SELECCIÓN ESTADÍSTICA DE BIOMARCADORES Y CARACTERÍSTICAS SOCIALES EN PACIENTES CON TRASTORNO AFECTIVO BIPOLAR TIPO 1, EN DOS CIUDADES DEL EJE CAFETERO

Luis Miguel Ramírez Sandoval

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍAS
PEREIRA
2015

SELECCIÓN ESTADÍSTICA DE BIOMARCADORES Y CARACTERÍSTICAS SOCIALES EN PACIENTES CON TRASTORNO AFECTIVO BIPOLAR TIPO 1, EN DOS CIUDADES DEL EJE CAFETERO

Luis Miguel Ramírez Sandoval

Proyecto de grado para optar al título de:
Ingeniero Electrónico.

Director
Mauricio Alexander Álvarez López, PhD.

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍAS
PEREIRA
2015

DEDICATORIA

Dedicado a familia, quienes siempre me han apoyado. Muchas Gracias.

AGRADECIMIENTO

Un especial agradecimiento al profe Mauricio Álvarez,
quien contribuyo enormente a alcanzar este sueño.

Agradezco también al Centro de Biología Molecular y Biotecnología
(CENBIOTEC) por permitirme el uso de su información,
sin la cual este estudio no hubiera sido posible.

Gracias a la Universidad Tecnológica de Pereira
por la financiación que le brindo a este estudio bajo
la convocatoria para la financiación de proyectos de grado de pregrado del 2013
y que gracias a esta fue posible obtener estos que resultados,
de forma que estos sirvan como contribución a la comunidad universitaria.

CONTENIDO

pág.

I. INTRODUCCION

El **trastorno afectivo bipolar tipo 1 (TAB-I)**, se encuentra clasificado por DSM-IV (*Diagnostic and Statistical Manual of Mental Disorders*), como una enfermedad en el grupo de los trastornos de ánimo, que se caracteriza principalmente por uno o varios episodios de manía acompañados por episodios de depresión [1]. Los episodios de depresión se caracterizan por baja autoestima, pérdida del apetito, pérdida del sueño, pensamientos negativos, de muerte y suicidio; en el estado de manía una persona con TAB-I puede presentar deficiencia en su auto control, incrementar el consumo de alcohol y de comida, pérdida de control en su temperamento, entre otros [2].

La fluctuación en el estado de ánimo en una persona afecta la relación con su entorno social, familiar y su desempeño laboral. Diferentes estudios epidemiológicos han encontrado un patrón de heredabilidad de la enfermedad en grupos familiares, lo que ha llevado a tratar de aislar los genes que influyen en el desarrollo de ésta. Cabe aclarar que los genes son partes cortas de ADN, que almacenan diferente tipo de información, como la forma de crear y procesar proteínas. El activador de la D-aminoácido oxidasa, el receptor δ de los activadores proliferativos de los peroxisomas, el transportador del neurotransmisor de serotonina (SLC6A4) también conocido como 5-HTT; son tan solo algunos de los genes que han sido identificados, como los que tienden a ser más susceptibles al TAB-I. No obstante y pese a la identificación de estos, los estudios que los han aislado, no presentan una relación estadísticamente significativa con la enfermedad. A pesar de los estudios realizados para comprender el perfil genético de la enfermedad, sus resultados no significan, en ningún caso, que la persona padezca o que padecerá la enfermedad; esto debido a que los resultados obtenidos no contemplan otro aspecto determinante en el desarrollo de la enfermedad, que es el componente social y personal del individuo; situaciones que ponen bajo diferentes presiones a las personas y que pueden ser detonantes de la enfermedad en una de ellas. El medio social en el que se desenvuelva una persona cumple un rol importante en el desarrollo de la enfermedad, por ejemplo, un entorno violento y agresivo puede causar en la persona depresión, estrés, e incluso consumo de drogas, entre otros; que sirven como desencadenante de episodios de depresión o manía, que junto con la predisposición existente en la persona, termina por convertirse en TAB. Sin embargo y pese a lo anteriormente descrito, los biomarcadores y características sociales y personales, que más inciden en la presencia y posterior desarrollo de la enfermedad continúan sin ser claros [2].

El gran impacto que posee la enfermedad en la vida diaria de una persona, y los riesgos potenciales que conlleva (como la muerte), ha incentivado estudios moleculares que permitan entender el componente genético de la enfermedad para así facilitar el diagnóstico y tratamiento. Estudios realizados en Colombia [2] [3], han analizado la relación entre el biomarcador 5-HTT con el desarrollo de la enfermedad; aunque ningúno presenta resultados estadísticamente significativos que reflejen una conexión fuerte entre el biomarcador y la enfermedad. No se encuentra en la literatura una especificación clara de los biomarcadores que permiten determinar la presencia o no de la enfermedad, pues

solo se han aislado algunos, que en ningún caso presentan una fuerte relación con la enfermedad. Sin embargo, otros estudios sostienen que la combinación de las reacciones de todos estos genes a la enfermedad, podría explicar el trasfondo genético de la misma. Encontrar los **biomarcadores genéticos** (respuesta de un gen o una secuencia de ADN que exhibe una especial susceptibilidad a la presencia de la enfermedad) y las características sociales o demográficas que mejor representan la enfermedad, permitirán determinar si una persona padece o no la enfermedad. De forma que al implementar algunas técnicas de análisis de datos, se puedan identificar esta características brindando una visión sobre el transfondo genético de la enfermedad.

Los estudios realizados sobre personas con TAB-I, han encontrado que parientes en primer grado de consanguinidad tienen una probabilidad promedio de 9 % de heredar la enfermedad [4]. Han sido también aislados los genes BDNF, DAOA, DISC1, GRIK4, SLC6A4 [5], sin embargo ningúno muestra una significativa relación con el TAB-I, pues su efecto sobre el perfil genético de una persona es similar al efecto producido por el TAB-II; también se han llevado a cabo estudios para tratar de diferenciar el genotipo de los trastornos bipolares de otras enfermedades mentales como la esquizofrenia [6], aunque se sugiere que la interacción conjunta de estos genes puede dar una mejor y más adecuada explicación del componente genético de la enfermedad [5].
El estudio Rengifo Ramos, et al., (2012), describió el comportamiento del gen SLC6A4, también conocido como 5-HTT, y su relación con la enfermedad, que en estudios previos ha mostrado relación con la enfermedad [7]; el estudio realizado en el Eje Cafetero, concluyó que la diferencia entre pacientes y controles en los genotipos y alelos del gen SLC6A4, fueron mayores en el alelo LL en los pacientes que en los controles, sin embargo, en los alelos SS, LS, L, S, las diferencias no fueron estadísticamente significativas; en comparación, con estudios realizados en Antioquia, el incremento del alelo LL fue ligeramente mayor, no obstante las frecuencias en los demás genotipos fueron similares; los resultados expuesto también son comparables con estudios realizados en Europa, España y Brasil, donde las diferencias no son significativas, sin embargo difirieron de estudios realizados en Japón, en donde la frecuencia del alelo SS fue mucho mayor a la registrada en el Eje Cafetero.

