principalmente usadas para el diagnóstico clínico o para evaluar la evolución de la severidad de la enfermedad a lo largo del tratamiento (Sawer Radloff, 1977). La escala CES-D fue diseñada para medir el nivel actual de sintomatología depresiva, haciendo énfasis en el componente cognitivo y conductual, con menor presencia de los síntomas de tipo somático (González Llana & Martín Carbonell, 2007)

La escala está compuesta por 10 ítems que pueden puntuar individualmente de 0 a 3; de esta manera el cuestionario puede presentar un puntaje dentro del rango de 0 a 30 puntos. Se aplica de manera auto administrada, seleccionando la respuesta que mejor refleje la situación en la semana previa. Los ítems están formulados en sentido negativo, exceptuando dos de ellos, con el fin de controlar el sesgo de respuesta. (Jaramillo-Toro, y otros, 2018)

De acuerdo con el puntaje obtenido se puede estimar si el paciente presenta o no síntomas depresivos. Cuando la escala presenta un puntaje mayor a 10, lo cual es equivalente a presentar seis de los síntomas durante la mayoría de la semana previa, el individuo es categorizado como con “Presencia de Síntomas Depresivos”, mientras que para puntajes menores o iguales a 10, el individuo se categoriza como con “Ausencia de Síntomas Depresivos”.

Según Andresen & Malmgren (1993) y el estudio inicial realizado por Radloff (1977) este punto de corte tiene un buen desempeño, pues genera pocos falsos positivos y una alta consistencia interna y confiabilidad.

2.3 Modelos de Clasificación

Es muy común en la práctica encontrarse con problemas en los que la variable de interés no es cuantitativa, sino categórica. Son problemas que únicamente aceptan como variable Y valores discretos no ordenados. En este sentido, la tarea de predecir una respuesta cualitativa es normalmente llamada como clasificar dicha observación, ya que busca asignarla a una categoría o clase (James, Witten, Hastie, & Tibshirani, 2017). En este sentido, los problemas de clasificación buscan identificar las características que indican a qué grupo o categoría pertenece cada observación. Esto puede utilizarse tanto para entender el comportamiento de los datos o para predecir cómo se comportarán nuevas observaciones. Mediante el uso de técnicas matemáticas y estadísticas es posible crear modelos de clasificación mediante el análisis de datos ya clasificados y encontrando un patrón predictivo (Imandoust & Bolandraftar, 2013).

Puesto que la variable de respuesta toma valores discretos dentro de un conjunto, es posible dividir el espacio generado por los predictores en un set de regiones con marcas acordes a las categorías de clasificación. Los límites de decisión que separan las categorías, pueden tener diversas características y dependen del algoritmo o clasificador que se utilice, entre los más conocidos están los modelos lineales como la Regresión Logística, Análisis Discriminante, *K Nearest Neighbor*, *Support Vector Machines*; modelos basados en la construcción de separaciones lógicas de manera secuencial como los árboles de clasificación y sus derivados (*Randon Forest* y *Bagging*) y modelos puramente estadísticos, cuya base es un modelo explícitamente probabilístico, como el caso de las Redes Bayesianas.

2.3.1 Regresión Logística

La regresión logística es una generalización de la regresión lineal para los casos en los que la variable de interés toma uno de dos posibles valores, los cuales representan un éxito o un fracaso o de manera más general, la presencia o ausencia de una característica de interés. En la regresión logística el interés recae sobre la modelación de la probabilidad de que una observación pertenezca a una clase dados los valores de las variables:

$$P(Y = 1|\bar{X}) = P(Y = 1|X = X_1, X_2, \dots, X_p)$$

En este modelo, Y es la variable binaria dependiente y \bar{X} representa el set completo de variables predictoras o explicativas. Dado que el modelo de regresión logística modela la relación entre las variables como una probabilidad, es necesario el uso de una función G , que garantice el cumplimiento de los axiomas de probabilidad para los resultados, es decir, G toma valores entre 0 y 1 para todos los números reales, representando la probabilidad de ocurrencia de la característica de interés, siendo G la función logística

$$P(Y = 1|\bar{X}) = G(\bar{X})$$

$$G(\bar{X}) = \ln\left(\frac{\pi(\bar{X})}{1 + \pi(\bar{X})}\right) = \beta_o + \beta_1 X_1 + \dots + \beta_n X_n$$

siendo $\pi(\bar{X})$ la probabilidad de ocurrencia de la característica de interés, dado un set de variables explicativas X

2.3.2 Análisis Discriminante

Análisis discriminante es una técnica paramétrica que busca determinar la mejor serie de ecuaciones que permita discriminar entre 2 o más grupos, con base en los valores de las variables predictores. Esta metodología, genera una función discriminante de una combinación lineal de pesos y puntajes de las variables predictoras. El número máximo de funciones discriminantes que se pueden obtener es el mínimo entre el número de predictores y el número de grupos menos uno (Ramayah, y otros, 2010).

La función discriminante tiene la siguiente forma general:

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + \dots + W_n X_{nk}$$

Donde Z_{jk} es el puntaje de la función discriminante j para la observación k , a es el intercepto, W_j el coeficiente discriminante del predictor j y X_{jk} es el valor del predictor j correspondiente a la observación k .

En general, existen tres principales usos para el análisis discriminante: explicar la relación entre las variables independientes y la variable categórica dependiente, hacer reducción de dimensionalidad, y predecir o clasificar una observación en una categoría particular.

2.3.3 K-Nearest Neighbor

El algoritmo *K-Nearest Neighbor* o KNN es un método para clasificar observaciones basado en los objetos de entrenamiento que estén más cerca a dichas observaciones en el espacio de características formado por los predictores. Se conoce como aprendizaje perezoso pues las funciones resultantes solo se aproximan de manera local, dado por el número de vecino que se quiera analizar.

KNN es el método de clasificación más básico cuando no se conoce de manera *a priori* la distribución de los datos con los que se va a trabajar. Para una nueva observación, la regla de decisión consiste en generar el espacio de características y obtener las distancias con todas las observaciones ya clasificadas dentro del espacio y de esta manera asignar la nueva observación a la clase mayoritaria dentro de los puntos más cercanos a esta (vecindario). En este método, cada observación deberá ser clasificada de manera muy similar a sus vecinos, de modo que si la clasificación no es conocida, esta podrá ser inferida por las categorías de sus vecinos más cercanos (Imandoust & Bolandraftar, 2013).

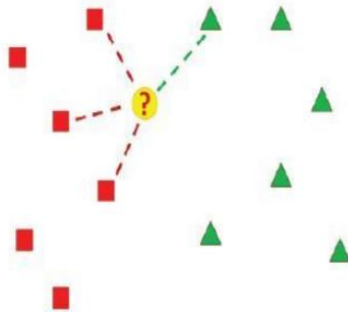


Ilustración 1: Algoritmo KNN Fuente: Imandoust & Bolandraftar (2013)

El desempeño del algoritmo dependerá en gran medida de la métrica de distancia utilizada y el número de observaciones a considerar en la determinación del vecindario.

2.3.4 Árboles de Clasificación

Árboles de clasificación y regresión (CART) es un algoritmo no paramétrico que utiliza datos históricos para crear particiones en el espacio de los predictores (o árboles de decisión), mediante el algoritmo de partición recursiva. Una característica atractiva de los árboles de clasificación es que debido a que el algoritmo genera preguntas o reglas de

decisión en una secuencia jerárquica, es relativamente simple de entender e interpretar (Kumar, 2011).

Para clasificación, CART utiliza una muestra de entrenamiento para organizar la base de datos en grupos, de acuerdo a las particiones generadas por el algoritmo recursivo. Busca en todas las variables y valores disponibles para encontrar la mejor partición, entendida como aquella que divide los datos en 2 grupos de manera que se maximice la heterogeneidad entre estos. Esto, de manera general, lo mide con respecto al *Gini Criterion* el cual permite la división de los datos basado en su homogeneidad (o heterogeneidad) mediante la impureza de cada división (Morgan, 2014), medido como el *Gini Index*:

$$i(t) = \sum_{i \neq j} p(i|t)p(j|t)$$

donde i y j son los niveles de la variable de interés y t representa la división. Así, $p(i|t)$ y $p(j|t)$ representan las probabilidades condicionales de que una observación pertenezca al niveles i o j , respectivamente, dada la división propuesta. En cada iteración se busca la división que maximice la suma de los índices.

El modelo CART no tiene problema en el manejo de variables numéricas o categóricas y no es sensible a los outliers ni datos faltantes.

2.3.5 Random Forest

Random Forest es un algoritmo no paramétrico de ensamblaje, que mejora el desempeño del clasificador y disminuye la varianza de las predicciones mediante el uso agregado de varios árboles de clasificación. *Random Forest* introduce aleatoriedad al algoritmo CART mediante la generación de árboles independientes entre sí con dos características fundamentales: la primera capa de aleatoriedad se introduce ya que para la creación de cada árbol se genera una muestra aleatoria de datos con reemplazo, proveniente de la muestra de entrenamiento original (*Bagging*); adicionalmente, para la creación de cada árbol, se selecciona una submuestra aleatoria de variables predictoras, de manera que cada árbol en el bosque es generado con un conjunto de predictores diferentes (Siwek & Osowski, 2016).

Cuando se ha generado el bosque completo (por default se generan 500 árboles y para cada uno se utiliza un total de \sqrt{p} variables, donde p es el número de predictores) la regla de decisión para determinar la clasificación de las observaciones en la muestra se da por voto mayoritario. Cada observación se evalúa en cada uno de los árboles generados y es asignada a la clase que cuente con la mayor cantidad de votos entregados por los árboles individuales.

Cada árbol del bosque es generado con el mismo algoritmo recursivo descrito para el método CART, con la diferencia que debido a que los árboles construidos en *Random Forest* utilizan un subconjunto aleatorio de todos los predictores disponibles, las posibles particiones a considerar se hacen únicamente sobre el subconjunto y no la totalidad de las variables de entrada (Liaw & Wiener, 2002).

2.3.6 Redes Bayesianas

Las redes bayesianas son un modelo matemático que modela las relaciones probabilísticas dentro de un set de variables (predictores). La estructura de la red bayesiana es un gráfico a cíclico dirigido, en el cual los nodos representan las variables, los arcos entre nodos representan las relaciones casuales entre variables y la falta de arcos representa la independencia condicional entre los predictores (Kotsiantis, 2007). Dado que los mapas causales representan el conocimiento de manera más descriptiva que otros modelos, se convierten en modelos de decisión más útiles, que sirven como puente de comunicación entre los usuarios y otras personas involucradas en el análisis de un problema de decisión determinado (Castillo Hernández, 2006).

Normalmente, la construcción de una red bayesiana puede ser dividida en dos tareas: inicialmente, se debe aprender la estructura causal de la red y luego determinar sus parámetros, en este caso las probabilidades condicionales resultantes del grafo. Los parámetros probabilísticos son consignados en tablas (una para cada variable) que contiene las distribuciones condicionales de cada variable dados sus padres. Dada las independencias condicionales consignadas en la red, la probabilidad conjunta puede ser reconstruida con la multiplicación de estas tablas (Kotsiantis, 2007).

Estos modelos tienen sus orígenes en el intento de representar el conocimiento de expertos en áreas donde este conocimiento es incierto, ambiguo y/o incompleto. Por esta razón, de manera general, la estructura de la red se define con base en el criterio de expertos en el tema a tratar y se asume que es correcta. Cuando se tiene la estructura, las distribuciones condicionales se estiman usando la muestra de entrenamiento.

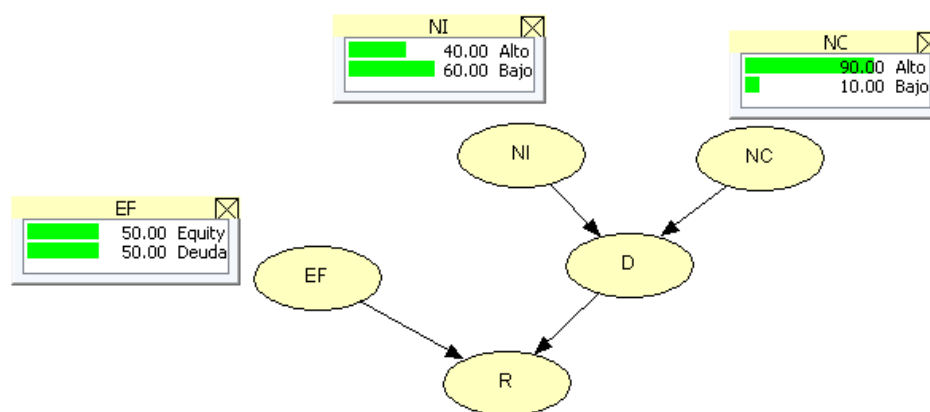


Ilustración 2: Diagrama de una red bayesiana. Elaboración propia

2.4 Teoría de la Decisión

Las decisiones estratégicas en las organizaciones son aquellas decisiones difícilmente reversibles, que no son replicables con facilidad, que involucran un componente de riesgo o incertidumbre, en las cuales usualmente transcurre un tiempo considerable entre el momento en que se toma la decisión y el momento en el cual se pueden observar sus resultados y como característica fundamental, las decisiones estratégicas representan un impacto importante sobre el futuro de la organización.

Dadas las fuertes implicaciones que tienen este tipo de decisiones, se hace necesario la utilización de un enfoque estructurado e integral que involucre al uso de metodologías y herramientas que soporten el proceso de decisión.

En este sentido, el análisis de decisiones es una disciplina que se ocupa de estudiar la estructura y las particularidades de un problema de decisión que comporta riesgo. El objetivo del análisis de decisiones es mejorar la calidad del proceso de toma de decisiones en las empresas, para lo cual es necesario contar con un marco conceptual apropiado, generar alternativas factibles, contar con información relevante y confiable, tener claridad

sobre los criterios de decisión y proveer soporte al proceso mediante el uso de un raciocinio lógico y modelos validos apropiados a la situación.

Castillo (2017) propone una metodología para el análisis de decisiones el cual puede ser usado de manera general sin importar la naturaleza u origen de la decisión a tratar. La metodología se basa en determinar en forma integral, cual es la situación problemática mediante su descripción y estructuración formal. La estructuración consiste en la identificación de todos los elementos clave en el problema como los aspectos, los actores que intervienen, las variables, las fuentes de incertidumbre, los criterios bajo los cuales se tomará la decisión y las alternativas que se consideran evaluar. Tomando como base el entendimiento integral del problema se busca la información requerida, de acuerdo a las variables involucradas. Cuando se tiene la información necesaria se procede al diseño de una metodología específica para el problema a tratar, lo cual consiste en el diseño de los pasos a seguir junto con los modelos a utilizar para obtener la solución del problema en cuestión. Construida la metodología, se aplica al caso específico de estudio y se analizan los resultados, para analizar la consistencia y robustez le la solución obtenida. Por último, y de acuerdo a los resultados obtenidos y su análisis, se formula un conjunto de conclusiones y recomendaciones acerca del problema, la metodología presentada y la implementación de la solución.

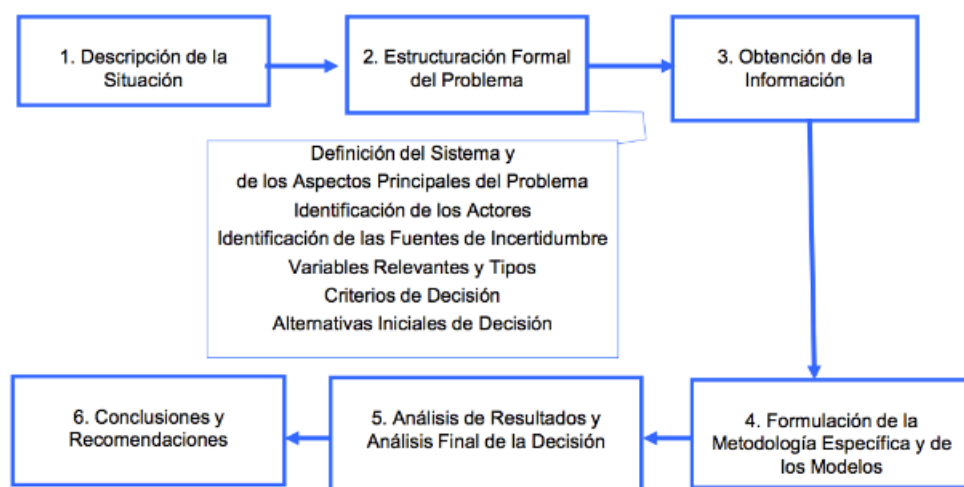


Ilustración 3: Metodología general para el análisis de decisiones. Fuente: Castillo (2017)

La metodología propuesta por Castillo (2017) pretende ser una guía general flexible, que puede ser adaptada a cada problema en particular. Por esta razón, y debido a que el problema a tratar en este documento no es propiamente un problema de decisión, se tomara como base los elementos de la metodología que aporten al planteamiento, entendimiento y desarrollo de los modelos necesarios para soportar el diagnóstico del trastorno depresivo mayor.