Para el tratamiento y análisis de datos biológicos se ha popularizado el uso de herramientas estadísticas [8], en donde el algoritmo *Lasso*, por sus siglas en inglés (*Least Absolute Shrinkage and Selection Operator*) [9], ha arrojado buenos resultados al momento de manejar bases de datos y obtener una buena regresión lineal entre un conjunto de variables y la salida o respuesta del problema a solucionar, por su precisión estadística y su viabilidad computacional. Con este algoritmo se logra reducir la dimensión de las bases de datos, por medio de parámetros conocidos como coeficientes Lasso, en donde se pueden eliminar por completo variables que no sean significativas y que no aporten a la solución del problema, esto mediante la reducción y la selección de las variables. Los resultados obtenidos mediante la implementación del Lasso, han sido ampliamente expuestos, además de desarrollarse técnicas para su optimización [10]. La combinación de selección y clasificación por medio de la implementación del Lasso, ha sido ya utilizada en estudios con información genética [11], en donde la implementación del Lasso

permitió reducir en más del 60 % la información genética utilizada en el estudio.
Es común en muchos casos de estudio, una vez se ha reducido la información a las características o variables más importantes, clasificar las personas objeto de estudio en grupos de acuerdo a los posibles resultados que presente uno u otro tipo de enfermedad, como es el caso de la mayoría de los ejemplos biológicos utilizados para mostrar el funcionamiento de los algoritmos estadísticos; dentro de los algoritmos de clasificación se pueden encontrar diferentes enfoques, que a su vez derivan en diferentes métodos de clasificación [12], con la regresión logística como uno de los métodos más utilizados y que arroja buenos resultados a la hora de clasificar problemas biológicos que involucran más de una variable [13]. La regresión logística estima la probabilidad de que una persona se encuentre en uno u otro grupo, sin embargo, ese modelo de regresión no realiza una selección de variables, de forma que para determinar esa probabilidad utiliza todas las variables involucradas en el problema. Un método derivado de esta, es la regresión logística regularizada, en donde se considera un parámetro de regularización que selecciona las variables que más contribuyen a la solución del problema [12]. Entre los métodos utilizados para seleccionar variables se pueden también mencionar Forward Stepwise Selection, que realiza la selección de variables dependiendo de si estas disminuyen o no el error del modelo lineal al momento de la clasificación [8]; también es posible mencionar un método que combina dos clases de regularización, llamado Elastic Net, al igual que Lasso, este método busca seleccionar las variables que mejor ajusten el modelo lineal a los datos [14].

II. OBJETIVOS

En esta capítulo se describen los alcances del proyecto.

1. OBJETIVO GENERAL

Determinar las características más relevantes de un conjunto específico de características, tanto genéticas como sociales, que permitan describir el Trastorno Afectivo Bipolar tipo 1 en pacientes de dos ciudades del eje cafetero y que a partir de estas se pueda clasificar a los pacientes con y sin la enfermedad.

2. OBJETIVOS ESPECÍFICOS

1. Realizar un análisis de estadística exploratoria en una base de datos con información sobre pacientes con TAB-I y población de control.
2. Estudiar la relevancia de las características biológicas y sociales en la clasificacíon de pacientes con TAB-I, empleando regresión regularizada.
3. Validar estadísticamente los resultados del algoritmo de regresión logística regularizada y métodos clásicos de selección de características.

III. TÉCNICAS DE ANÁLISIS DE DATOS

En este capítulo se expondrán brevemente algunas técnicas para el análisis de datos; estas permiten encontrar un modelo que simule el comportamiento de un conjunto de datos y a partir de estos estimar el valor de la respuesta. Los modelos lineales entregados por estas técnicas permiten ser utilizados tanto para resolver problemas de regresión como de clasificación. En los problemas de regresión se desea estimar el valor de una variable continua, esto en función del tiempo o de una variable independiente, para asi predecir el comportamiento de la respuesta ante el ingreso de nuevos datos. Los problemas de clasificación esperan determinar la probabilidad de una variable estar en una u otra clase, permitiendo separar el conjunto de datos, para realizar la separación de los datos se establece un umbral que se debe superar para ser clasificado en una clase o en la otra.
Los métodos utilizados se pueden ver de forma general en las ecuaciones (1) y (2) [8].

$$\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} [y_i(\beta_0 + \beta^T x_i) - \log(1 + e^{(\beta_0+\beta^T x_i)})] \right\} \tag{1}$$

$$\operatorname*{argmin}_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} (\alpha|\beta| + \frac{(1-\alpha)}{2}\beta_j^2) \right\} \tag{2}$$

La ecuación (1) representa el modelo general para el método de regresión logística, esta ecuación representa la máxima verosimilitud que existe entre las variables de entrada y la salida del problema, esto es la probabilidad de estar en una u otra clase. Los parámetros β representan la contribución o peso de cada variable en la solución del problema.
En la ecuación (2) los términos acompañados por λ son las penalizaciones L_1 y L_2. Esta ecuación es una forma general de los métodos Lasso y Elastic Net. Los términos β representan las contribuciones de cada variable al modelo lineal de regresión. El parámetro α indica si se implementarán ambas penalizaciones o solo una de ellas. El término N indica el número de observaciones [8].

1. REGRESIÓN LOGÍSTICA

La regresión logística es un modelo de clasificación, frecuentemente utilizado cuando la respuesta que se desea obtener es de tipo binaria, es decir, que solo pueda tomar dos valores. La regresión logística permite estimar la probabilidad de que la respuesta se encuentre dentro de una u otra clase. El modelo general se muestra en (1)[8]. En (1) los términos β_0 y β hacen referencia al término independiente y a los coeficientes estimados por el modelo; estos últimos representan la importancia o el peso que tiene cada variable del conjunto de datos para el modelo de selección. La regresión logística regularizada es una derivación de la regresión logística normal; esta es útil en situaciones en las cuales no solo se desea clasificar sino también seleccionar variables. Este método realiza primero un selección de variables y a partir de estas calcula los coeficientes para construir

el modelo de probabilidad. Para realizar la selección se requiere del cálculo de un parámetro de regularización (λ) que determinará si la variable permanece o es removida del modelo de probabilidad. Esto se logra a traves de técnicas, como validación cruzada, que buscan estimar el valor de este parámetro, el cual estima la cantidad de reducción aplicada a cada variable, de forma que se removerán algunas variables del modelo de probabilidad; para encontrar este paraḿetro, se separan en n subconjuntos los datos de entrenamiento y se escoge un grupo a partir del cual se realiza una predección sobre los otros datos, dando lugar así a un error de predicción, este procedimiento se repite de forma que se van incluyendo y retirando variables, hasta que la variable incluida no mejore el valor del error estimado. Con este valor, el algoritmo estima los coeficientes de las variables que permanecieron en el modelo. La ecuación (1) permite ser utilizada como modelo para la regresión logística regularizada, esto se realiza estimando un valor de λ mayor que cero, de forma que se aplique una reducción a las variables del problema. La ecuación (3) muestra la ecuación final utilizada para la regresión logística regularizada. El tipo de reducción aplicada a este método es la penalización L_1, la cual estima valores de reducción iguales a cero para aquellas variables que serán removidas de la solución y otros coeficientes pequeños para aquellas que permanecen [8].

$$\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} [y_i(\beta_0 + \beta^T x_i) - \log(1 + e^{(\beta_0+\beta^T x_i)})] - \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{3}$$

La probabilidad hallada en la ecuación (3) podrá tomar cualquier valor en un rango entre [0 1]. Sin embargo, dado la naturaleza binomial de la respuesta, se debe establecer desde que valor será considerado de una clase o de otra. Para dar igual probabilidad a una como a la otra, el valor será de 0.5, esto quiere decir que cada una tendrá un 50 % de probabilidad de estar en cualquiera de las clases [8].

2. LASSO

Si en la ecuación (2) se hace $\alpha = 1$ se obtiene el método LASSO (*Least Absolute Shrinkage and Selection Operator*), que es una técnica de análisis que permite disminuir la dimensión de un conjunto de datos, de forma que algunas variables tendrán un coeficiente igual a cero y otras se verán reducidas. De esta forma se mantendrán las variables que son más importantes para la solución del problema y con las cuales se podrían obtener los mismos resultados que si se retuvieran todas las variables [8].
Un método cercano a Lasso, es *Ridge Regression*, este es un método de análisis que disminuye el conjunto de datos. Para realizar esta reducción, el modelo devuelve unos parámetros conocidos como *Coeficientes Ridge* (β_{ridge}), estos coeficientes determinan el peso de cada variable para el modelo, por lo que el método devuelve el vector de coeficientes con una componente por cada variable del conjunto de datos, definiendo de esta forma la contribución de cada variable a la respuesta del problema. Los coeficientes Ridge se pueden hallar haciendo en la ecuación (2), *n = 1/2* y $\alpha = 0$ [8].
El parámetro λ en (2) es el encargado de determinar que tanto se debe reducir la contribución de una variable al modelo. Mientras más grande sea λ mayor será la reducción

de la variable.

Los métodos Lasso y Ridge, se diferencian en el tipo de penalización utilizada para estimar el porcentaje de reducción que se le aplicará a las variables. Para la regresión tipo Ridge, la ecuación utilizada por la regresión ridge utiliza la penalización L_2, $\lambda \sum_{j=1}^{p} \beta_j^2$, que devuelve un vector de coeficientes en donde su valor tiende a ser pequeño pero no exactamente cero, razón por la cual se mantienen todas las variables dentro del modelo. La penalización L_1, $\lambda \sum_{j=1}^{p} |\beta|$,que es utilizada por la ecuación implementada por lasso retorna coeficientes reducidos a cero y otros con una reducción no tan grande [8].
Una de las razones para implementar una reducción de dimensión al conjunto de variables, es la interpretación, es considerablemente más fácil interpretar un modelo con pocas variables, puesto que se podría relacionar directamente el comportamiento de cada término con la respuesta esperada del problema. Métodos como la regresión logística tradicional estiman un coeficiente para cada una de las variables, resultando una ecuación lineal con una gran cantidad de términos que dificultan la interpretación del modelo [8].

3. ELASTIC NET

Elastic Net es una técnica de análisis que combina las penalizaciones del método Ridge y de Lasso. La ecuación que describe al método de Elastic Net, se puede obtener al hacer en (2) $\alpha = 0{,}5$ [10].
Elastic net presenta ciertas ventajes sobre el Lasso y Ridge, puesto que esta selecciona grupos de variables de acuerdo a su correlación, por consiguiente, las variables serán conservadas o desechadas por grupos, al contrario de como lo hacen Lasso y Ridge, que a partir de un par de variables correlacionadas, escogerán una de las dos sin importar que la variable desechada represente mejor el conjunto de datos; Elastic net conservará las dos, escogiendo implícitamente la variable que más contribución tenga para que el modelo se ajuste mejor a los datos [14].

4. FORWARD STEPWISE SELECTION

Forward stepwise selection, es una técnica de análisis derivada de la selección del mejor subconjunto, que consiste en escoger un subconjunto de variables que describán la respuesta del problema igual a si se tuvieran todas las variables involucradas, este procedimiento sin embargo, es inviable cuando el número de variables es muy grande y mucho mayor al número de observaciones; Es un método que combina los métodos de Eliminación sucesiva hacia atras y Adición sucesiva hacia adelante; La eliminación sucesiva hacia atras, inicia con un modelo que involucra todas las variables medidas, con estas se ajusta un modelo lineal de regresión que trate de predecir la respuesta, de forma que se van eliminando variables a medida que estas no reduzcan el error en la clasificación. Por otra parte, la adición sucesiva hacia adelante realiza el procedimiento opuesto al método descrito anteriormente, esta inicia con un modelo sin ningúna variable y estas se adicionan a medida que las variables reduzcan el error de clasificación. La selección por pasos empieza con el modelo sin ningúna variable y va agregando variables,

luego se determina la constribución de todas las variables presentes en el modelo hasta ese momento, si la contribución de la variable se ha visto reducida, esta es decartada. Sin embargo, esto no significa que la variable no pueda regresar al modelo, de hecho, si mas adelante la variable aumenta su contribución esta será reintegrada al modelo [8].

IV. METODOLOGÍA

En este capítulo se describe el diseño metodólogico llevado a cabo durante el desarrollo del proyecto.

1. Base de datos

Como primer paso para el desarrollo del proyecto, se seleccionó la información que iba a ser utilizada por los algoritmos de análisis. La base de datos utilizada fue la recolectada por el Centro de Biología Molecular y Biotecnología de la Universidad Tecnológica de Pereira. La base de datos cuenta con información de 138 pacientes con TAB-I, 45 hombres y 93 mujeres, con edades de entre 12 y 77 años; también se encuentra información de 124 controles (personas que no padecen TAB-I), 62 hombres y 62 mujeres. Esta información fue recolectada en dos ciudades del Eje Cafetero, Armenia y Pereira, en dos centros de salud mental, Hospital Mental de Risaralda (HOMERIS) y la Clínica Especializada en Salud Mental El Prado en Armenia.

Las variables objeto de estudio fueron 22 en total, de las cuales 2 fueron variables genéticas (Promotor e Intrón del gen transportador del neurotransmisor de serotonina), estas están relacionadas con comportamientos suicidas y desarrollo tardio del TAB-I [2] y el resto fueron variables demográficas como el genero, la edad, etc. Se separó la información de la base de datos en dos sets. Uno que se utilizó como datos de entrenamiento y a partir de los cuales se determinaron los modelos para la clasificación. El otro juego de datos se utilizó como datos de prueba (test) para la clasificación y para con estos estimar el error de clasificación de los modelos encontradados. Los datos fueron separados y seleccionados de forma aleatoria. Como datos de entrenamiento se seleccionaron 106 de los cuales 55 fueron pacientes de control. Para los datos de test se utilizaron 151 datos, 78 fueron pacientes de control.

2. Análisis exploratorio

Con la base de datos seleccionada y catalogada se realizó un análisis exploratorio de la misma, que consistió en determinar la media o promedio y la desviación de cada una de las variables estudiadas. Se estimo de igual forma la correlación de Spearman entre cada una de las variables y la respuesta del problema (padece o no TAB-I). También se obtuvo un *boxplot* para cada variable; estos son útiles para mostrar ciertas propiedades de cada uno de los grupos de datos graficados como la media, la desviación, los valores máximos y mínimos, así como también datos atípicos, entre otros. Se realizó también un grafíco de dispersión de las variables estudiadas, de forma que se pudiese ver la relación entre las variables.

3. Normalización

Uno de los primeros procedimientos llevados a cabo en el desarrollo del proyecto fue el cálculo de la media o promedio y la desviación de cada una las variables. Con estos

datos se procedió a normalizar los datos, esto se realizo restando a cada uno de los datos el valor medio de cada variable y dividiendo por la desviación.