3 Revisión de literatura

Apartándose de la línea clásica de diagnóstico, la construcción de modelos estadísticos y computacionales estructurados para la identificación de síntomas y enfermedades en pacientes clínicos es un campo que se ha venido desarrollando durante los últimos años, debido a la cantidad de información disponible, los avances en la capacidad computacional y la importancia que cobra el diagnóstico temprano, para la prevención y el tratamiento. Sin embargo, el uso de este tipo de modelos, tiene preminencia en la literatura para la clasificación de enfermedades físicas, mientras que su utilización para la clasificación de enfermedades psicológicas esta hasta ahora empezando a tomar fuerza.

La literatura especializada presenta una diversidad de formas en las cuales se han abordado este tipo de problemas, pues se tienen una gran variedad de modelos a utilizar y así mismo, una gran cantidad de variables que pueden alimentar dichos modelos. En este sentido, la información disponible es bastante extensiva; para efectos de este documento se mostrarán los trabajos más relacionados con el tema y la metodología de investigación

En el campo de la detección de enfermedades psicológicas está el trabajo “*Classification of the most Significant Psychological Symptoms in Mental Patients with Depression using Bayesian Network*” de Galiatsatos, et. al (2015). Dicho trabajo, busca categorizar en modelos flexibles los síntomas más relevantes que presenten pacientes diagnosticados con depresión, de manera que asista a los profesionales en salud mental en el diagnóstico de sus pacientes, en menos tiempo de lo usual. Para ello desarrollan modelos de redes bayesianas las cuales se calibran mediante validación cruzada, debido a su flexibilidad y poder predictivo.

Para la construcción de los modelos, utilizaron los perfiles psicológicos de los estudiantes que habían atendido al consultorio psicológico de la Universidad de Patras (Grecia). Para el diagnóstico, los autores alimentan el modelo con variables demográficas como la edad, el género y el estado civil, junto con variables diagnósticas relacionadas con los cambios de humor como la falta de interés, problemas al dormir o comer, dificultad para concentrarse, entre otras. Debido a la gran cantidad de variables que tenían a su disposición, los autores, con ayuda de un grupo de psicólogos determinaron cuatro grupos de factores relevantes, que consistían en diferentes combinaciones de las variables mencionadas anteriormente, cada uno de los cuales fue utilizado para ajustar una red bayesiana.

El artículo encuentra que la mezcla de metodologías estadísticas junto con un conocimiento estructurado del tema a tratar puede proporcionar una herramienta lógica, útil y con un gran desempeño para el diagnóstico temprano de enfermedades mentales.

En la misma línea de investigación se encuentra el trabajo *“Automatic Diagnosis of Depressive Disorders using Ensemble Techniques”* de Ojeme, Akazue, & Nwalih (2016) el cual busca enfrentar las dificultades de estructurales que existen para el diagnóstico de enfermedades mentales en ciertas regiones de África. Para ello, investigan la utilización de distintos métodos de Machine Learning de manera separada y su posterior ensamble para mejorar las predicciones de diagnóstico.

Los datos utilizados por el estudio corresponden a 580 casos obtenidos del Hospital Universitario de Benin. Los casos relacionan el diagnóstico con 23 variables de interés, que son utilizadas posteriormente para la construcción de los modelos. Estas variables son en su mayoría sintomatologías relacionadas con el trastorno depresivo como la pérdida de placer, insomnio, pérdida de apetito, pérdida de energía, pasar por eventos estresantes y cambios en el peso, entre otras. También se consideran en menor medida variables demográficas como la edad, el género, el estado civil y la condición socioeconómica.

En cuanto a los modelos, los autores hacen uso cinco técnicas de aprendizaje estadístico para el diagnóstico automático de la depresión: Redes bayesianas, Redes neuronales, lógica difusa, *Support Vector Machines* y *K Nearest Neighbors*. Todos los modelos fueron calibrados por validación cruzada y comparados contra la misma muestra de test de manera individual. Una vez se tenían los modelos individuales, se utilizaron técnicas de ensamblaje

para aprovechar las fortalezas de cada uno de los algoritmos y combinarlas en un predictor con una mayor calidad diagnóstica.

El documento encuentra un aumento significativo en las métricas de desempeño predictivo, especialmente el área bajo la curva ROC, cuando se utilizan ensambles de modelos, en comparación con los modelos individuales.

El artículo “*Mobile cloud based depression diagnosis using Ontology and Bayesian network*” (2015) Chang, et. al. Los autores utilizan una metodología un poco alejada de los demás artículos en esta revisión. El trabajo se basa en la combinación de una ontología para entender el diagrama causal y relacional de la depresión, junto con una herramienta predictiva, una red bayesiana en este caso. La ontología resulta la herramienta principal en el estudio, pues mediante el conocimiento de expertos y la literatura, les permite a los autores estructurar el problema del diagnóstico y estudiar las posibles relaciones entre el tener o no depresión, y las variables sintomáticas frecuentemente citadas en la literatura, como; los sentimientos de frustración, culpa, ira, poca energía, fatiga, contemplación del suicidio, ansiedad, tristeza y problemas alimenticios y del sueño. Una vez se estructuran las relaciones de los variables incluidas en el estudio, se utiliza una red bayesiana como herramienta de inferencia para la estimación de la probabilidad de entrar en depresión. Con la estructura inferencia y las probabilidades entregadas por la red, los autores desarrollan una aplicación móvil, en el sistema operativo Android que le permite a los usuarios evaluar su estado emocional y proporcionar el nivel o la probabilidad de estar deprimido.

El trabajo utiliza una base de datos para inferir las probabilidades condicionales para poder evaluar la red, pero no realiza ningún tipo de validación en las métricas clásicas de calidad predictiva. Está enfocado, no a la construcción de un modelo perfecto, sino en la combinación de la ontología y la red bayesiana, subiendo el trabajo en la aplicación móvil para demostrar su utilidad.

Como último trabajo consultado en esta línea de investigación se encuentra el artículo “*Predicting Anxiety and Depression in elderly patients using machine learning technology*” de Sau y Bhakta (2016). Dicho trabajo se enfoca en la identificación de las dos enfermedades psicológicas más recurrentes en la población de tercera edad, la ansiedad y la depresión. El artículo se basa en la importancia del diagnóstico temprano para el adecuado

tratamiento, el cual, al tratarse de población de tercera edad, es bastante delicado. El fin del artículo consiste en la generación de un modelo predictivo adecuado, que sirva para el diagnóstico temprano en la India, donde existen grandes restricciones en la disponibilidad de especialistas en trastornos psicológicos.

La fuente para la construcción de los modelos es una base de datos de 510 pacientes geriátricos en el hospital de Kolkata. La base contiene información socio demográfica como la edad, el género, si sabe o no leer, el estado civil, si trabaja o no, estado socioeconómico, junto con variables diagnósticas como el historial familiar de ansiedad y depresión, insomnio, problemas visuales y/o auditivos y si tiene adicciones, entre otros. Debido a la gran cantidad de predictores que se consideran, los autores hacen uso de métodos de selección automática para reducir el conjunto de predictores a los más significativos; se prueban cinco métodos de selección y después de la comparación se reduce en conjunto de predictores de 20 a 11. Este conjunto de predictores es utilizado para alimentar de los modelos de clasificación binaria que consideran, siendo estos: Redes Bayesianas, Regresión logística, Redes Neuronales, *Naive Bayes*, *Random Forest*, *Random Tree* y *Sequential Minimal Optimization*.

Cada uno de los modelos es calibrado con validación cruzada y comparado mediante métricas como el porcentaje de predicción correcta, tasa de falsos positivos y falsos negativos y el área bajo la curva ROC. Adicionalmente, todos los modelos son comparados con la prueba HADS catalogada como estándar de oro para el diagnóstico de la depresión.

El artículo encuentra que la selección estructurada de predictores, junto con las técnicas apropiadas de clasificación puede entregar muy buenos resultados que incluso compiten con los test y métodos clásicos de diagnóstico.

Separándose del uso de variables socioeconómicas y sintomatológicas, se encuentran en la literatura trabajos de un nivel técnico más avanzado como “*Studying depression using imaging and machine learning*” de Patel, Khalaf y Aizenstein (2016). El estudio hace énfasis en la complejidad del diagnóstico de la depresión y la importancia de una identificación temprana para su apropiado tratamiento. Para responder a esta problemática, los autores hacen un barrido por artículos e investigaciones que hacen uso de datos anatómicos y psicológicos obtenidos de neuroimagenes (MRI) para construir modelos de

Machine Learning, que sean capaces de discernir entre pacientes con y sin depresión. Los artículos revisados hacen uso de bases de datos conformada por imágenes de diferentes regiones cerebrales de una muestra balanceada de pacientes que presentan depresión y pacientes sanos. Debido a la cantidad de información contenida en las imágenes, cada estudio utiliza distintas metodologías de reducción de predictores, como componentes principales, ANOVA o la utilización de información experta. Los datos son utilizados para entrenar distintos modelos que resultan de variaciones de las máquinas de vectores de soporte (SVM), los cuales varían en desempeño y dependiendo de la calidad de la información y la metodología usada para su construcción, con un porcentaje de aciertos desde 67.6% hasta 94.3%.

Los autores, realizan una adecuada recopilación de estudios, en los cuales analizan la diversidad de metodologías que existen para la construcción de modelos de diagnóstico. Sugiere, debido a la complejidad del problema, utilizar bases de datos con un número significativo de observaciones y algoritmos de optimización para la selección de parámetros, para mejorar la generalización de los modelos, junto con la diversificación de modelos a implementar, no solo variaciones de SVM.

De estos artículos se puede observar una tendencia en el uso de modelos estadísticos para la predicción de pacientes con el trastorno depresivo mayor, en especial de las Redes Bayesianas (que son utilizadas en todos los artículos), seguido de los modelos de caja negra como Redes Neuronales y Support Vector Machine.

De la misma manera, parece existir un consenso en las variables consideradas predictores útiles, separándolos en dos grandes categorías:

- Variables Socio-Demográficas: las cuales describen las características propias del individuo como la edad, género, estado civil.
- Variables Sintomatológicas: aquellas que son citadas en la literatura psicológica y en las pruebas de diagnóstico como factores de riesgo para el desarrollo de la depresión, como falta de apetito, insomnio, pérdida de placer, baja autoestima y experimentar eventos estresores, entre otros.

Esta síntesis permite la construcción de una guía para la selección preliminar de variables adecuadas a medir sobre la población de interés, así como información acerca de modelos y metodologías frecuentemente usadas por la academia para abordar y resolver problemas de diagnóstico psicológico, mediante la aplicación de modelos matemáticos y estadísticos.

4 Situación actual de la depresión en Colombia y el mundo.

4.1 Prevalencia en la Población general

Como se ha mencionado antes, se estima que el 10% de la población mundial padece algún trastorno mental. Más específicamente, el trastorno depresivo mayor llega a afectar un estimado de 1 en 15 adultos (6,7%) en un año; adicionalmente, se estima que una de cada seis personas experimentara depresión en algún momento de su vida, así mismo, entre 50 y 70% de las personas que presentan episodios depresivos, tienen una predisposición mayor a desarrollar uno nuevo en los próximos 5 años (Jaramillo-Toro, y otros, 2018). Esta prevalencia de la enfermedad mental genera un impacto significativo sobre la economía global por los costos que estos episodios generan sobre el sistema de salud en cuanto a terapias, consultas y fármacos los cuales se proyectan en 6 trillones de dólares para el año 2030 a nivel mundial (Bloom, Cafiero, E, S, & Bloom LR, 2012).

La depresión puede manifestarse en cualquier momento a lo largo de la vida de una persona, pero en promedio aparece durante la adolescencia tardía, hacia los 20 años.

Los trastornos mentales y del comportamiento afectan entre 10-15% de los niños y adolescentes de Latinoamérica, de los cuales menos de la mitad reciben atención especializada. Para Colombia, datos del segundo estudio nacional de salud y consumo de sustancias psicoactivas muestran una prevalencia de la depresión general del 25.1%, donde los jóvenes en el rango de 16 a 21 años muestran los más altos índices de depresión al ser comparados con los demás grupos poblacionales.

El Sistema Integral de Información de la Protección Social (SISPRO) reporta que las atenciones por depresión se han incrementado desde el año 2009 y el diagnóstico de depresión se realizó en 48.100 atenciones en 2015.

Año						
2009	2010	2011	2012	2013	2014	2015
Depresión Moderada						
17.711	18.534	21.633	26.860	28.467	38.638	36.584
Depresión Grave sin síntomas Psicóticos						
4.703	4.801	6.647	6.499	6.879	20.111	8.385
Depresión Grave con síntomas Psicóticos						
3.311	3.100	3.963	4.755	2.546	3.137	3.131

Tabla 1: Número de personas atendidas por depresión en Colombia de 2009 a 2015. Datos reportados por el SISPRO

Los casos de depresión grave con síntomas psicóticos son menos frecuentes y representan entre el 6 y el 13% de las atenciones por depresión. Según el boletín de salud mental, la mayor proporción de personas atendidas por depresión moderada o severa correspondió a mujeres (Ministerio de Salud, 2017). En 2015, el 70,4% de las personas atendidas con diagnóstico de depresión moderada fue de mujeres, mientras que el 29,6% restante fue de hombres.

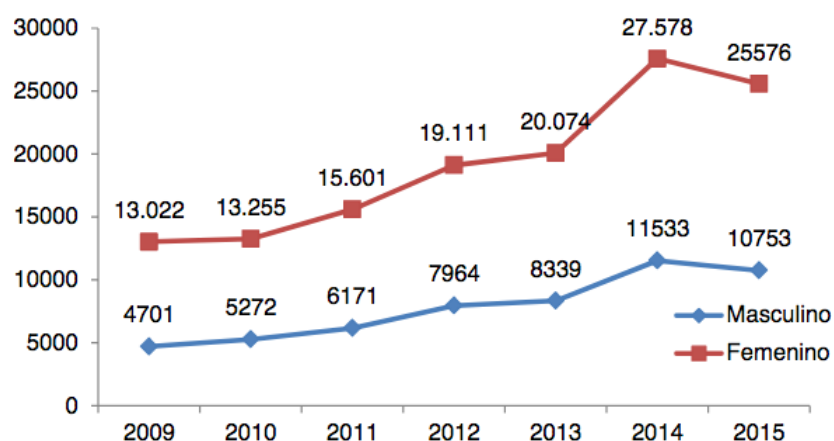


Ilustración 4: Número de personas atendidas por depresión moderada de 2009 a 2015. Fuente: Boletín de salud mental 2015

De la misma manera, el boletín reporta las tendencias del trastorno de acuerdo a los rangos de edad, donde los casos atendidos aumentaron a medida que aumentó la edad de las personas y alcanzaron su punto máximo en el rango de los 50 a 54 años de edad y luego la tendencia fue al descenso, para luego aumentar nuevamente de los 75 años en adelante.

Para el trastorno depresivo mayor, se encuentra que su prevalencia es pronunciada en dos rangos de edades: de 20 a 24 años y de 50 a 54 años.