4. Implementación

El programa utilizado para el desarrollo e implementación de los algoritmos de selección y clasificación fue MATLAB® R2013a , bajo licencia de la Universidad Tecnológica de Pereira, para el desarrollo de los algoritmos se utilizaron los comandos y funciones integradas con el programa; estos algoritmos corrieron sobre un computador portátil marca DELL® con 4GB de memoria RAM, 500GB de disco duro, procesador Intel Core i3 de 2.13GHz y con sistema operativo Windows 7 Home Premium Service Pack 1 de 64 bits.

Para utilizar los comandos de MATLAB®, se deben ingresar como argumentos una matriz X, en donde cada columna corresponde a una variable y cada fila a una observación y un vector de salida o respuesta Y, que para el caso del proyecto es un vector binario, "0" para pacientes de control (No padece TAB-I) y "1" para pacientes diagnosticados (Padece TAB-I).

- El primer método implementado fue la Regresión Logística, el comando utilizado fue *mnrfit*, este retorna una matriz de coeficientes estimados para una regresión logística.

- Con la matriz encontrada se implementó el modelo de probabilidad de la ecuación (3). Se estableció que una probabilidad mayor al 50 % sería tratado como un caso con TAB-I y una menor como un paciente de control.

- Luego se implementó el método de Regresión Logística Regularizada, para esto se utilizó la instrucción *lassoglm*. De esta instrucción obtenemos un modelo lineal que contiene la contribución de cada variable al modelo de regresión. Una vez obtenido esto, se determinan los coeficientes que se hicieron cero y el λ o parámetro de regularización que minimiza el error cuadrático medio; el parámetro de regularización es escogido mediante el método de validación cruzada. Este método separa en sub grupos el conjunto de datos, una vez separados utiliza *K-1* de sub grupos, donde K es el número de sub grupos obtenido, esto con el fin de dejar el último sub grupo como sub grupo de test, con los demás sub grupos se calculan varios modelos de regresión que mejor se ajusten al sub grupo de datos K, de esta forma obteniendo el λ que minimiza el error. Con el λ también se extrae el término constante del modelo. Con los coeficientes encontrados obtenemos un modelo de probabilidad de regresión binomial que permite realizar predicciones sobre nuevos datos.

- Una vez determinado el modelo de probabilidad, se pasó a estimar el comportamiento del modelo para clasificar nuevos datos. Al aplicar el modelo, se separaron los datos de test en dos grupos, padece o no padece, de forma que si la probabilidad superaba el 50 % sería una persona con TAB-I, pero si fuese menor, la persona

sería parte del grupo de control. Con los datos clasificados, se compararon las respuesta del modelo con la de la base de datos y a partir de esto se calculó el error de clasificación del método.

- El siguiente método utilizado fue el método Lasso, para la implementación de esta tecnica de regresión se hizo uso de la instrucción *lasso*.
  A partir del modelo retornado por la instrucción se determinó cuantas variables fueron reducidas a cero y cuántas permanecieron en el modelo de regresión. Con el modelo lineal obtenido a partir de los coeficientes retornados por la instrucción Lasso, se clasificó el juego de datos de test. Posteriormente, se calculó el error de clasificación con respecto al vector de respuesta real de los datos de test.

- El tercer método implementado en el análisis de la base de datos, fue el Elastic net. Este método es un variante del método Lasso, por lo que para su implementación se utlizó la misma instrucción *lasso*, pero con el valor de alpha, α, igual a 0.5. Este valor representa el peso que tendra la contribución de las penalizaciones L_1 y L_2, es decir que tanto se va a utilizar una o la otra, para este caso se implementó la doble penalización. Con el modelo devuelto por la instrucción anterior, se identificaron los coeficientes que permanecieron despues de la regresión, identificando así cuales fueron las variables de mayor contribución para el problema. Con el modelo lineal obtenido, se procedió a clasificar los datos de test.

- El ultimo método utilizado fue el Forward Stepwise Regression, para la implementación de este método se utilizó la instrucción *stepwisefit*.
  Esta instrucción retorna un modelo lineal con las variables que más contribuyen en la solución del problema. Consiste en ir adicionando variables a un sub set de variables de forma que esta nueva variable minimize el error cuadrático medio; el modelo retornado contará con los coeficientes respectivos para las variables que permanecieron en la solución. Con el modelo lineal ya determinado, se procedió a clasificar la base de datos de test y a comparar el resultado del modelo con la respuesta real de la base de datos, calculando así el error de clasificación.

Como prueba adicional, se realizó el mismo proceso anterior pero en esta ocasión apliacado a los datos de entrenamiento y de test definidos previamente, pero separando la información en dos sub sets adicionales, uno con las variables genéticas y otro con las variables sociales. Para la implementación de cada uno de los modelos entregados por los métodos utilizados se deben multiplicar las variables seleccionadas por el método y los coeficientes entregados por este más el término constante, las demás variables se despreciarán, para obtener asi la respuesta de si padece o no TAB-I.

5. Validación

Para probar la significación estadística de los resultados, seguimos el procedimiento propuesto para la selección de modelos en [15]. Se dividio cada conjunto de datos en un conjunto de entrenamiento y un conjunto de validación. Se entrenaron los diferentes métodos que utilizarón el conjunto de entrenamiento y luego medir la precisión sobre

el conjunto de validación. Se repitio este procedimiento 10 veces con un conjunto de entrenamiento diferente y validación establecido por la repetición. Para estudiar si hay diferencias que sean estadísticamente significativas entre los clasificadores, se aplico primero una prueba Lilliefors de normalidad en los 10 repeticiones de cada clasificador. Si la hipótesis nula de normalidad es rechazada, se realiza una prueba de Kruskal-Wallis para comparar los rendimientos promedio entre los clasificadores. Si la hipótesis nula de igualdad de medias se rechaza, se realiza una prueba de comparación múltiple utilizando Tukey-Kramer para estudiar más a fondo que los clasificadores son diferente. Todos los niveles de significación se miden al 5 %.

V. ANALÍSIS Y RESULTADOS

En este capítulo se describirán y analizarán los resultados obtenidos una vez seguida la metodología del capítulo IV.

1. Análisis exploratorio

En esta sección se exponen los resultados obtenidos del análisis exploratorio.

Variables medidas en los pacientes con TAB-I

Variable	Media	Desviación
PRO	1.7	0.8
INT	2.2	0.8
GEN	1.3	0.5
EDA	43.4	15.5
LDN	17.8	18.5
DEP	2.6	1.9
ESC	3.3	1.9
NUH	1.3	1.7
CON	3.6	1.2
OCU	16.8	13.9
TRA	11.2	16.3
ESN	3.4	1.8
DIA	0.0	0.2
HIE	0.0	0.2
HIO	0.1	0.3
CUS	0.0	0.1
MIG	0.1	0.3
ULC	0.2	0.4
HTA	0.0	0.2
EPI	0.0	0.1
TRD	0.0	0.1
ALT	0.0	0.2

Tabla 1: Media y desviación de las variables medidas para los pacientes. Las variables se encuentran representadas por tres letras que son una abreviación de su nombre real.Tabla (19).