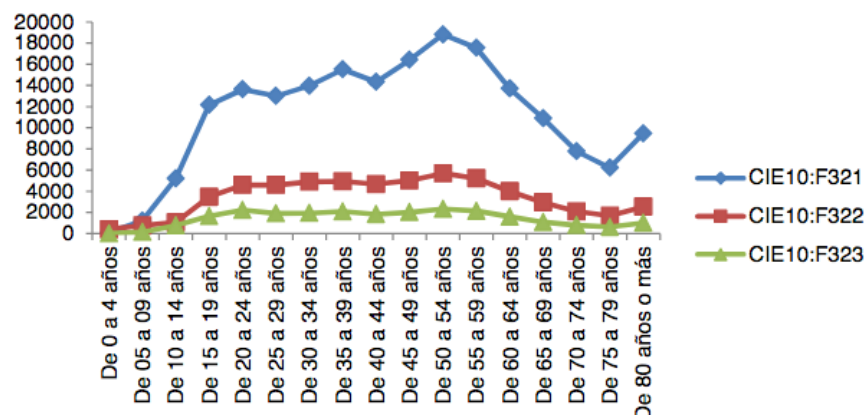


Ilustración 5: Número de personas atendidas por rango de edad. Fuente: Boletín de salud mental 2015

En cuanto a la distribución geográfica de los casos diagnosticados de depresión, se tiene que los departamentos de Antioquia, Bogotá y Valle del Cauca ocupan los primeros lugares, de manera independiente a la severidad del trastorno. Mientras que departamentos como Quindío, Santander, Atlántico y Nariño presentan un número elevado de casos en tipos específicos del trastorno depresivo (Ministerio de Salud, 2017) .

4.2 Prevalencia en la población universitaria

Adicional a los reportes nacionales, donde se reporta a los jóvenes como aquellos colombianos en los cuales se da una mayor prevalencia de la enfermedad, existen varios artículos académicos que exploran esta misma relación.

El trabajo Campo et. al (2005) presenta un estudio en la población universitaria femenina de la universidad autónoma de Bucaramanga donde mediante la aplicación de la escala de Zung, se pretende estudiar la prevalencia de la depresión en esta población específica. La escala de depresión de Zung es un instrumento psicológico auto administrado tipo Likert, que está compuesto por 20 ítems que presentan diversos síntomas asociados a la depresión, donde el individuo responde de acuerdo a la frecuencia con que ha experimentado cada uno de ellos. El estudio contó con una muestra de 110 universitarias, las cuales contestaron el cuestionario y posteriormente fueron citadas a una entrevista con un profesional con el fin

de validar lo expuesto en el instrumento de medición. El trabajo obtuvo como resultado un 43.6% de las entrevistadas presentabas síntomas depresivos con importancia clínica, resultado que fue validado con la entrevista estructurada.

También se encuentra el trabajo de Gutiérrez, et al (2010) el cual estudia la prevalencia del trastorno depresivo mayor de estudiantes matriculados en pregrado de la Universidad CES de Medellín y la relación que tiene la presencia de la enfermedad con la severidad del estrés generado, en los estudiantes, por diversos estresores académicos. La información de la depresión y el estrés fue capturada mediante instrumentos de medición como lo son la escala de Zung (para depresión) y la escala Dusi (para estrés). La población estudiada se compuso de 1344 estudiantes entre los 16 y 37 años de edad, de los cuales el 68.7% eran mujeres. El estudio encontró una prevalencia de la depresión de 47.2% dividida en un 29.9% de los estudiantes presentan depresión leve, un 14.2% de los estudiantes presentaron depresión moderada y un 3.2% presentaron depresión severa. Así mismo, se hace evidente la relación entre la severidad de la depresión y el nivel de estrés académico generado por dificultades en la universidad.

Por último, el trabajo de Jaramillo-toro et al. (2018) estudia la prevalencia de la depresión y los factores que pueden desencadenarla en la población universitaria, además de realizar una validación de dos instrumentos diagnósticos utilizados para la identificación de síntomas depresivos la escala de Zung y el CES-D. El estudio utiliza una muestra de 350 estudiantes de pregrado de la universidad tecnológica de Pereira. Los autores encuentran una prevalencia del 36.3% de la depresión en la población estudiada, dividida en: un 16.9% con depresión leve, un 14% con depresión moderada y un 5.4% con depresión severa. La presencia de la sintomatología estuvo relacionada a la inactividad física y el consumo de alcohol y sustancias psicoactivas.

Los estudios, fueron realizados de manera independiente, con diversos autores, con poblaciones universitarias de distintas partes del país, los cuales al utilizar muestras aleatorias significativas y encontrar resultados similares, entregan confiabilidad en los resultados encontrados. Así, los estudios concuerdan en una prevalencia de la depresión de un 45% aproximadamente sobre la población universitaria del país, lo cual al ser

comparada con la prevalencia en la población general (21.5%) representa una cifra alarmante.

5 Diseño de la metodología específica

De acuerdo con la información obtenida de la revisión de literatura y las herramientas descritas en el marco teórico, se propone la metodología específica que se define a continuación para abordar el problema del soporte de diagnóstico del trastorno depresivo mayor en estudiantes universitarios.

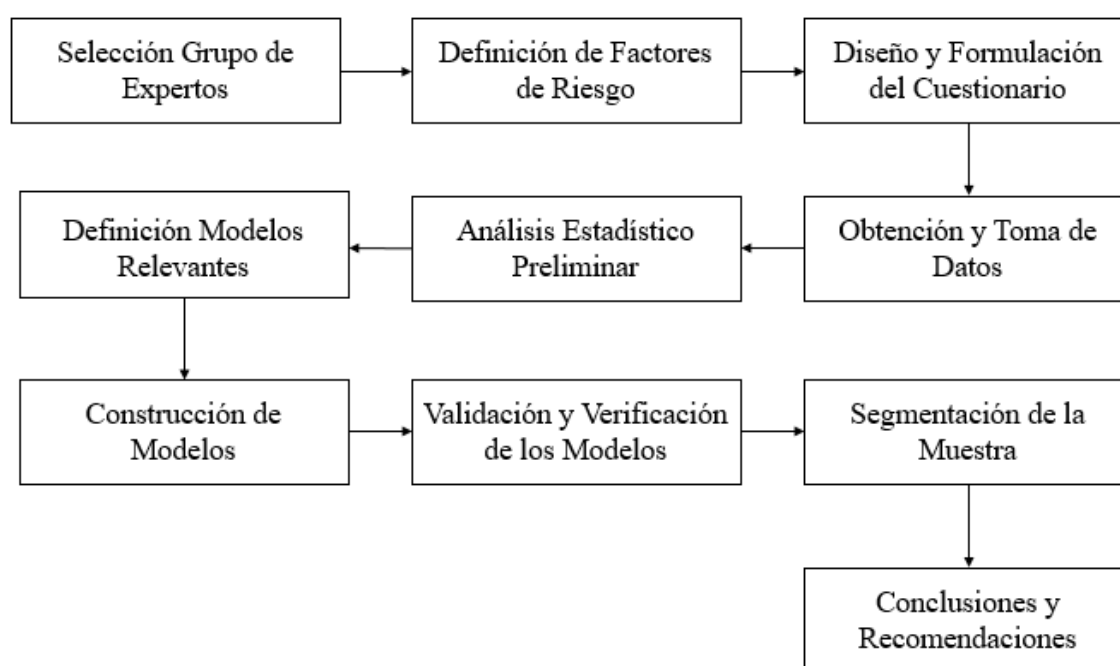


Ilustración 6: Metodología Específica Propuesta

- a) Selección del grupo de expertos, el cual consistirá en profesionales en psicología clínica con altos de experiencia en el manejo y diagnóstico de depresión, los cuales serán la base para la selección de variables a considerar y la construcción de los instrumentos necesarios para el análisis.
- b) Definición de los factores de riesgo a medir mediante revisión de literatura y reuniones con el grupo de expertos. Esto es la definición del conjunto de variables que son consideradas como factores de riesgo para el estudiante y pueden resultar significativas en el diagnóstico.

- c) Diseño y formulación de un cuestionario que apoyara la toma de datos, en las variables seleccionadas anteriormente, de los estudiantes participantes del estudio.
- d) Obtención y toma de datos, mediante la aplicación del cuestionario construido a una muestra aleatoria de estudiantes de la universidad. Esto es el trabajo de campo de la consecución de estudiantes dispuestos a participar en el estudio, la preparación de los mismos y la toma de datos para la construcción de la base de datos, la cual será el punto de partida para la construcción de modelos.
- e) Análisis estadístico preliminar de los datos, con el fin de caracterizar la problemática en la población objetivo. Esto es, el cálculo y realización de medidas descriptivas y pruebas estadísticas relevantes que permitan la caracterización de la muestra.
- f) Definición de los modelos relevantes de acuerdo a la naturaleza de los datos y la revisión de literatura. Ya que se está tratando con un problema de clasificación se utilizarán algoritmos de minería de datos que buscan categorizar las observaciones basados en variables relevantes.
- g) Construcción de los modelos y estimación de parámetros. Dependiendo de la naturaleza de los modelos seleccionados, puede requerir reuniones periódicas con el grupo de expertos con el fin de que su conocimiento guíe la estructuración y la estimación.
- h) Validación y verificación de los modelos. Basado en la muestra de datos, se buscará validar el comportamiento de cada modelo construido con el fin de realizar comparaciones y juicios de valor acerca de la calidad diagnóstica de cada modelo, identificado sus fortalezas y debilidades, mediante validación cruzada y su desempeño sobre pruebas de test.
- i) Segmentación de la Muestra. Con los resultados obtenidos del modelo de clasificación se buscará segmentar la muestra con el fin de encontrar grupos diferenciados de riesgo y sus características representativas, lo cual facilitará el diseño de programas preventivos y de tratamiento.
- j) Formulación de conclusiones y recomendaciones.

6 Aplicación de la metodología específica

6.1 Selección del grupo de expertos

El conocimiento técnico acerca del tratamiento y diagnóstico del trastorno depresivo mayor se obtuvo de un grupo de psicólogos clínicos con varios años de experiencia como docentes y como terapeutas privados. El grupo estuvo conformado principalmente por Diego Leal y Emilio Herrera actualmente docentes en la Universidad del Rosario. Con ambos, se realizaron varias reuniones donde se discutía del tema y sobre la relevancia o mérito de cada una de las variables que terminarían en el modelo.

6.2 Definición de las variables relevantes

Para el diseño de un modelo que permita predecir la probabilidad de que un estudiante en particular sufra depresión, es indispensable realizar una selección rigurosa de variables que aporten información relevante y resulten herramientas útiles en la discriminación de estudiantes con y sin depresión. Con el fin de obtener las variables más relevantes se utilizaron dos fuentes de información: primero la revisión de literatura, donde se creaba una lista preliminar con las variables más usadas tanto en la literatura psicológica como estadística; después de construida, dicha lista era depurada de acuerdo al criterio del experto, el cual la validaba algunas variables, rechazaba otras y proponía una que según su criterio y experiencia resultaban relevantes según distintos modelos psicológicos.

Después de la revisión de literatura, descrita brevemente en la Sección 3, y de varias reuniones con el grupo experto, en la que se revisaba y replanteaban las variables, se definió un total de 35 variables, distribuidas en demográficas, hábitos de salud del estudiante, eventos estresores que al pasar por ellos pueden desencadenar sentimientos depresivos y factores de riesgo, siendo las variables tipo cognitiva y conductual que se miden para dar el diagnóstico clínico de depresión. Las variables se pueden observar en la Tabla 2.

Dependiente	Depresión		Falta de apetito
Demográficas	Edad	Hábitos Salud	Ejercicio
	Genero		Consumo de drogas/ alcohol
	Facultad		Historia familiar de depresión
	Relación		Desbalance de sueño
	Vive Solo		Actividades Evitación
	Semestre		Actividades de Activación
	Promedio	Factores de Riesgo	Pensamientos recurrentes de muerte
	¿Fuera de Bogotá?		Perdida de placer
	Amigos en la Universidad		No vale la pena
	Red Social		Inutilidad
	Satisfacción Red Social		Falta de Concentración
	Condición Socioeconómica		Tristeza
Eventos Estresores	Insatisfacción Carrera		Soledad
	Atrasado carrera		Culpa
	Muertes de un familiar		
	Terminación de relación		
	Heridas o enfermedad personal		
	Problemas económicos		
	Bajo Desempeño académico		
	Problemas en el Hogar		

Tabla 2: Variables definitivas por tipo de variable

6.3 Diseño y formulación del instrumento de medición.

Dada la naturaleza abstracta de la variable de interés, es necesario el uso de escalas e instrumentos de medición con el fin de capturar los comportamientos del individuo y poder generar una variable proxy que dé indicativos si un estudiante presenta o no síntomas depresivos. Para este estudio, tal y como se mencionó en el marco teórico, se hizo uso de la Escala de Depresión del Centro de Estudios Epidemiológicos en su versión corta (CES-D10), la cual fue diseñada para medir el nivel actual de sintomatología depresiva, haciendo énfasis en el componente cognitivo y conductual, con menor presencia de los síntomas de tipo somático (González Llana & Martín Carbonell, 2007).

De esta manera se tomó la versión en español de la escala y se introdujo como una de las secciones de la encuesta para ser aplicada a los participantes del estudio, durante la toma de datos.

Habiendo definido el grupo de variables relevantes que influyen sobre la variable de interés, se procede a diseñar un cuestionario que permita medir dichas variables para cada uno de los participantes. Con la asistencia del grupo experto se formularon preguntas que

representaran cada una de las variables a medir, siendo meticuloso en la redacción de las preguntas con el fin de no introducir un sesgo de respuesta a la pregunta y así, alterar los resultados.

Depresión en Estudiantes Universitarios

*Obligatorio

Para la realización de esta escala en la primera parte usted deberá responder una serie de preguntas para informar sobre sus datos sociodemográficos. Gracias por responder este cuestionario.

¿Cuántos años tiene? *

Elige ▼

¿Cuál es su género? *

☐ Hombre

☐ Mujer

☐ Otro: _____

Depresión en Estudiantes Universitarios

*Obligatorio

Para la segunda parte, usted deberá leer una lista de frases que describen la forma en como pudo haberse sentido o comportado durante la última semana. Por favor lea con cuidado y responda de manera honesta.

Me molestaron cosas que normalmente no me molestan *

☐ Raramente o Nunca (menos de 1 día)

☐ Algo o un poco del tiempo (1-2 días)

☐ Ocasionalmente o Moderadamente (3-4 días)

☐ Siempre o Todo el tiempo (5-7 días)

Tuve dificultad para mantener mi mente en lo que estaba haciendo *

☐ Raramente o Nunca (menos de 1 día)

Ilustración 7: Vista del cuestionario construido para la toma de datos.

El cuestionario final consistió de 45 preguntas, organizadas en cinco secciones de acuerdo con la naturaleza de las variables que se miden en cada una de ellas: el instrumento de medición que actúa como proxy de comportamiento depresivo, que consta de 10 preguntas; hábitos de salud, que consta de 7 preguntas; eventos estresores que consta de 8 preguntas, factores de riesgo, que consta de 8 preguntas y, por último, variables sociodemográficas, que consta de 12 preguntas.

6.4 Obtención y toma de datos

Con el fin de facilitar el registro de las respuestas de cada participante, el formulario se subió a la herramienta Google Forms. Con el cuestionario en línea se hicieron dos aproximaciones a la toma de datos. Ambos métodos se llevaron a cabo por una semana y al finalizar, se obtuvo una muestra de 610 estudiantes de contexto social, económico y académico variado.

El primero de estos caminos fue, el trabajo de campo presencial, el cual involucraba recorrer la universidad (principalmente los edificios ML, W y estudiantes de la clase de Probabilidad y Estadística I) preguntando a los grupos de estudiantes ubicados en las mesas

de cada piso, si estaban dispuestos a participar en la investigación, se les explicaba acerca de esta y se les aplicaba el formulario.

El segundo método de recolección de datos, consistió en la difusión masiva del formulario por medio de los grupos de estudiantes de la universidad con una publicación en redes sociales, en la cual se les pedía su colaboración con la investigación y se les explicaba a los participantes la finalidad de la encuesta y el manejo de los datos a recolectar.

6.5 Análisis estadístico de los resultados de la encuesta.

6.5.1 Comparación de la muestra contra la población real de la universidad

En primer lugar, es necesario comparar la muestra obtenida con la población total de la universidad con el fin de garantizar de cierta manera que la muestra obtenida es representativa de la población de la universidad en términos de distribución por departamentos y distribución por género, de manera general y por departamento.

Se puede observar que la muestra presenta desbalance en cuanto al género de los participantes con respecto a la universidad. Mientras que la universidad tiene un porcentaje de hombres y mujeres estudiando muy similar (55% hombres y 45% mujeres), en la muestra tomada existe una clara prevalencia de las mujeres siendo casi el doble que los hombres encuestados (65% mujeres y 35% hombres).

	Mujer	Hombre
Universidad	45,38%	54,62%
Muestra	64,05%	35,95%

Tabla 3: Distribución por género Universidad vs Muestra

Cuando se hace la comparación por facultad, la muestra parece ser mucho más similar a la realidad de la universidad. La composición de la muestra es prácticamente idéntica a la de la universidad en las facultades de Arquitectura y Diseño, Artes y Humanidades, Derecho, Economía, Educación, Gobierno, Ingeniería y Medicina. Las mayores discrepancias se encuentran en la facultad de ciencias que compone el 6.9% de la población de la universidad y el 10.8% de la muestra; o en la facultad de ciencias sociales la cual compone el 10% de la población de la universidad y el 17.16% de la muestra. En general, cuando se

observa por facultades la muestra resulta ser más balanceada y significativa de la población universitaria.

Facultad	Universidad	Muestra
Administración	7,67%	4,08%
Arquitectura y Diseño	9,91%	9,31%
Artes y Humanidades	5,15%	5,39%
Ciencias	6,93%	10,78%
Ciencias Sociales	10,00%	17,16%
Derecho	6,95%	5,56%
Economía	6,19%	4,90%
Educación	0,13%	0,16%
Escuela de Gobierno	1,73%	0,49%
Estudios Intercambio	1,23%	0,00%
Estudios Dirigidos	0,85%	0,00%
Ingeniería	37,83%	37,58%
Medicina	5,42%	4,58%

Tabla 4: Distribución por Facultad Universidad vs Muestra

La discrepancia en la muestra sugiere que, aunque tiene un tamaño muestral significativo, una variación adecuada entre facultades y género, los datos obtenidos presentan la mayor debilidad del estudio, lo que era de esperarse dado que la composición y diseño experimental adecuado es un arduo trabajo por sí solo, que para efectos de la finalidad del trabajo y disposiciones de tiempo no pudo desarrollarse a cabalidad.

Aun así, la muestra resulta ser representativa de la población general de la universidad por género y facultad por separado, variables de suma importancia en el estudio de la variable de interés, permitiendo que el trabajo genere análisis y conclusiones validas adaptables a la población general en la Universidad de los Andes.

6.5.2 Análisis Descriptivo de la Muestra

De los 610 estudiantes entrevistados se encontró una prevalencia de los síntomas del trastorno depresivo mayor en un poco más de la mitad de participantes, con 346 estudiantes en riesgo de depresión y 264 estudiantes sin riesgo, haciendo la prevalencia del trastorno en los estudiantes de la Universidad de los Andes de 56.72%. Debido a la naturaleza de los datos recolectados, se tienen variables discretas y continuas, para las cuales se requieren distintos tipos de análisis.

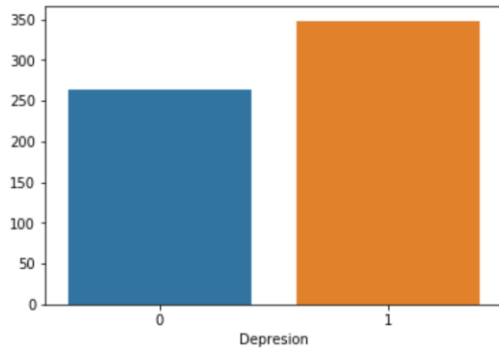


Ilustración 8: Prevalencia de la depresión en la muestra

En primer lugar, se tienen las variables categóricas o discretas para los cuales se registran los niveles de la variable y el porcentaje de datos faltantes. Este procedimiento se puede observar en la Tabla 3.

Variable	Categorías	% Faltantes
Genero	[Hombre, Mujer]	0%
Facultad	[Medicina, Ingeniería, Ciencias, Administración...]	0%
Estrato	[4, 3, 5, 6, 2, 1]	0%
Bogotá	[No, Sí]	0%
Falta Concentración	[Sí, No]	0%
Tristeza	[No, Sí]	0%
Soledad	[No, Sí]	0%
Atrasado	[No, Sí]	0%
Insatisfecho	[No, Sí]	0%
Relación	[No, Sí]	0%
Vivir Solo	[No, Sí]	0%
Amigos	[Sí, No]	0%
Pensamientos de la Muerte	[No, Sí]	0%
Perdida Placer	[No, Sí]	0%
Vale Pena	[Sí, No]	0%
Inútil	[No, Sí]	0%
Culpa	[No, Sí]	0%
Actividades Activación	[Sí, No]	0%
Red Apoyo	[Sí, No]	0%
Falta Apetito	[No, Sí]	0%
Ejercicio	[Sí, No]	0%
Alcohol Drogas	[No, Sí]	0%
Historial Depresión	[No, Sí]	0%
Desbalance de Sueño	[No, Sí]	0%
Muerte Familiar	[No, Sí]	0%
Terminación Relación	[No, Sí]	0%
Heridas Enfermedad	[No, Sí]	0%
Problemas Económicos	[No, Sí]	0%
Problemas Académicos	[No, Sí]	0%
Problemas Hogares	[No, Sí]	0%
Otro Estrés	[No, Sí]	0%
Depresión	[No, Sí]	0%

Tabla 5: Análisis Descriptivo Variables Categóricas

Adicionalmente, para cada categoría de cada variable se graficó el número de casos de depresión con el fin de entregar una primera estimación de la incidencia de cada una de

estas variables sobre la probabilidad de presentar depresión. A continuación, se presentan solo algunas de las variables. Las gráficas completas se pueden encontrar en Anexos.

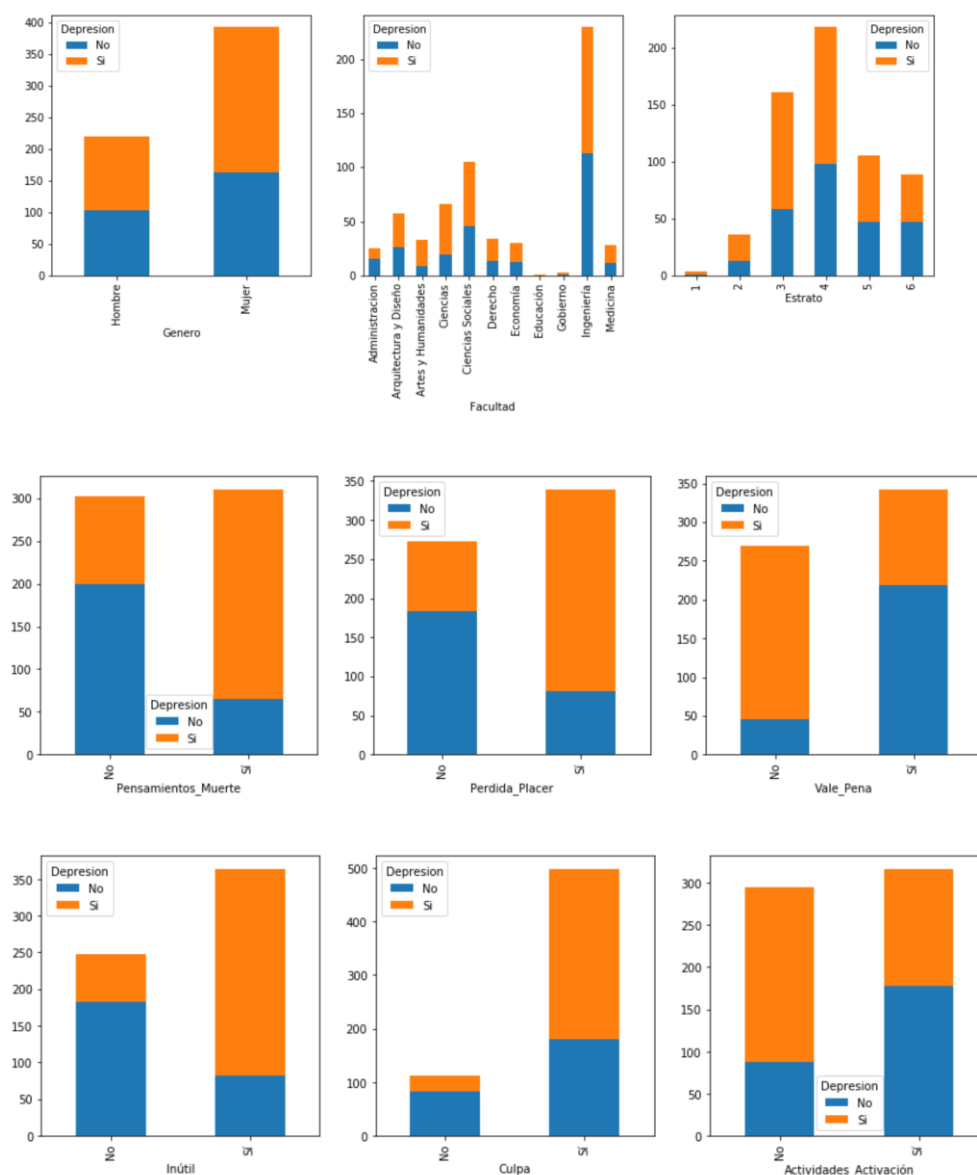


Ilustración 9: Prevalencia Depresión por variables categóricas

De un análisis preliminar parecen existir variables que permiten una clara separación entre aquellos estudiantes que presentan depresión y los que no. Por ejemplo, puede verse que los comportamientos depresivos son más frecuentes en estudiantes que presentan: sentimientos de culpa, que sienten que lo que hacen no vale la pena, que presentan pensamientos recurrentes acerca de la muerte, pérdida de placer en las actividades que consideraban sus favoritas o sentimiento de inutilidad, todas estas mencionadas en la literatura como factores

de riesgo en la presencia de trastorno depresivo. Así mismo, se puede observar que la prevalencia de la depresión es mayor en mujeres que en hombres y es más probable su aparición en carreras de Ciencia e Ingeniería.

Para las variables numéricas como la edad, semestre o promedio se calcularon las principales estadísticas descriptivas, así como las correlaciones entre cada par de variables. La tabla 4 y la Ilustración 9 muestran estos resultados.

	Cuenta	Media	Desviación	Mínimo	25%	50%	75%	Máximo
Edad	610	20.62	1.93	16.00	19.0	20.00	22.0	28.0
Semestre	610	6.49	2.61	1.00	5.0	6.00	8.0	20.0
Promedio	610	3.93	0.35	2.94	3.7	3.98	4.2	4.8
Actividades de Evitación	610	3.49	2.97	0.00	2.0	3.00	5.0	25.0
Satisfacción de la Red	610	3.51	1.02	1.00	3.0	4.00	4.0	5.0

Estadísticas Descriptivas Variables Numéricas

El estudiante promedio tiene 20.6 años, está entre 6 y 7 semestre y posee un promedio de 3.93. Se puede observar que las variables no tienen alta variabilidad, pues poseen una desviación estándar baja y en general están acumulados alrededor de la media.

En cuanto a las correlaciones, se observa que la única correlación que parece existir entre las variables numéricas es entre el semestre y la edad, lo cual tiene sentido ya que a medida que aumenta el semestre en el que se encuentra el estudiante, se espera que su edad también aumente. La distribución de las variables de acuerdo al riesgo de depresión parece ser muy similares, a excepción de la satisfacción de la red, lo cual nos da un indicio de que estas variables pueden no ser muy útiles a la hora de determinar el grado de riesgo de los estudiantes.

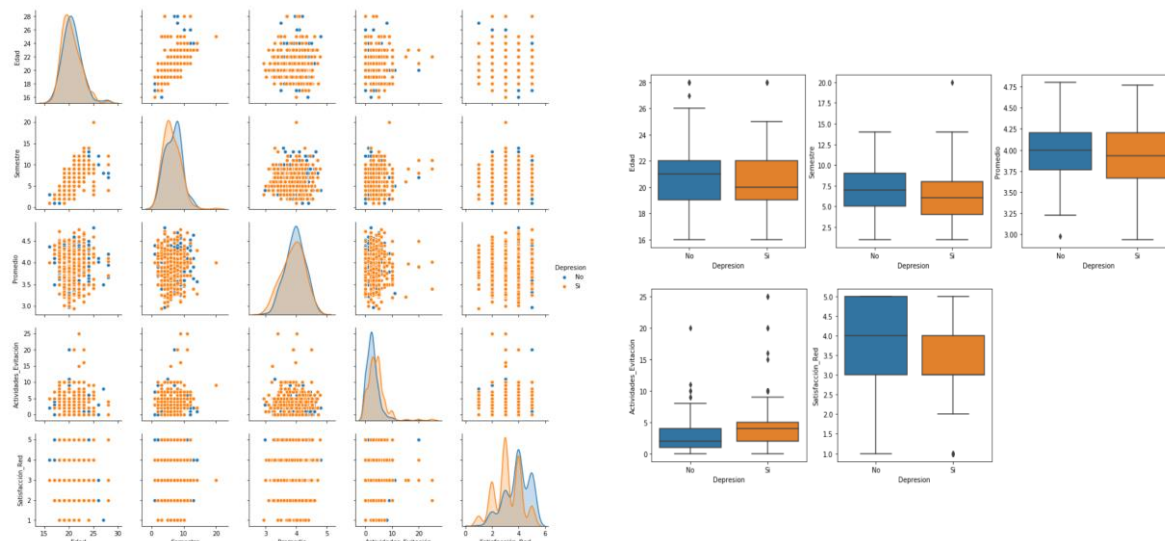


Ilustración 10: Correlación y Distribución Variables Numéricas

Por otro lado, el instrumento psicométrico que nos sirve como proxy de depresión presenta un alfa de Cronbach de 0.839 y un índice lambda de Guttman de 0.828 sobre la muestra de datos, indicando una buena consistencia interna, con una correlación promedio entre los ítems de 0.33. Así mismo, los índices obtenidos para cada uno de los ítems de la encuesta son siempre menores que los índices totales, indicando que cada uno de los ítems provee información relevante para la validez y consistencia interna del instrumento.

#	Ítem	Alfa	Lambda
I1	Me molestaron cosas que normalmente no me molestan	0.81	0.81
I2	Tuve dificultad para mantener mi mente en lo que estaba haciendo	0.81	0.80
I3	Me sentí deprimido	0.79	0.79
I4	Tuve la impresión de que todo lo que hice necesitó esfuerzo	0.83	0.83
I5	Me sentí esperanzado acerca del futuro	0.82	0.82
I6	Me sentí con miedo	0.81	0.81
I7	Mi sueño fue intranquilo	0.82	0.82
I8	Yo estuve feliz	0.81	0.80
I9	Me sentí solo	0.81	0.81
I10	No pude ponerme en marcha	0.81	0.81

Tabla 6: Alfa de Cronbach y Lambda de Guttman para cada uno de los ítems del instrumento

6.5.3 Pruebas de Hipótesis

Como medida preliminar al diseño de los modelos de clasificación se busca determinar si cada una de las variables de las cuales se tienen registros, presentan una diferencia

significativa estadísticamente entre los dos grupos de estudiantes, los que están en riesgo de sufrir depresión y los que no.

Esta diferenciación se realizará mediante la prueba de Wilcoxon, para las variables continuas y mediante la prueba de independencia Chi cuadrado para las variables categóricas. La prueba de Wilcoxon es el equivalente no paramétrico de la prueba t para la diferencia de medias, en la cual no es necesario que las poblaciones provengan de una distribución Normal.

Variable	P-Value	¿Afecta?
Relación	0.8754	No
Alcohol.Drogas	0.3338	No
Historial_Depresión	0.2953	No
Muerte_Familiar	0.7278	No
Terminación_Relación	0.2253	No
Heridas_Enfermedad	0.4615	No
Problemas_Económicos	0.2615	No
Otro_Estres	0.3013	No
Genero	0.0262	Si
Facultad	0.0506	Si
Estrato	0.0158	Si
Bogotá	0.1000	Si
Falta_Concentración	0.0000	Si
Tristeza	0.0000	Si
Soledad	0.0000	Si
Atrasado	0.0029	Si
Insatisfecho	0.0000	Si
Vivir_Solo	0.0910	Si
Amigos	0.0053	Si
Pensamientos_Muerte	0.0000	Si
Perdida_Placer	0.0000	Si
Vale_Pena	0.0000	Si
Inútil	0.0000	Si
Culpa	0.0000	Si
Actividades_Evitación	0.0000	Si
Red_Apoyo	0.0000	Si
Falta_Apetito	0.0000	Si
Ejercicio	0.0000	Si
Desbalance_Sueño	0.0393	Si
Problemas_Acamémicos	0.0001	Si
Problemas_Hogares	0.0094	Si

Variable	P-Value	¿Afecta?
Edad	0.0130	Si
Semestre	0.0077	Si
Promedio	0.0446	Si
Actividades_Activación	0.0000	Si
Satisfacción_Red	0.0000	Si

Tabla 7: Pruebas de independencia Chi Cuadrado y Prueba de Rangos de Wilcoxon

Como se quiere probar si las variables afectan de manera significativa a la probabilidad de tener o no depresión, se busca un p-valor pequeño con el fin de rechazar la hipótesis nula que establece que no existe diferencia estadística entre los dos grupos.