La tabla (1) muestra los valores de las medias y la desviación de cada variable estudiada, en ésta se puede ver que la mayoría de variables tienen una media igual o cercana a 0 al igual que su desviación, lo que indica que estas variables tienen una poca dispersión, es decir que la mayoría de sus datos se encuentran cerca a la media. Variables como EDA,LDN, OCU y TRA, entre otras, tienen una gran dispersión, esto debido al amplio rango que poseen estas variables. En la figura (1) se pueden apreciar los valores de la media y la desviación de forma gráfica. Se puede observar que las variables x_5, x_6, x_8, x_9, x_{10} y x_{11} cuentan con puntos por fuera del recuadro, estos puntos indican datos atípicos, estos son datos que estan por fuera de la desviación normal, sin embargo, se

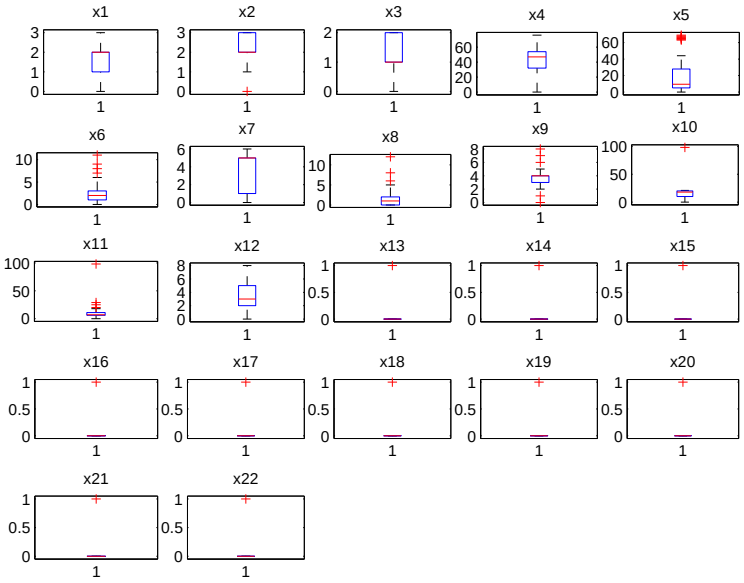

Figura 1: Gráfica de las medias de las variables para los pacientes con TAB-I. La letra x_i representa a cada una las variables de estudio.Tabla (19).

puede observar que la mayoría de datos se encuentran dentro de los limites establecidos, es decir, se encuentran dentro de la desviación estándar estimada. La figura (2) muestra la relación existente entre 8 de las variables estudiadas (de la variable x_4 a la variable x_{11}. Tabla (19)), es posible observar que las variables no guardan ningúna relación entre ellas, esto debido a que no es posible identificar un patrón de distribución en los datos.

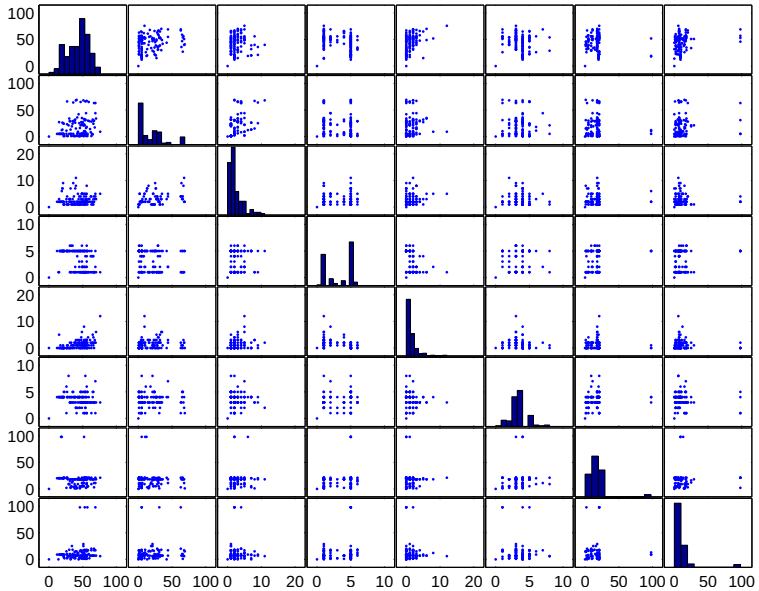

Figura 2: Scatterplot de 8 variables estudios. En la figura se muestra la relación existente entre las variables seleccionadas (variable x_4 a la variable x_{11}. Tabla (19)).

Variables medidas en controles (Sin TAB-I)

Variable	Media	Desviación
PRO	1.5	0.7
INT	2.3	0.7
GEN	1.5	0.5
EDA	38.3	13.0
LDN	21.9	22.4
DEP	2.4	2.2
ESC	3.0	1.9
NUH	1.3	1.3
CON	3.3	1.1
OCU	11.3	13.6
TRA	14.8	14.8
ESN	5.0	1.7
DIA	0.0	0.1
HIE	0.0	0.1
HIO	0.0	0.1
CUS	0.0	0.0
MIG	0.1	0.3
ULC	0.1	0.3
HTA	0.0	0.1
EPI	0.0	0.1
TRD	0.0	0.0
ALT	0.0	0.0

Tabla 2: Media y desviación de las variables medidas para los controles (Sin TAB-I). Las variables se encuentran representadas por tres letras que son una abrevación de su nombre real. (Tabla (19)).

La tabla (2) muestra los valores de las medias y la desviación de cada variable estudiada, tanto en ésta tabla como en la tabla (1), las variables (EDA,LDN, OCU y TRA) posean desviaciones similares, lo que indica que ambos grupos (las personas que padecen y las que no), tienen una distribución de datos similares. En la figura (3) se pude ver que a diferencia de la figura (1), se encontraron menos datos atípicos, es decir, todos los datos fueron estadísticamente similares. Se puede observar también que desde la variable x_{13} hacia adelante, las variables son de tipo binomial, esto también se puede observar en la figura (1). En la figura (4) se muestran la relación existente entre 8 de las 22 variables de estudio (de la variable x_4 a la variable x_{11}. Tabla (19)). En ella se puede ver que los datos no siguen una distribución normal y por el contrario los datos no siguen ningún patrón reconocible, por lo cual estas variables no guardan una relación aparente.

En la tabla (3) se exponen los valores de la correlación de Spearman para cada una de las variables estudiadas con respecto a la salida del problema, que para el caso de estudio es si padece o no TAB-I.

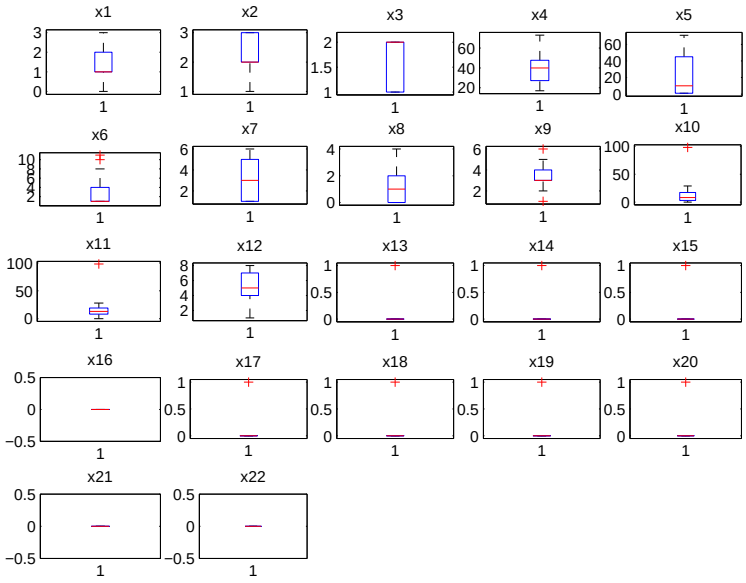

Figura 3: Gráfica de las medias de las variables para la población de control (sin TAB-I). La letra x_i representa a cada una las variables de estudio. (Tabla (19)).