El análisis preliminar arroja que para todas las variables numéricas existe diferencia entre las medias de los estudiantes con riesgo de sufrir depresión y los estudiantes sin riesgo. Por otro lado, en las variables categóricas no todas resultan significativas; las variables de si el estudiante está o no en una relación, si consume alcohol o drogas de manera regular y las variables asociadas a los eventos estresores, resultan no significativas en cuanto a su relación con la variable de interés. La no significancia estadística de estas variables puede responder al hecho de que dentro de las personas encuestadas (independientemente de su nivel de riesgo) son muy pocas las que han experimentado cada uno de estos eventos estresores, por lo cual no alcanza a ser suficiente para presentar diferencias entre un grupo y el otro.

6.6 Definición de los modelos relevantes.

De acuerdo con la revisión de literatura realizada y con el fin de tener una amplia variedad de predictores, cada uno con sus fortalezas y debilidades, se implementarán varios algoritmos de aprendizaje estadístico con el fin de extraer información de cada uno y encontrar cuál de ellos se acomoda mejor a nuestro caso particular. En total se probarán y compararán 12 metodologías de predicción diferentes: Regresión Logística, *Classification Trees*, *Random Forest*, Discriminante Lineal, tres tipos de *Support Vector Machine*, cada uno con un Kernel diferente, *Adaptative Boosting*, *Gradient Boosting*, *Xtreme Gradient Boosting*, *K-Nearest Neighbour* y *Naive Bayes*.

6.7 Construcción de los modelos.

Para la construcción de cada modelo, se dividió aleatoriamente la muestra tomada en dos: el 80% de los datos para entrenar los modelos y el 20% de los datos para hacer el test de calidad predictiva de los mismos. Cada modelo fue entrenado y sus hiperparametros calibrados mediante *10 fold cross-validation*, método mediante el cual se realizó una búsqueda exhaustiva dentro de un rango de parámetros, de manera que sea seleccionada la

combinación de los mismos que produjera el mejor modelo en términos del área bajo la curva ROC.

La división de los datos y la validación cruzada, se realiza con el fin de prevenir el sobre ajuste del modelo, de manera que sea útil como herramienta diagnóstica con nuevos casos que no hagan parte de la muestra de entrenamiento. De utilizarse el conjunto completo de los datos para construir los modelos, se corre el riesgo de que cada modelo, memorice la estructura de los datos en lugar de aprender de ellos las generalidades del problema; es decir, no se tendría un estimado del rendimiento predictivo del modelo por fuera de la muestra, haciendo que eventualmente prediga bastante mal nuevas observaciones, imposibilitando su uso como herramienta de apoyo al diagnóstico del trastorno depresivo mayor.

6.7.1 Regresión Logística

Para la regresión logística se optó por un modelo con regularización LASSO, con el fin de que el modelo seleccionara las variables que resultan más útiles a la hora de realizar la predicción y descartara las otras. Adicionalmente, se calibró el parámetro C, que es el inverso de la fuerza de regularización.

6.7.2 Discriminante Lineal

Para el modelo LDA se calibró el parámetro del número de componentes para reducir la dimensionalidad de la muestra, junto con el parámetro de reducción de los coeficientes, lo cual permite disminuir el sobre ajuste del modelo. De esta manera, el modelo final utiliza un único componente y un parámetro de reducción.

6.7.3 Classification Tree

En el árbol de clasificación se calibran 4 parámetros de manera simultánea: la profundidad máxima, el número máximo de variables a utilizar para realizar las particiones, el número mínimo de observaciones por hoja del árbol y el número mínimo de observaciones necesarias para realizar una partición.

6.7.4 Random Forest

En modelo *Random Forest* se calibran 4 parámetros de manera simultánea: la profundidad máxima, el número máximo de variables a utilizar para realizar las particiones, el número mínimo de observaciones por hoja del árbol y el número mínimo de observaciones necesarias para realizar una partición.

6.7.5 Support Vector Machine

Como se mencionó en la sección anterior, para el SVM se realizan tres modelos cada uno con un kernel diferente. Se prueban los kernel lineal, radial y polinómico. Para cada uno de los tres modelos se calibra el parámetro C, que representa el parámetro de penalización del término de error y el parámetro gamma o el coeficiente del kernel. Adicionalmente, para el modelo con el kernel polinómico se calibra el grado del polinomio.

6.7.6 AdaBoost

En el modelo construido con *Boosting* Adaptativo se calibraron los parámetros de la tasa de aprendizaje y el número de estimadores a usar.

6.7.7 Gradient Boosting

En el modelo construido con *Gradient Boosting* se calibraron los parámetros de la tasa de aprendizaje, el número máximo de variables a usar, el número de estimadores bajo los cuales hacer el *boosting*, la profundidad máxima, el número máximo de variables a utilizar para realizar las particiones, el número mínimo de observaciones por hoja del árbol, el número mínimo de observaciones necesarias para realizar una partición y la fracción de la muestra a utilizar para entrenar los modelos base.

6.7.8 Xtreme Gradient Boosting

En el modelo construido con *Xtreme Gradient Boosting* se calibraron 8 parámetros: la fracción de la muestra a utilizar para entrenar los modelos base, el parámetro lambda y alfa que controlan la regularización con el fin de evitar el sobre ajuste, la profundidad máxima de los árboles, la tasa de aprendizaje, el parámetro gamma que controla la reducción mínima sobre la función objetivo que debe haber para crear una nueva partición, el peso mínimo que cada hoja debe tener y el número de variables a utilizar por cada árbol.

6.7.9 K-Nearest Neighbours

En modelo KNN fue construido de una manera diferente a los anteriores. Debido a que este método utiliza distancias para agrupar las observaciones y en la muestra se tienen variables categóricas y variables continuas no se pueden agrupar bajo una única distancia. Por esta razón, se separan los datos de acuerdo a su tipo y se ajusta un modelo a cada grupo, con la medida de distancia adecuada.

Para el modelo con los datos categóricos se calibraron los parámetros del número de vecinos en un rango de 1 a 10 y la distancia a utilizar, entre *Jaccard*, *Dice* y *Matching*. Por otro lado, para el modelo con los datos continuos se calibraron los parámetros del número de vecinos en un rango de 1 a 10 y la distancia a utilizar, entre *Euclidean*, *Manhattan* y *Chebyshev*.

Una vez se entrenan los dos modelos por separado, se utilizan las predicciones generadas por cada uno como input para un modelo combinado. Ya que las variables de entrada son todas continuas, pues son probabilidades generadas como predicción de cada uno de los modelos anteriores, se construye con ellas un modelo KNN donde se calibran el número de vecinos y la distancia a utilizar dentro de las continuas (*Euclidean*, *Manhattan* y *Chebyshev*).

6.7.10 Naive Bayes

Al igual que el modelo KNN, el modelo que hace uso de *Naive Bayes* fue construido de una manera diferente a los demás. Debido a que este método asume distribuciones de probabilidad para las variables de entrada y en la muestra se tienen variables categóricas y variables continuas no se pueden agrupar bajo una única distribución, pues se tendría una distribución Bernoulli para las variables binarias y una distribución Normal para las variables continuas. Por esta razón, se separan los datos de acuerdo a su tipo y se ajusta un modelo a cada grupo, con la distribución adecuada.

Para el modelo con los datos categóricos se utilizó la distribución Bernoulli y se calibró el parámetro alfa, que representa la suavización que se desea para las estimaciones. Por otro lado, para el modelo con los datos continuos se utilizó la distribución Normal y se calibró también el parámetro de suavización.

Una vez se entrenan los dos modelos por separado, se utilizan las predicciones generadas por cada uno como input para un modelo combinado. Teniendo en cuenta que, las variables de entrada son todas continuas, pues son probabilidades generadas como predicción de cada uno de los modelos anteriores, se construye con ellas un modelo *Naive Bayes* Gaussiano donde se calibra el parámetro de suavización.

7 Análisis de Resultados

Los resultados de los modelos se pueden agrupar en dos grandes categorías: la calidad predictiva de los modelos, en términos del porcentaje de predicciones correctas, así como, de la sensibilidad y la especificidad y la información que es posible extraer de cada uno de ellos sobre la relación entre las variables y la relevancia de cada uno de los inputs sobre la variable de interés. A continuación, se analizará cada una de estas categorías.

7.1 Importancia de las Variables

Junto con la calidad predictiva, los modelos lineales utilizados y aquellos basados en árboles entregan herramientas adicionales para entender el problema. Mediante medidas de pureza e importancia, estos algoritmos permiten obtener las variables que después de la calibración resultaron relevantes a la hora de determinar los individuos que presentan depresión y los que no. Esto permite obtener una mirada más a fondo e identificar las variables fundamentales a la hora de que un estudiante desarrolle depresión. Con esta información se puede mantener controles más exactos y enfocar esfuerzos por la mejora de dichos factores, con el fin de disminuir la prevalencia de la enfermedad en la universidad.

De los modelos utilizados, los que permiten generar un ranking de importancia de las variables de las que se alimentaron son: Regresión Logística, *Classification Trees*, *Random Forest* *Support Vector Machine* con Kernel lineal, *Adaptive Boosting*, *Gradient Boosting* y *Xtreme Gradient Boosting*.

7.1.1 Regresión Logística

Como se trata de un modelo Logit-Lasso, mediante el coeficiente de regularización el modelo escoge las variables que más información entregan para la estimación de la variable de interés. Para este caso, el modelo eliminó las variables del estrato, las variables binarias

de facultad, presentar o no desbalance de sueño, consumir alcohol o drogas y haber experimentado recientemente la muerte de un familiar, todas las demás variables las conserva el modelo para determinar si un estudiante presenta o no depresión.

Para el modelo de regresión logística, la importancia de las variables está dada por la magnitud del coeficiente que acompaña a cada una de las variables que resultan relevantes en el modelo a la hora de hacer la predicción. Así, para el modelo las variables correspondientes a factores de riesgo (tristeza, soledad, falta de concentración, pensamientos recurrentes de la muerte, insatisfacción, sentir que lo que hace vale la pena, entre otros) son las que tienen un mayor impacto sobre la probabilidad de desarrollar depresión, siendo la tristeza y la soledad mucho más importantes en término del coeficiente que las acompaña en el modelo estimado, con coeficientes entre 1.5 y 2 contra los coeficientes de menos de 1 de las demás variables.

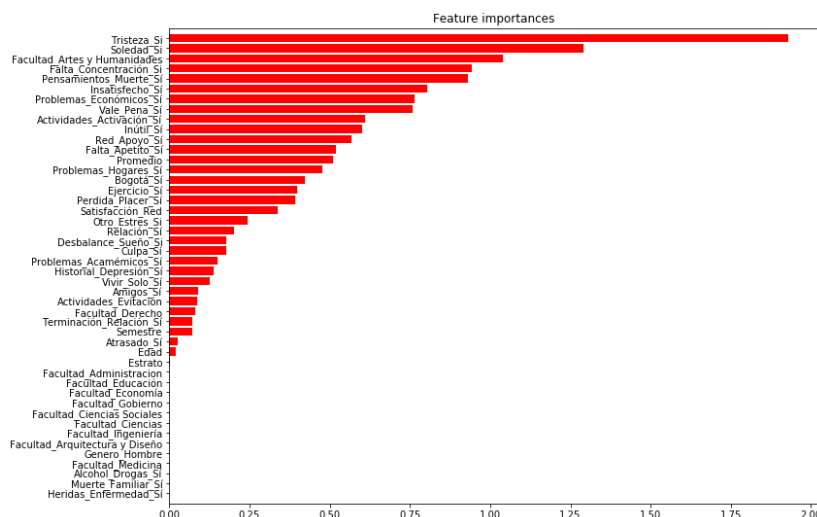


Ilustración 11: Importancia de las Variables modelo Logit

7.1.2 Clasification Tree

El modelo del árbol de clasificación, debido a la gran varianza que posee el algoritmo, cada vez que se corre el modelo, se logra una participación diferente. Por esta razón, el modelo utiliza la gran mayoría de las variables en distintas combinaciones. En promedio, las variables que menos utiliza el modelo son las binarias de facultad de Economía, Gobierno, Educación y Arquitectura y Diseño, así como algunos eventos estresores como, la muerte de un familiar, terminación de una relación, problemas académicos y económicos. Por otra

parte, las variables más relevantes son una vez más los factores de riesgo como experimentar tristeza, pérdida de placer, pensamientos recurrentes de la muerte, el sentimiento de satisfacción e inutilidad, falta de concentración y el sentimiento de soledad.

Una de las ventajas de los árboles de decisión, sobre las otras metodologías utilizadas es que nos provee una herramienta visual de la manera en que realiza las particiones. Nos muestra de manera gráfica las variables utilizadas y las reglas de decisión de cada una para dar el diagnóstico. Resulta relevante, pues se muestra como una información adicional al problema y puede servir de herramienta de apoyo al profesional a la hora de determinar los puntos críticos de ciertas variables o las conexiones entre las variables que resultan en el desarrollo de la enfermedad.

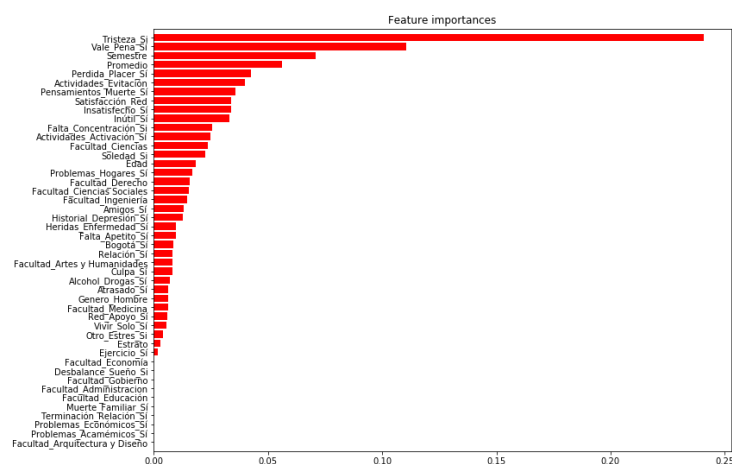


Ilustración 12: Importancia de las Variables modelo CART

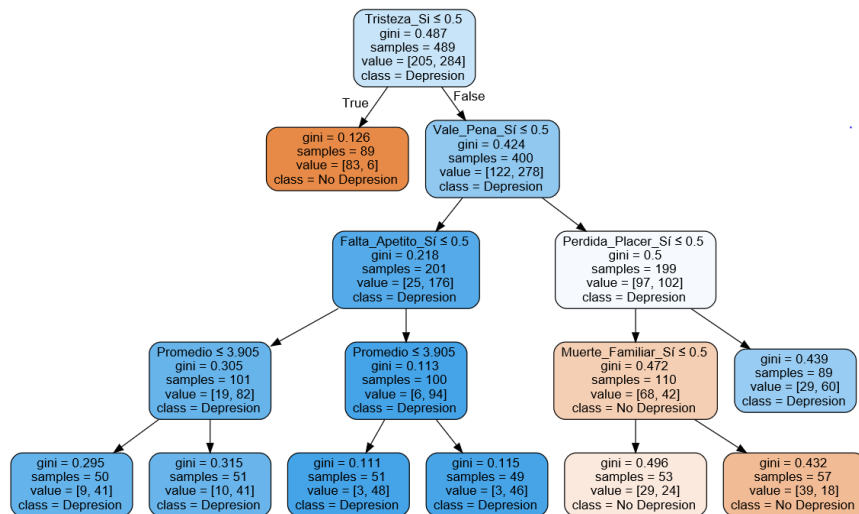


Ilustración 13: Representación reglas de decisión bajo modelo CART

7.1.3 Random Forest

Dada la naturaleza del algoritmo *Random Forest*, se espera que el modelo utilice la mayoría de las variables, pues cada uno de los árboles del bosque es construido con un set diferente y en cada uno de ellos, hay variables que resultan relevantes en la predicción. La única variable que resulta no relevante en ninguno de los arboles es la variable binaria referente a la Facultad de Educación. Las variables con mayor relevancia vuelven a ser los factores de riesgo tristeza, pensamientos recurrentes de la muerte y pérdida de placer. Variables referentes a la vida académica del estudiante cobran mayor importancia en este modelo, con variables como el promedio, la edad y el semestre estando entre las 5 variables más importantes.