Variable	Correlación
PRO	0.1
INT	-0.1
GEN	0.2
INT	0.2
LDN	0.0
DEP	0.2
ESC	0.1
NUH	-0.1
CON	0.1
OCU	0.3
TRA	-0.3
ESN	-0.4
DIA	0.1
HIE	0.1
HIO	0.2
CUS	0.1
MIG	0.0
ULC	0.1
HTA	0.1
EPI	0.1
TRD	0.1
ALT	0.1

Tabla 3: Correlación de Spearman entre cada una de las variables de estudio y la respuesta del problema (padece o no TAB-I). Las variables se encuentran representadas por tres letras que son una abrevación de su nombre original. (Tabla (19)).

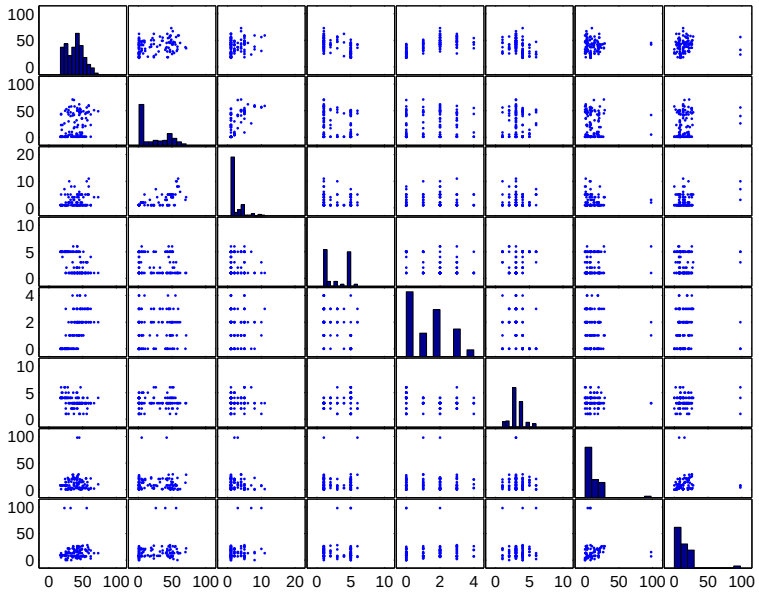

Figura 4: Scatterplot de 8 variables estudios. En la figura se muestra la relación existente entre las variables seleccionadas (variable x_4 a la variable x_{11}. Tabla (19)).

Se puede observar en la tabla (3) que las variables INT,NUH,TRA y ESN, poseen un coeficiente de correlación negativo entre -1 y 0, lo cual indica que existe una correlación negativa entre las variables y la salida. La variable LDN posee un coeficiente de correlación igual a 0, lo que indica que no existe una relación lineal entre esta variable y la presencia de la enfermedad, no obstante, si puede existir relaciones no lineales entre ambas. El coeficiente de las variables restantes son valores entre 0 y +1, lo cual indica una relación positiva entre variables, cuando una aumenta la otra también lo hará, aunque no en proporciones constantes; este mismo comportamiento se presenta en las variables con coeficiente de correlación negativa, cuando una disminuye la otra variable también lo hará aunque no en la misma proporción.

2. Análisis y resultados utilizando todas las variables biológicas y demográficas.

En esta sección se expondrán y analizarán los resultados obtenidos al utilizar el set de datos que contenia todas las variables (genéticas y demográficas).

Regresión Logística

Dado que este método no realiza reducción en los datos, la probabilidad de padecer o no la enfermedad es entonces una combinación lineal de todas las variables medidas. Con este modelo de probabilidad se procedio a clasificar nuevos datos; los resultados de la clasificiación se muestran en la tabla (4). La tabla (5) muestra de forma más detallada como se clasificarón los datos para uno de los sets de tests.

Porcentaje clasificación correcta	$27.9 \pm 2.9\,\%$

Tabla 4: Porcentajes de Clasificación

		Clases	
		No Bipolar	Bipolar
Predicción	No Bipolar	30	48
	Bipolar	39	34

Tabla 5: Predicción de datos para uno de los sets de test.

Se puede observar en la tabla (5) que sin realizar ningún tipo de reducción en la dimension de los datos, el algoritmo de regresión logística tradicional no es efectiva al momento de clasificar nuevos datos y por el contrario su comportamiento es bastante deficiente. Lo que indica que tener un conjunto amplio de variables no garantiza en ningún caso que estas describan de forma adecuada la presencia o no de la enfermedad, por el contrario, se podría decir que la interacción entre todas las variables contrarestan el efecto de poder realizar una predección con mayor relevancia estadística.

Regresión Logística Regularizada

Al ejecutar las instrucciones descritas anteriormente se extrajo de la información del modelo lineal que el valor de λ para el cual la desviación se encuentra dentro un error estandar minimo es de 0.1281; también fue posible determinar que solo cuatro variables permanecieron en el modelo entregado por la regresión logística regularizada, estas fueron:*GEN*, *EDA*, *LDN*, *TRA* y *ESN*.

Se puede observar en la figura (5) el comportamiento del valor del error cuadrático medio al momento de predecir nuevos datos. En la figura (5) se identifican los valores de λ más relevantes para el modelo. La linea de trazo largo marca el valor de lambda para el cual la desviación se encuentra dentro de un error cuadrático medio mínimo; por otra parte, la linea punteada pequeña demarca el valor de λ para el cual la desviación es mínima, esto sin importar el valor del error. Cabe resaltar, que el valor de λ determina el tamaño de la regularización que será aplicada a las coeficientes de regresión del modelo.

El valor de λ utilizado para hallar las variables importantes en el modelo, es el valor identificado por la linea de trazo largo; esto debido a que si bien la desviación es mayor a la desviación presentada por el otro valor (linea punteada), el error determinado mediante validación cruzada es mínimo.

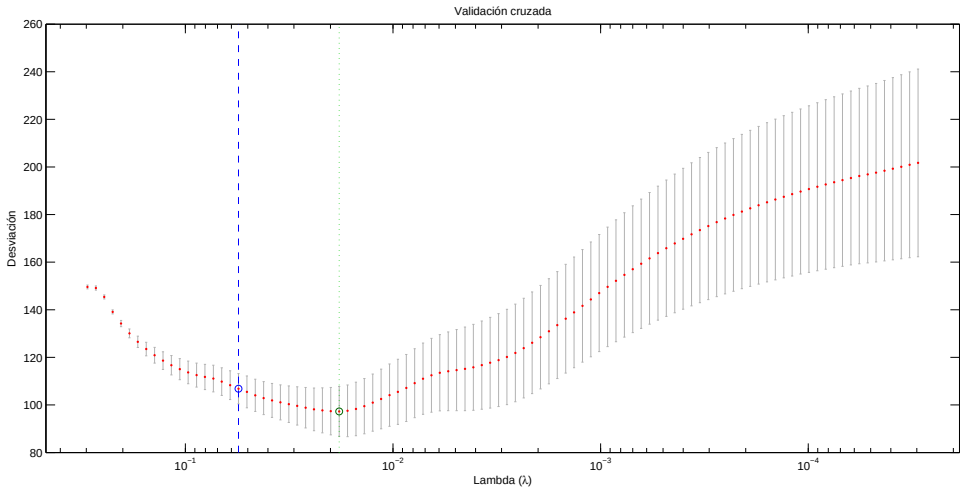

Figura 5: Gráfica del error cuadrático medio al momento de predecir nuevos datos con un modelo lineal ajustado a un valor de λ determinado, a medida que el parámetro de regularización λ disminuye.