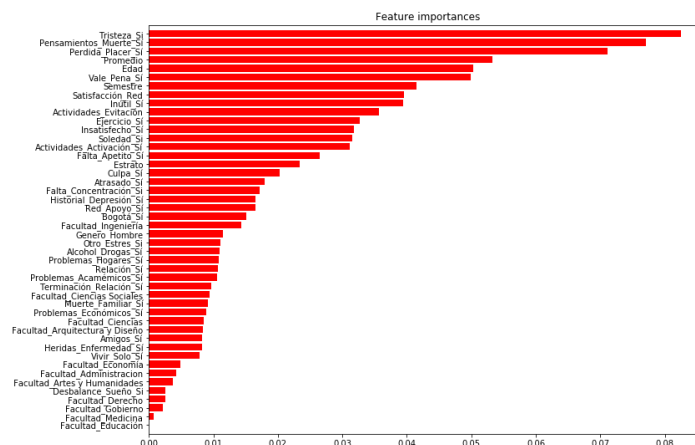


Ilustración 14 Importancia de las Variables modelo Random Forest

7.1.4 Support Vector Machine Lineal

En el modelo de *Support Vector Machine* con kernel lineal se acentúa mucho la importancia de las variables pertenecientes a los factores de riesgo como las más importantes. Dentro de las 10 variables más importantes, todas pertenecen a este grupo. El sentimiento de tristeza, pensamientos recurrentes de la muerte, sentimiento de inutilidad, pérdida de placer, sentimiento de inutilidad e insatisfacción, todas son variables que resultan de mayor relevancia a la hora de hacer las particiones entre los dos grupos de estudio. Por otro lado, las variables de género, Facultad de Derecho y Facultad de Educación son las variables que no aportan información relevante para el modelo.

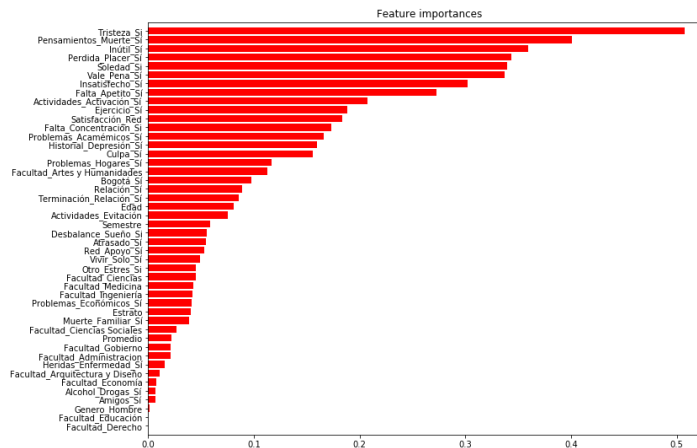


Ilustración 15: Importancia de las Variables modelo Support Vector Machine Lineal

7.1.5 Adaptive Boosting

Para el modelo construido con *Adaptive Boosting*, un gran número de variables resultan no significativas: la mayoría de binarias del programa del estudiante, la mayoría de eventos estresores (muerte familiar, terminación relación, heridas o enfermedad personal). Toman un rol central las variables académicas y demográficas del estudiante como: edad, estrato, promedio y semestre, así como una vez más las variables de factores de riesgo presentan relevancia a la hora de realizar la predicción.

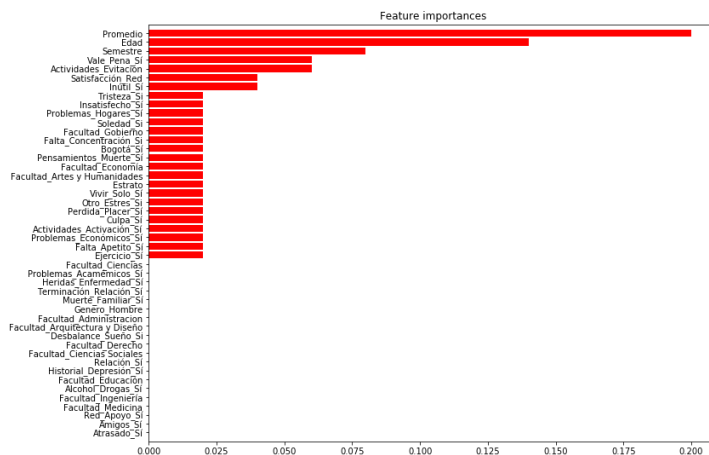


Ilustración 16: Importancia de las Variables modelo AdaBoost

7.1.6 Gradient Boosting

El modelo construido con *Gradient Boosting* muestra resultados muy similares al construido con *Adaptative Boosting*. Este modelo se enfoca en las variables de riesgo, sociodemográficas y académicas, a excepción de la facultad, junto con ciertos eventos estresores. De manera similar a los anteriores modelos, las 5 variables más importantes resultan ser el sentimiento de tristeza, insatisfacción, soledad, inutilidad y los pensamientos recurrentes de la muerte.

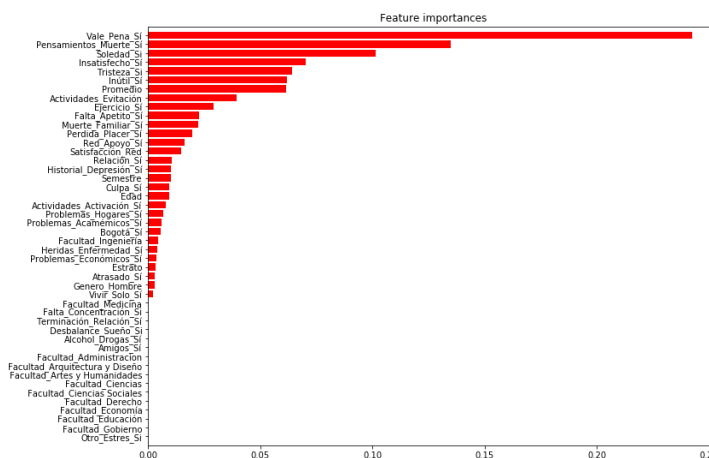


Ilustración 17: Importancia de las Variables modelo Gradient Boosting

7.1.7 Xtreme Gradient Boosting

Por último, el modelo construido con el algoritmo *Xtreme Gradient Boosting* utiliza todas las variables a excepción de las binarias de Facultad de Gobierno y la Facultad de

Educación, posiblemente debido a que son las dos que menor número de estudiantes tienen de acuerdo con los resultados de la encuesta. De manera similar al modelo de *Gradient Boosting*, las variables con mayor importancia resultan ser las sociodemográficas y académicas de los estudiantes como el estrato, el promedio, el semestre y la edad; seguido de las variables consideradas en la literatura psicológica como factores de riesgo como la tristeza, soledad, falta de concentración, pérdida de placer, falta de apetito, entre otras.

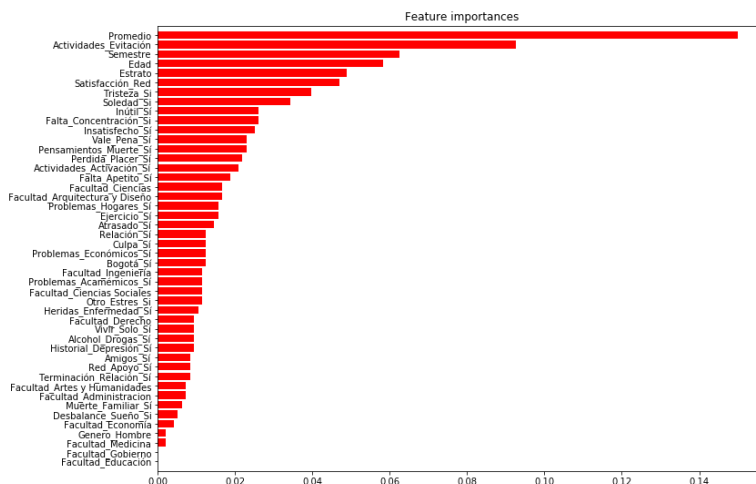


Ilustración 18: Importancia de las Variables modelo Xtreme Gradient Boosting

Como se puede observar en las ilustraciones anteriores los modelos representan el conocimiento de experto con el que fueron contruidos. Todos ellos utilizan como predictores principales las variables clásicas que, en la práctica psicológica son consideradas como factores de riesgo para el desarrollo de la enfermedad. Estas variables son combinadas con características propias del individuo y eventos estresores que pudo haber experimentado, por los modelos que, utilizan lo mejor de ambos grupos de variables para entregar clasificadores flexibles que representan adecuadamente la realidad diagnóstica.

Así mismo, se presentan como una herramienta en la que el profesional clínico puede confiar y en la cual apoyarse para el diagnóstico, pues combina el conocimiento psicológico con el poder estadístico de predicción, propio de este tipo de modelos.

7.2 Calidad Predictiva de los Modelos

El porcentaje de predicciones correctas nos muestra cómo se comportan los modelos de manera general, prediciendo tanto individuos con depresión como sin depresión. La sensibilidad muestra como el modelo se comporta en cuanto a la predicción de individuos con depresión de manera relativa a únicamente los casos de individuos con depresión, se define como la probabilidad que el modelo prediga un individuo con depresión, dado que el individuo efectivamente tenía depresión. Por último, la especificidad muestra como el modelo está prediciendo los casos de individuos que no presentan depresión y se define como la probabilidad de que el modelo prediga que un individuo no tiene depresión dado que efectivamente no presenta depresión.

Debido a que en este caso se tiene prácticamente la misma cantidad de casos de estudiantes con y sin depresión, puesto que se logró obtener una muestra balanceada, se puede guiar la escogencia de los modelos haciendo uso del porcentaje de aciertos de cada uno. Sin embargo, se prefiere utilizar una medida que balancee la calidad que presente el modelo en términos de sensibilidad y especificidad y los errores de predicción de casos positivos y negativos. Con esto en mente, la medida para seleccionar el modelo será la curva ROC.

La curva ROC es una gráfica que simultáneamente muestra los dos tipos de errores para todos los umbrales de clasificación (*Threshold*); el desempeño predictivo general de un algoritmo, para todos los posibles valores del umbral, está dado por el área bajo la curva ROC. Una curva ROC ideal estará en la esquina superior izquierda, por lo que entre mayor área bajo la curva se tenga, mejor será el clasificador. Esta medida es muy útil para la comparación de modelos debido a que toma en consideración ambos errores y todos los posibles Thresholds (James, Witten, Hastie, & Tibshirani, 2017).

La Tabla 7 contiene los resultados de cada uno de los modelos en cada una de las métricas mencionadas. Se busca que los modelos se comporten de manera satisfactoria, con puntajes, al menos, por encima de 85 en todas las métricas.

Modelo	Precisión	Sensibilidad	Especificidad	AUC
Regresión Logística	84,0%	89,0%	78,0%	0,911
Discriminante Lineal	86,0%	91,0%	80,0%	0,904
CART	79,0%	88,0%	68,0%	0,864
Random Forest	83,0%	92,0%	71,0%	0,900
AdaBoost	88,0%	92,0%	81,0%	0,929
SVM Radial	85,0%	91,0%	78,0%	0,904
SVM Lineal	83,0%	89,0%	76,0%	0,899
SVM Polinómico	83,0%	88,0%	78,0%	0,890
Gradient Boosting	85,0%	92,0%	75,0%	0,898
XGBoosting	86,0%	89,0%	81,0%	0,912
KNN	80,0%	89,0%	69,0%	0,792
Naive Bayes	86,0%	91,0%	81,0%	0,899

Tabla 8: Calidad predictiva de los modelos de clasificación

En cuanto a la capacidad de predicción de los modelos, se obtuvo un mínimo de 79% de predicciones correctas con el modelo de árbol de clasificación, lo que era de esperarse debido a su gran variabilidad y los supuestos que hace sobre el espacio de los predictores, mientras que el límite superior se situó en 88%, puntaje el cual se obtuvo con un modelo de ensamblaje, construido mediante el *Boosting* Adaptativo de árboles, seguido por los modelos de *XGBoosting*, *Gradient Boosting*, *Naive Bayes* y SVM con kernel radial con puntajes alrededor de 85%.

En la métrica de sensibilidad, se obtienen resultados muy prometedores con valores que oscilan entre 88% y 92%, indicando que los modelos construidos son capaces de identificar la enfermedad de manera correcta en alrededor del 90% de los casos. El mayor puntaje obtenido de sensibilidad se dio para varios modelos todos construidos desde el ensamblaje de árboles: *AdaBoost*, *Gradient Boosting* y *Random Forest*.

En cuanto a la facultad de los modelos de identificar correctamente las observaciones que no poseen la enfermedad es el área donde peor desempeño muestran los modelos. La sensibilidad no logra superar la barrera del 81%, la cual no es un valor bajo, pero comparado con la sensibilidad tampoco es un valor bueno. Adicionalmente, los valores resultan ser más variables, yendo desde 69% hasta 81%, el cual se obtiene con los modelos de *Boosting* adaptativo y *XGBoosting*.

Por último, se estudia el área bajo la curva ROC. Como se mencionó anteriormente, esta métrica presenta una ponderación entre la sensibilidad y la especificidad para cualquier *threshold* de predicción, por lo cual es una de las métricas más utilizadas para medir la calidad del modelo. En el área bajo la curva ROC se obtienen resultados muy prometedores, indicando que a pesar de posibles falencias que puedan tener los modelos en cada métrica individual, de manera conjunta todos poseen una calidad predictiva significativa. Todos los valores obtenidos están en el rango de 0.89 a 0.93 (a excepción del modelo CART que resulta el peor en esta característica con 0.87) siendo el mejor modelo en esta característica el obtenido con el método de *Adaptive Boosting*.

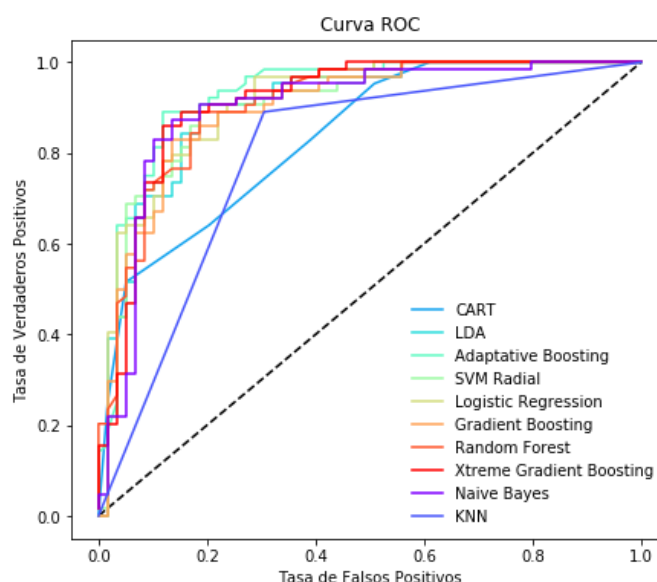


Ilustración 19: Curvas ROC de cada uno de los modelos construidos

Aunque se obtienen resultados prometedores en la mayoría de modelos, especialmente los modelos de ensamblaje, el máximo porcentaje de predicciones correctas que se logra es de 88%, con una sensibilidad y especificidad de 92% y 81% respectivamente, no resultan ser del todo satisfactorias dado que con estas métricas se deja un número considerable de observaciones mal clasificadas (aproximadamente 72 dentro de la muestra). Adicionalmente, se considera que debido a que la enfermedad a tratar suele ser diagnosticada de manera coloquial y suele a menudo confundirse con el sentimiento de tristeza aguda, se requiere una herramienta que permita al profesional, la adecuada identificación de la enfermedad en la mayoría de casos posibles, siendo este número lo más cercano posible a la totalidad de los casos a estudiar.

7.3 Stacking

Debido a que no se logra un porcentaje de acierto por encima del 90% se busca explorar las áreas de mejora de los modelos con el fin de optimizar la predicción, mediante el stacking de los modelos ya construidos. El ensamblaje de los modelos se hará en primer lugar con el promedio simple de las predicciones y en caso de lograr una mejoría, se construirá un meta algoritmo que aprenda a combinarlas complementando, llenando los vacíos y reforzando las debilidades que cada modelo por separado presenta.

Para el primer método de ensamblaje únicamente se tomarán las predicciones de cada modelo individual sobre el mismo conjunto de datos, la muestra de test, y se tomará como predicción final aquella en la que coincidan la mayoría de los modelos. Este método se utiliza como prueba de que efectivamente existe capacidad de mejora en los modelos mediante el ensamblaje.

El segundo método de ensamblaje no es tan sencillo y directo como el primero. Para este, teniendo los modelos individuales ya entrenados mediante validación cruzada, se busca utilizar las predicciones de cada uno de ellos sobre la muestra de *train* como las variables que serán utilizadas para el entrenamiento de un algoritmo de segundo nivel o meta algoritmo. Una vez se entrena el meta algoritmo, también mediante validación cruzada para seleccionar el conjunto óptimo de parámetros, se utiliza para generar la una predicción final sobre el conjunto de datos de *test*, el cual se espera diferencie mejor los individuos que presentan depresión y aquellos que no.

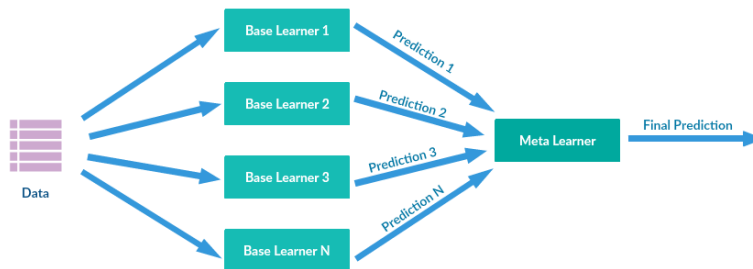


Ilustración 20: Stacking de algoritmos de clasificación. Fuente: Medium.com

Antes de construir los modelos se debe verificar que efectivamente se complementen unos con otros, es decir, que donde unos poseen debilidades, los otros presenten fortalezas y viceversa. Mediante el estudio de las predicciones por observación que resultan de cada uno de los modelos, se pretende encontrar aquellas con la menor correlación, de manera que al combinarlas se pueda obtener una mejoría en la capacidad de los modelos de identificar todo tipo de individuos, tanto los que presentan como los que no presentan la enfermedad.

La Ilustración 20 muestra la matriz de correlación entre todos los pares de modelos con el fin de estudiar, si existe la posibilidad de mejorar la calidad predictiva del modelo mediante la combinación (o stacking) de los modelos. Idealmente todos los modelos deberían tener correlaciones muy bajas, lo que indica que no todos predicen de la misma manera y que por ende existe posibilidad de mejorar las predicciones al combinarlos.

El análisis de correlación se hace con únicamente 10 modelos, todos de metodologías diferentes, es decir se eliminan 2 modelos construidos mediante SVM con los kernels lineal y polinómico, dejando únicamente el modelo de kernel radial. Esto debido a que al ser modelos con la misma metodología presentan desempeños muy parecidos y en el análisis preliminar mostraban correlaciones muy altas, lo cual convierte el hecho de considerar los tres modelos en algo redundante.

En este caso, se puede observar que las correlaciones entre los modelos no son tan bajas, pero a su vez no son lo suficientemente altas, para que el ensamblaje de los modelos genere resultados pobres en cuanto a la calidad de la clasificación. Estas rondan entre 0.4 y 0.8, por lo que parece haber aún cabida a mejorías en la predicción. Los modelos parecen tener la capacidad de suplir sus debilidades entre sí, y generar un modelo con mayor poder predictivo y con mejor comportamiento en todos los índices de calidad.

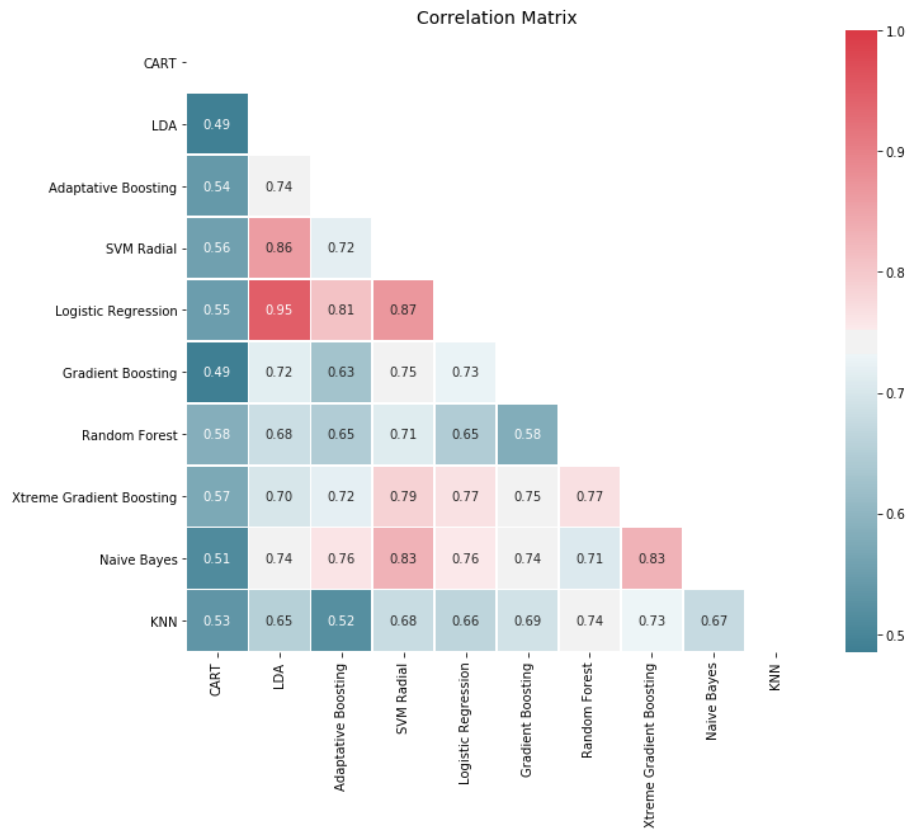


Ilustración 21: Matriz de Correlación de las predicciones de cada modelo individual

7.3.1 Clasificador Final Votación

El primer intento a un clasificador que combine las predicciones de cada modelo creado anteriormente, se construye únicamente como el promedio de los resultados de los modelos individuales. La predicción final para cada individuo será la clase que mayor número de votos tenga dentro del conjunto de clasificadores, es decir, la clase que se asignará a cada individuo de la muestra de test será aquella que la mayoría de modelos pronostique. En este caso, y como se entrenaron 10 modelos individuales se clasificará cada observación en la clase que posea 6 o más votos. La Tabla 8 muestra el desempeño del clasificador construido.

	Precisión	Sensibilidad	Especificidad
Voto Mayoritario	90,24%	93,75%	86,44%

Tabla 9: Desempeño clasificador por voto mayoritario

De primera mano se puede observar que, al realizar un ensamblaje, incluso uno tan simple como el que utiliza el promedio de las predicciones de los modelos individuales, presenta una mejora significativa sobre el desempeño del mejor modelo que se tenía hasta el momento, el construido mediante *Adaptive Boosting*. El clasificador de voto mayoritario obtiene una mejora de aproximadamente 2% en el porcentaje de observaciones correctamente clasificadas, 1.75% sobre la identificación acertada de estudiantes en riesgo de depresión y un 5% en la identificación acertada de estudiantes que no están en riesgo de depresión.

Con el primer ensamblaje básico se obtiene una mejora significativa en el área donde los modelos presentaban más falencias, la especificidad. Con los modelos individuales no se logró identificar correctamente a más del 80% de los estudiantes sin riesgo de depresión, lo cual se sobrepasó con creces mediante el modelo de voto mayoritario, logrando un incremento de aproximadamente el 6%. Este resultado, junto con los aumentos en las otras métricas, apoya la idea de que es posible combinar las predicciones de los modelos individuales satisfactoriamente, permitiendo crear modelos de ensamblaje que suplan las falencias de estos modelos individuales.

7.3.2 Stacking mediante XGBoosting

Sabiendo que es posible obtener mejores resultados mediante el stacking de modelos, se pretende construir un clasificador de XGBoosting el cual, mediante las predicciones de los 10 modelos, aprenda a diferenciar los individuos de acuerdo a su nivel de riesgo de desarrollar la enfermedad, mejor que los modelos individuales. Adicionalmente, se espera que al estar utilizando un clasificador sofisticado sobrepase el ensamblaje de voto mayoritario creado en la sección anterior.

Al igual que el modelo de primera fase construido con el algoritmo, se utilizó validación cruzada para hacer la estimación del modelo y la calibración de sus hiperparámetros. Después de entrenar el modelo se realizaron las predicciones sobre la muestra de test, con el fin de indagar si al combinar las predicciones de los modelos individuales mediante *boosting* de árboles, se obtiene una mejoría sobre la capacidad diagnóstica del modelo para identificar ambos grupos de estudiantes. La tabla 9 muestra el desempeño del clasificador construido

	Precisión	Sensibilidad	Especificidad
XGBoosting	93,50%	96,88%	89,83%

Tabla 10: Desempeño algoritmo de Stacking mediante XGBoosting

El ensamblaje construido mediante uno de los algoritmos más poderosos de clasificación, sobrepasa en calidad predictiva al ensamblaje mediante voto mayoritario y en una gran medida a todos los modelos individuales con una precisión de 93.5% presenta una mejora de 3% sobre el ensamblaje anterior, y de 5% sobre el clasificador individual *AdaBoost*. La sensibilidad llega a 96.8%, mostrando mejoría de 3% sobre el ensamblaje anterior y de 4.7% sobre el modelo *Adaboost*; por último, la especificidad, que es la métrica donde peor desempeño tienen los clasificadores individuales, tiene un valor de casi 90%, siendo este valor un 4% superior al ensamblaje de voto mayoritario y un 9% mayor al modelo *AdaBoost*.

El stacking de modelos resultó ser una herramienta muy útil para el problema bajo estudio. Al entrenar el algoritmo predictivo con las predicciones iniciales de los modelos individuales, se aprendió la estructura ideal de las mismas, permitiendo al modelo combinar de manera satisfactoria los outputs individuales y resultando en predicciones mucho más acertadas y confiables en el diagnóstico de la enfermedad.

Este modelo se destaca en todas las métricas, con sus altos puntajes en precisión, sensibilidad y especificidad. Este desempeño lo convierte en una herramienta confiable de apoyo a la decisión, en el diagnóstico del trastorno depresivo mayor en estudiantes universitarios. De esta manera, el uso del modelo construido puede ayudar a la identificación temprana y eficaz de un mayor número de casos, lo cual permitirá diseñar planes de tratamiento adecuados y así como programas de prevención enfocados a cada una de las poblaciones analizadas.

7.4 Comparación

Habiendo obtenido un modelo de clasificación satisfactorio, es de interés hacer una pequeña comparación de dicho modelo con los construidos por otros autores y que están presentes en la literatura. Es importante tener en cuenta, que no es posible realizar una

comparación completa debido a la gran disparidad que existe entre los métodos, variables y especialmente las bases de datos utilizadas por cada uno de los trabajos.

Por esta razón, el análisis y la comparación que se harán a continuación se hace únicamente con la intención de dar un proxy del desempeño y la competitividad relativa del modelo construido con respecto a aquellos existentes en la literatura.

Para la comparación se toman cinco de los trabajos con mayor similitud a este en cuanto a las variables y métodos de predicción utilizados. Todos estos trabajos fueron consultados durante la fase de investigación de este proyecto y su análisis se encuentra en la sección de revisión de literatura.

Trabajo	Modelo	Precisión	Sensibilidad	Especificidad	AUC
Ojeme, Akazue & Nwalih (2016)	BN+KNN	94,60%	94,70%	97,20%	98,90%
Sau & Bhakta (2016)	RF	89,00%	89,00%	89,10%	94,30%
Galiatsatos et al. (2015)	BN	83,51%	-	-	79,80%
Bhakta & Sau (2016)	MLP	90,00%	92,00%	-	98,00%
Mwangi et al. (2012)	SVM	90,30%	93,30%	87,50%	-
Mejía (2019)	XGBoost	93,50%	96,88%	89,83%	97,80%

Tabla 11: Comparación Modelo contra la literatura

De manera general, el modelo que se desarrolló en este trabajo presenta no solo índices de calidad muy buenos, sino que son a su vez bastante competitivos con los presentes en la literatura. De los trabajos analizados, únicamente existe uno con mejor desempeño y es el desarrollado por Ojeme, Akazue, y Nwalih (2016) el cual fue también construido mediante el ensamblaje de distintos modelos; de los modelos construidos con únicamente un algoritmo, la literatura presenta el mejor en el trabajo de Bhakta y Sau (2016) logrando una precisión de 90% y una sensibilidad de 92% mediante un modelo de red neuronal simple o MLP, estando ligeramente por debajo del modelo final, construido durante esta investigación.

7.5 Identificación de grupos de riesgo

Después de haber obtenido un algoritmo que modela satisfactoriamente el problema y es viable como herramienta en el apoyo diagnóstico, se busca mediante las probabilidades generadas para cada individuo según el modelo, segmentar a los estudiantes en grupos de riesgo. Esto permitiría identificar las características promedio de cada grupo de riesgo,

permitiendo determinar con facilidad los casos pertenecientes a cada rango y poder desarrollar planes de acción enfocadas a cada uno de estos, basados en su nivel de riesgo; bien siendo de prevención o atención.

Para lograr la identificación de grupos de riesgo dentro de la muestra, se utilizará el modelo construido mediante ensamblaje y XGBoosting para estimar la probabilidad de cada uno de los individuos de la muestra presente depresión. Esta nueva variable se anexará a la base de datos y se llevará a cabo el algoritmo de clúster de *K-Medoids* de manera que se puedan identificar si existen grupos de observaciones semejantes, sus características principales y comprobar si poseen similares niveles de riesgo.

Se decide utilizar *K-Medoids* sobre cualquier otra metodología de agrupación debido a dos razones particulares: es bastante sencillo adaptar el algoritmo de *K-Medoids* para que trabaje con variables categóricas y continuas mediante el uso de la distancia de gower y especialmente por la capacidad de interpretación de los resultados que provee esta metodología. A diferencia de *K-Means*, el centro de los clústers encontrados mediante *K-Medoids* no es un punto cualquiera, sino es una de las observaciones, el individuo representativo de cada grupo, lo cual permite caracterizar a los estudiantes pertenecientes a cada grupo de riesgo encontrado, facilitando su identificación y posterior tratamiento.

Como se mencionó anteriormente, para encontrar la similitud entre las observaciones se utilizó la distancia de gower, la cual es particularmente útil cuando se tiene una base de datos mixta, es decir, con variables tanto continuas como categóricas. La distancia gower combina la distancia de manhattan para las variables continuas y la distancia dice para las variables categóricas, las cuales agrega y escala al rango 0-1.

Una vez se aplica la distancia, se agrupan las observaciones de acuerdo al algoritmo seleccionado, iterando el número de grupos a crear entre 2 y 10, con el fin de seleccionar el número “óptimo” entendido como aquel que maximiza la silueta, una medida de validación interna que mide la similitud de las observaciones dentro de un mismo clúster en comparación con el clúster más cercano.



Ilustración 22: Silhouette Index para cada número de clusters.

La Ilustración 21 muestra los resultados del índice de validación interna de acuerdo al número de clusters creados por el algoritmo. Dado que la silueta es un índice que se quiere maximizar, se determina que dentro de la muestra hay dos grupos distintivos de observaciones.

La Ilustración 22 contiene la separación de los individuos de acuerdo a su grupo graficados sobre los dos primeros componentes principales. Como se puede observar el algoritmo de *K-Medoids* logra agrupar las observaciones de la muestra en prácticamente dos grupos totalmente separados. Sigue existiendo un ligero margen de error pues se ven unos pocos puntos mezclados en los clusters opuestos, pero en general se puede apreciar que existe una clara separación entre los individuos pertenecientes a cada grupo.