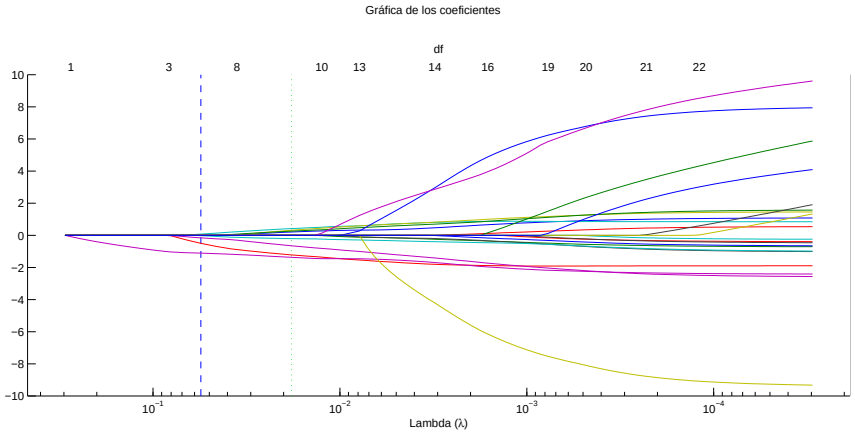

Figura 6: Número de variables no nulas con su respectivo valor estimado a medida que λ aumenta.

La figura (6) muestra el índice de las variables y su valor estimado (coeficiente) a medida que λ aumenta. En esta también se puede observar los dos valores de λ descritos anteriormente.
El modelo lineal de probabilidad obtenido se describe en (4):

$$\begin{aligned} Respuesta \sim\ & 0{,}8553 - 0{,}4966GEN + 0{,}0772EDA \\ & - 0{,}1538LDN - 0{,}0220TRA - 1{,}1096ESN \end{aligned} \tag{4}$$

En la ecuación (4) se muestran las variables que más contribuyen al modelo de regresión logística regularizada más un término constante (+1), en donde *Respuesta*, representa

la probabilidad de padecer o no la enfermedad. De la ecuación (4) se puede notar que el signo menos (-) que precede al coeficiente de cada variable indica una relación inversamente proporcional al valor de la variable medida, esto significa, que los hombres tienen una probabilidad menor de padecer la enfermedad y que las personas con un nivel de educación bajo tienen por el contrario una probabilidad mayor. Para la variable de la edad, variable para la cual el coeficiente esta precedido por un signo (+), la relación con la probabilidad de padecer la enfermedad es directamente proporcional, por lo cual a mayor edad mayor es la probabilidad de tenerla.
Con el modelo logístico de regresión (4) se pasó a clasificar la información de la base de datos de test. La tabla (6) muestra los porcentajes de error al momento de clasificar nuevos datos.

Porcentaje clasificación correcta	$68.2 \pm 3.9\,\%$

Tabla 6: Porcentajes de Clasificación

		Clases	
		No Bipolar	Bipolar
Predicción	No Bipolar	38	35
	Bipolar	27	51

Tabla 7: Predicción de datos para uno de los sets de test.

Es de notar que si bien la ecuación (4) solo presenta cuatro variables de todas las variables medidas, este modelo obtuvo un rendimiento superior al modelo de probabilidad encontrado por el método de regresión logística, al aumentar el porcentaje de acierto de nuevos datos a cerca al 60 %. De esta forma, la implementación del método de regresión logística regularizada presenta una reducción en el conjunto de variables en un 77 %. La selección de características mejoró notablemente el porcentaje de acierto en la clasificación de nuevos datos aumentando el porcentaje de clasificación correcta en casi un 20 %.

Forward Stepwise Selection

El modelo lineal obtenido mediante Forward Stepwise se muestra en la ecuación (5), en donde *Respuesta* es la estimación de si se padece o no la enfermedad.

$$Respuesta = -0{,}3200GEN - 0{,}2600ESN \tag{5}$$

Las variables *GEN* y *ESN* son las únicas que permanecen como parte de la solución del conjunto de variables estudiadas (tabla (19)). Ambos coeficientes están precedidos por un signo menos (-), lo que implica que la relación existente entre ellas y la respuesta es inversamente proporcional. A medida que la persona tiene un alto nivel educativo, menor es la contribución de esta variable a la respuesta del modelo, igual comportamiento presenta la variable que representa el género de la persona, para la cual una mujer tiene

mayor probabilidad de padecer la enfermedad.

La tabla (8) expone los resultados de la clasificación de nuevos datos mediante el método de Forward Stepwise.

Porcentaje clasificación correcta	$45.2 \pm 4.7\,\%$

Tabla 8: Porcentajes de Clasificación

		Clases	
		No Bipolar	Bipolar
Predicción	No Bipolar	36	37
	Bipolar	48	30

Tabla 9: Predicción de datos para uno de los sets de test.

De la tabla (8) se puede observar que el método de forward stepwise selection tiene un comportamiento deficiente al momento de clasificar nuevos datos, dado que solo el $43.7\,\%$ de los datos fueron clasificados correctamente. De esto se puede inferir que las dos variables escogidas por este método no son suficientes para describir la respuesta del problema. La tabla (9) muestra en detalle como se realizó la clasificación de los datos para unos de los sets de test.

Lasso

En la figura (7) se puede observar gráficamente los resultados obtenidos al aplicar el método Lasso. Del resultado entregado por el método Lasso, se establece que el número

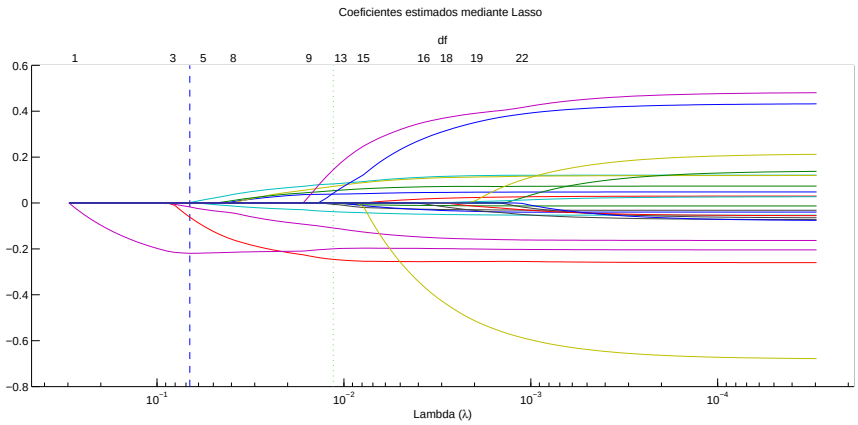

Figura 7: Coeficientes estimados por Lassso.

de variables que permanecen en el modelo y que por consiguiente son las que más contribuyen a la solución del problema son cuatro. El valor de λ para el cual la desviación se encuentra dentro de un error mínimo, es 0.0667.
El modelo lineal encontrando mediante el método Lasso se muestra en la ecuación (6).

$$Respuesta = 0{,}6060 - 0{,}0616GEN + 0{,}0020EDA - 0{,}0172LDN - 0{,}2193ESN \quad (6)$$

A diferencia del método de regresión logística regularizada, las variables más relevantes para el método lasso son *GEN*, *EDA*, *LDN* y *ESN*, dejando por fuera la variable *TRA*. No obstante, las variables presentan comportamientos similares, esto es, que las variables con un signo menos (-) en frente de ellas, presentan una relación inversamente proporcional con las respuesta del problema.
Con el modelo descrito en la ecuación (6) se realizo el proceso de predicción frente a la entrada de nuevos datos. Los porcentajes de la clasificación se muestran en la tabla (10).