Ilustración 23: Visualización clusters de acuerdo a los dos primeros componentes principales

Sabiendo que existen dos grupos fuertemente delimitados, es de interés analizar si los individuos dentro de cada grupo comparten características similares y más importante, si poseen probabilidades de desarrollar la enfermedad lo suficientemente separadas para que sean considerados con niveles de riesgo diferentes. La Tabla 10 contiene estos resultados.

Clúster	Probabilidad de sufrir Depresión	Riesgo
1	78,81%	Alto
2	27,98%	Bajo

Tabla 12: Resultados de la agrupación en función del nivel de riesgo

La tabla muestra que los grupos encontrados mediante el algoritmo efectivamente pueden ser clasificados en dos niveles de riesgo: los individuos del grupo 1 tienen una probabilidad promedio de sufrir depresión de 78.81%, presentando un riesgo alto y los individuos del grupo 2 tiene una probabilidad promedio de sufrir depresión de 27.98%, presentando un riesgo bajo.

Habiendo identificado los grupos de riesgo presentes en la muestra, es de especial interés el poder interpretar los resultados de cada clúster, con el fin de identificar las características generales que los individuos pertenecientes a cada uno presentan. Eso es posible, gracias a la metodología de agrupación utilizada; examinando los puntos centrales de cada clúster

(los medoids) es posible encontrar el perfil y las características del individuo representativo de cada grupo.

Género	Facultad	Promedio	Estrato	Bogotá	Falta_Concentración	Tristeza	Perdida_Placer	Vale_Pena
Mujer	Ingeniería	3.76	3	Sí	Si	Si	Sí	No
Soledad	Atrasado	Insatisfecho	Relación	Vivir_Solo	Amigos	Pensamientos_Muerte	Ejercicio	Alcohol_Drogas
Si	Sí	Sí	No	No	Sí	Si	No	No
Inútil	Culpa	Act_Evitación	Act_Activación	Red_Apoyo	Satisfacción_Red	Falta_Apetito	P_Hogar	Otro_Estres
Sí	Sí	4	No	Sí	4	No	Sí	No
Historial_Depresión	Desbalance_Sueño	Muerte_Familiar	Terminación_Relación	Heridas_Enfermedad	P_Económicos	P_Acamémicos	Probabilidad	
No	No	Sí	Sí	Sí	Sí	Sí	98,06%	

Tabla 13: Individuo Representativo Clúster de Alto Riesgo

El grupo de alto riesgo está representado por una mujer de la facultad de ingeniería, de estrato 3, con un promedio por debajo de 4, que está atrasada e insatisfecha con su carrera; que durante la última semana ha experimentado falta de concentración, sentimiento de tristeza, pérdida de placer en actividades que antes consideraba sus favoritas, que siente que lo que hace no vale la pena, a pesar de poseer una red de apoyo y sentirse a gusto con ella siente que está sola; ha experimentado pensamientos recurrentes acerca de la muerte, no hace ejercicio ni consume alcohol o drogas de manera regular. Lo más interesante de este grupo es la cantidad de eventos estresores por los que ha pasado últimamente: muerte de un familiar, terminar una relación, heridas personales problemas económicos, académicos y en el hogar. Dadas estas características es de esperar un nivel alto de riesgo, lo cual se confirma con la predicción del modelo de un 98% de probabilidad de desarrollar depresión.

Género	Facultad	Promedio	Estrato	Bogotá	Falta_Concentración	Tristeza	Soledad	Atrasado
Mujer	Ciencias Sociales	4.38	4	Sí	Si	Si	Si	No
Insatisfecho	Relación	Vivir_Solo	Amigos	Pensamientos_Muerte	Perdida_Placer	Vale_Pena	Inútil	Culpa
No	No	No	Sí	No	No	Sí	No	Sí
Act_Evitación	Act_Activación	Red_Apoyo	Satisfacción_Red	Falta_Apetito	Ejercicio	Alcohol_Drogas	Historial_Depresión	Desbalance_Sueño
3	Sí	Sí	4	No	Sí	No	No	No
Muerte_Familiar	Terminación_Relación	Heridas_Enfermedad	P_Económicos	P_Acamémicos	P_Hogares	Otro_Estres	Probabilidad	
No	No	Sí	No	No	Sí	No	2,21%	

Tabla 14: Individuo Representativo Clúster de Bajo Riesgo

El grupo de bajo riesgo está representado por una mujer de la facultad de ciencias sociales, de estrato 4, con un promedio por encima de 4, la cual no está atrasada o insatisfecha con su carrera; que durante la última semana ha experimentado falta de concentración, sentimiento de tristeza, sin pérdida de placer en actividades que antes consideraba sus favoritas, que siente que lo que hace si vale la pena, a pesar de poseer una red de apoyo y sentirse a gusto con ella siente que está sola; no ha experimentado pensamientos recurrentes acerca de la muerte, hace ejercicio regularmente y no consume alcohol o drogas de manera regular. A diferencia del grupo de alto riesgo, los individuos de este grupo se caracterizan por no

haber experimentado casi ningún evento estresor, en este caso únicamente problemas en el hogar. Dadas estas características es de esperar un nivel bajo de riesgo, lo cual se confirma con la predicción del modelo de un 2% de probabilidad de desarrollar depresión.

8 Conclusiones y Recomendaciones

La salud mental de la persona cobra cada vez más importancia debido a que esta, afecta en gran medida la forma como se relaciona con su ambiente, sus relaciones sociales, su desempeño en actividades diarias y puede desencadenar en serios problemas físicos de salud.

Sin embargo, el diagnóstico y tratamiento de los trastornos que afectan la salud mental del individuo resulta ser una tarea complicada debido a las comorbilidades que resultan como producto de estas y las restricciones de tiempo y cantidad de profesionales entrenados para identificar y tratar estas enfermedades. Así mismo, la falta de conocimiento por parte del paciente acerca de las enfermedades dificulta la visita a un profesional y, por ende, el diagnóstico correcto y tratamiento adecuado según el caso.

En la muestra seleccionada se encontró una prevalencia de la enfermedad del 56.72%, una cifra bastante alta y algo preocupante pues resulta ser mucho más alta que la prevalencia identificada en la población colombiana (25.1%) y más alta que la prevalencia, reportada por varios autores, en distintas universidades del país (aproximadamente 43%).

Mediante el trabajo realizado en este proyecto se buscó plantear una alternativa a las restricciones de tiempo y recursos existentes a la hora de realizar el diagnóstico de la enfermedad, por medio de una metodología estructurada basada en modelos matemáticos debidamente contruidos y validados, que sirviera como apoyo al profesional médico en la identificación y tratamiento de una de las enfermedades mentales de mayor prevalencia en la población mundial: la depresión.

Debido a la naturaleza del tema a tratar, fue necesario un arduo trabajo de investigación preliminar y un buen número reuniones con el grupo de expertos con el fin de entender la problemática y contextualizar el trabajo. Como resultado de esta etapa fue posible la selección del conjunto de variables de mayor relevancia en cuanto a su incidencia en que los estudiantes universitarios se encuentren en diferentes estados de riesgo de la

enfermedad; lo anterior resultó de gran utilidad, pues el desempeño de los modelos de predicción depende en gran medida al *input* que se les entrega; habiendo determinado un buen conjunto de variables se garantiza en cierta medida la competitividad de los modelos a construir.

Desarrollando los pasos de la metodología propuesta fue posible desarrollar un modelo con un desempeño muy alto y que además representa la realidad diagnóstica y el criterio experto. Gracias al tamaño y composición significativa de la muestra, los modelos fueron capaces de aprender la estructura de los datos y las relaciones entre las variables que llegaban a desencadenar en la presencia de la enfermedad en los estudiantes.

Después de probar varias metodologías predictivas, se optó por la combinación de predictores individuales, los cuales por separado mostraban de altos desempeños por sí solos, de modo que un nuevo modelo obtuviera las fortalezas de cada una y las utilizara para optimizar el poder predictivo del algoritmo. Gracias a la baja correlación entre las predicciones, se logró un aumento significativo en la calidad predictiva mediante el ensamblaje de modelos, presentándose como pieza fundamental en el desarrollo de un problema y una metodología predictiva.

Al haber logrado un desempeño diagnóstico satisfactorio, se utilizó los resultados del modelo para segmentar la muestra basado en su nivel de riesgo. Mediante la aplicación del algoritmo de agrupación *K Medoids* fue posible la identificación de dos grupos de riesgo prominentes dentro de los datos, así como las características representativas de cada uno de estos grupos. Mediante la identificación de individuos con riesgo alto o bajo desde sus características demográficas y hábitos de salud, se abre la posibilidad a un mejor diseño de programas y políticas preventivas y de tratamiento enfocados a cada nivel de riesgo, siendo adaptados a las necesidades particulares de sus individuos, facilitando su manejo y maximizando su utilidad.

Es importante también, resaltar las limitaciones que posee el estudio. La más notable siendo el diseño muestral recolectado. Se debe buscar una muestra de mayor tamaño totalmente aleatorizada y balanceada de acuerdo a las distintas poblaciones que componen el campus universitario; esto garantizaría que se trabaja con una muestra significativa y representativa de la realidad estudiantil en la universidad. Dadas las restricciones de recursos y tiempo del

estudio, se construyó una encuesta la cual fue aplicada a los participantes, quienes fueron seleccionados aleatoriamente de 610 transeúntes de los edificios ML y W durante la semana de toma de datos, la cual resulto ser representativa de las poblaciones generales de la universidad, pero presentaba falencias al ser comparada de manera más específica, todos problemas que pueden ser superados mediante un diseño muestral más adecuado.

Por otro lado, el panel de expertos que apoyaron el trabajo puede ser enriquecido mediante la adición de más profesionales con distintos enfoques psicológicos y psiquiátricos, de modo que la discusión permita nutrir mejores reglas de decisión, variables e inputs de entrada al modelo, así como instrumentos de mediciones y mejores prácticas para el registro de variables psicológicas. Como se mencionó anteriormente, mejorando la calidad de los elementos de entrada al modelo se garantiza de cierto modo un buen desempeño del mismo, pues existe relación entre las variables que puede ser usada por el modelo a la hora de realizar la predicción.

Como trabajo futuro se propone el análisis de distintas formas de predicción mediante modelos más complejos como una red bayesiana construida con base al criterio de expertos o redes neuronales profundas. Así mismo, es de gran relevancia garantizar la implementación del modelo construido en una herramienta computacional de interfaz amigable y de fácil uso y acceso, de modo que pueda ser utilizado por los profesionales sin necesidad de que tengan conocimiento detallado de la matemática del modelo que funciona por detrás.

9 Bibliografía

- Andresen, E., Malmgren, J., Carter, W., & Patrick, D. (1994). Screening for depression in well older adults: Evaluation of a short form of the CES-D. *American journal of preventive medicine*, 77-84.
- Arco, J., López, S., Heilborn, V., & Fenández, F. (2005). Terapia breve en estudiantes universitarios con problemas de rendimiento académico y ansiedad: eficacia del modelo. *International Journal of Clinical and Health Psychology*, 589-608.
- Bland, M. J. (2002). Validating scales and indexes. *BMJ*, 606-607.
- Bloom, D., Cafiero, E., E, J.-L., S, A.-G., & Bloom LR, e. a. (2012). The global economic burden of noncommunicable diseases. . *Program on the Global Demography of Aging*.

- Campo, A., Díaz, L., Rueda, G., & Barros, J. (2005). Validación de la Escala de Zung para depresión en Universitarias de Bucaramanga. *Revista Colombiana de Psiquiatría*, 54-62.
- Campo-Arias, A., & Oviedo, H. (2008). Propiedades Psicométricas de una Escala: La Consistencia Interna. *Revista de Salud Pública*, 831-839.
- Castillo Hernández, M. (2006). *Toma de Decisiones en las Empresas: Entre el Arte y la Técnica*. Bogotá: Uniandes.
- Galiatsatos, D., Konstantopoulou, G., Anastassopoulos, G., Nerantzaki, M., Assimakopoulos, K., & Lymberopoulos, D. (Septiembre de 2015). Classification of the most Significant Psychological Symptoms in Mental Patients with Depression using Bayesian Network.
- García, A. (2009). *La depresión en Adolescentes*. Madrid: Universidad Complutense de Madrid.
- González Llana, F., & Martín Carbonell, M. (2007). Instrumento de Evaluación Psicológica. *La Habana: Ciencias Médicas*.
- Gutiérrez Rodas, J. A., Montoya Veléz, L. P., Toro Isaza, B. E., Briñón Zapata, M. A., Rosas Restrepo, E., & Salazar Quintero, L. E. (2010). Depresión en estudiantes universitarios y su asociación con el estrés académico. *CES Medicina*, 7-17.
- Hospital Clínic de Barcelona. (2007). *La depresión: guía intermedia para pacientes con enfermedades de larga duración*. Barcelona: Fundación BBVA.
- Imandoust, S., & Bolandraftar, M. (2013). Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background. *International Journal of Engineering Research and Applications* , 605-610.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R*. Nueva York: Springer.
- Jaramillo-Toro, C., Martínez, J. W., Gómez-González, J. F., P, T. M., Otálvaro, S., & Sánchez-Duque, J. A. (2018). Sintomatología depresiva en una población universitaria de Colombia: Prevalencia, factores relacionados y validación de dos instrumentos para tamizaje . *Revista Chilena de Neuro-Psiquiatría*, 18-27.
- Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Techniques. En I. Maglogiannis, *Emerging Artificial Intelligence Applications in Computer Engineering* (págs. 3-24). IOS Press.
- Kumar, A. (2011). Classification and Regression Tree (CART) Techniques. En A. Arora, P. Malhotra, S. Marwah, A. Bahardwaj, Dahiya, & S, *Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets*. Nueva Delhi: Indian Agricultural Statistics Research Institute.

- Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 18-22.
- Lucas, P. (Diciembre de 2001). Bayesian Networks in Medicine: a Model-based Approach to Medical Decision Making.
- Ministerio de Salud. (2017). *Boletín de salud mental: Depresión, Subdirección de Enfermedades No Transmisibles*. Bogotá D.C.
- Morgan, J. (2014). *Classification and Regression Tree Analysis*. Report No 1. Boston University School of Public Health.
- Ojeme, B., Akazue, M., & Nwalih, E. (2016). Automatic Diagnosis of Depressive Disorders using Ensemble Techniques . *African Journal Of Computing & ICT*, 31-38.
- Organización Panamericana de la Salud. (1998). *Plan de acción de desarrollo y salud de adolescentes y jóvenes en las Américas, 1998-2001*. Washington, DC: OPS/OMS.
- Ramayah, T., Ahmad, N., Halim, HA, Mohamed Zainal, S., & Lo, M. (2010). Discriminant Analysis: An Illustrated Example. *African Journal of Business Management*, 1654-1667.
- RODAS, J. A., VÉLEZ, L. P., ISAZA, B. E., ZAPATA, M. A., RESTREPO, E. R., & QUINTERO, L. E. (2010). Depresión en estudiantes universitarios y su asociación con el estrés académico. *CES Medicina*, 7-17.
- San Molina, L., & Arranz Martí, B. (2010). *Comprender la Depresión*. Barcelona: Amat.
- Sau, A., & Bhakta, I. (2016). Predicting anxiety and depression in elderly patients using machine learning technology. *Healthcare Technology Letters*, 238-243.
- Sawer Radloff, L. (1977). The CES-D Scale: A Self-Report Depression Scale for Research in the General Population . *Applied Psychological Measurement*, 385-401.
- Siwek, K., & Osowoki, S. (2016). Data Minig Methods for Prediction of Air Pollution. *International Journal of Applied Math and Computerl Science*, 467-478.
- Sorias, S. (2014). Overcoming the Limitations of the Descriptive and Categorical Approaches in Psychiatric Diagnosis: A Proposal Based on Bayesian Networks. *Turkish Journal of Psychiatry*.
- Vázquez, C., Hernangómez, L., Hervás, G., & Nieto, M. (2005). Evaluación de la depresión. En *Manual para la evaluación cognitivo-conductual de los trastornos psicopatológicos* (págs. 1-80). Madrid: Pirámide.