Porcentaje clasificación correcta	$67.1 \pm 4.8\,\%$

Tabla 10: Porcentajes de Clasificación

		Clases	
		No Bipolar	Bipolar
Predicción	No Bipolar	42	31
	Bipolar	24	54

Tabla 11: Predicción de datos para uno de los sets de test.

El método lasso presentó un mejor porcentaje de acierto al momento de clasificar nuevos datos que los métodos de regresión logística regularizada, forward stepwise selection y regresión logística. En comparación a los métodos ya implementados se puede ver, que el método lasso, escoge un grupo de variables que describen mejor la presencia de la enfermedad.

Elastic Net

La figura (8) muestra el comportamiento de los coeficientes en función de λ. La ecuación (7) describe el modelo lineal formado por las variables seleccionadas por el método de elastic net, en donde *Respuesta* representa si la persona padece o no TAB-I.

$$\begin{aligned} Respuesta = {} & 0{,}6202 - 0{,}0716GEN + 0{,}0139EDA \\ & - 0{,}0256LDN - 0{,}0018TRA - 0{,}1468ESN \end{aligned} \quad (7)$$

Los resultados obtenidos al implementar el método de elastic net, son muy similares a los hallados por medio del método de regresión logística regularizada. Las variables con

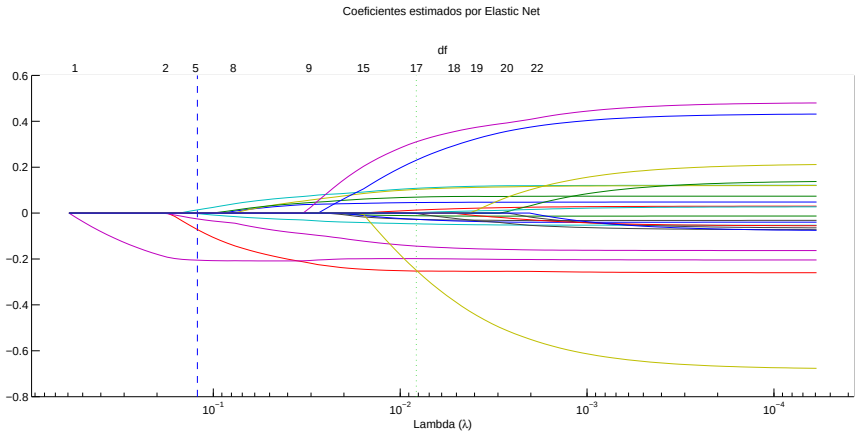

Figura 8: Coeficientes estimados por Elastic Net .

coeficiente negativo presentan una relación inversamente proporcional con la respuesta del problema.
Con el modelo de la ecuación (7) se realizo una predicción de clasificación sobre nuevos datos, los resultados de la clasificación se resumen en la tabla (12). De forma más detallada se puede observar en la tabla (13) como se realizó la clasificación de los datos mediante el método de Elastic Net.

Porcentaje clasificación correcta	$66.2 \pm 8.0\,\%$

Tabla 12: Porcentajes de Clasificación

		Clases	
		No Bipolar	Bipolar
Predicción	No Bipolar	39	34
	Bipolar	26	52

Tabla 13: Predicción de datos para uno de los sets de test.

Los resultados de la clasificación descritos en las tablas (4) (6) (8) (10) (12), muestran un comportamiento similar al momento de clasificar nuevos datos; con excepción de los métodos Forward Stepwise Selection y la Regresión Logística tradicional, los demas métodos implementados tienden a clasificar a un poco más de la mitad de los datos nuevos de forma correcta, aproximadamente el 60 % de ellos. La tabla (14) muestra los porcentajes de clasificación obtenidos por medio de la implementación del modelo lineal hallado para cada método.

Analizando la tabla (14), se puede ver que los métodos que mejor clasificaron nuevos

Clasificación \ método	Regresión Logística	Regresión Logística Regularizada	Forward Stepwise	Lasso	Elastic Net
Correcta	27.8 ± 2.9 %	68.2 ± 3.9 %	45.2 ± 4.7 %	67.1 ± 4.8 %	66.2 ± 8.0 %

Tabla 14: Porcentajes de Clasificación

datos fueron los métodos de Lasso, Elastic Net y Regresión Logística Regularizada. Estos métodos escogieron 4 (Regresión Logística Regularizada y Elastic Net) y 5 (Lasso) variables, estos tres métodos entregarón porcentajes de clasificación similares, esto es esperado por cuanto los tres métodos comparten la misma base para el cálculo del parámetro de regularización λ, como se describió en el capítulo III. Los métodos de Regresión Logística tradicional y Forward Stepwise Selection, tuvieron un comportamiento pobre al momento de clasificar nueva información, esto se debe principalmente a la estimación de los coeficientes que hace el modelo internamente, es decir, al momento de aplicar las ecuaciones expuestas en el capítulo (III). No obstante, todos los métodos tuvieron un comportamiento bastante bajo al momento de la clasificación.

La tabla (15) muestra las variables que fueron comunes para los métodos que mejor comportaminto tuvieron al momento de clasificar nuevos datos, los espacios vacios denotan que la variable no fue seleccionada por ese método. Cabe resaltar que las variables *GEN*, *EDA*, *LDN* y *ESN* fueron variables comunes para todos los métodos, lo que resalta su importancia para la predicción de nuevos datos, de forma general estas variables presentan un comportamiento inversamente proporcional a padecer o no la enfermedad (con excepcion de la variable *EDA*, que es directamente proporcional a la respuesta del modelo), de esto se puede concluir que son las mujeres las que tienen una mayor probabilidad de padecer la enfermedad, situación que es evidente en [2]. Es de notar que las variables geneticas *PRO* y *INT*, quedaron por fuera de todos los modelos encontrados, lo que podría indicar que el componente de herabilidad descrito[4] no esta correlacionado de forma estadísticamente significante como para ser tomado en cuenta al momento de predecir sobre nuevos datos.

3. Análisis y resultados utilizando solo las variables biológicas.

Se aplico la metodología propuesta en la capítulo (IV), separando del set de datos de entrenamiento las variables genéticas (*PROMOTOR* e *INTRON*) y dejando las variables demográficas y sociales como un set de datos independiente.
La tabla (16) muestra los porcentajes de clasificación de nuevos datos de los métodos que mejor porcentaje de acierto obtuvieron.

Variable \ método	Regresión Logística Regularizada	Lasso	Elastic Net
GEN	-0.4966	-0.0616	-0.0716
EDA	0.0772	0.0020	0.0139
LDN	-0.1538	-0.0172	-0.0256
TRA	-0.0220		-0.0018
ESN	-1.1096	-0.2193	-0.1468

Tabla 15: Coeficientes Estimados

Clasificación \ método	Regresión Logística Regularizada	Lasso	Elastic Net
Correcta	50.1 ± 2.9 %	50.2 ± 3.1 %	50.1 ± 2.7 %

Tabla 16: Porcentajes de clasificación para las variables genéticas.

Los resultados obtenido mediante el procedimiento de clasificación, no muestran mejoría con respecto a los descritos en la tabla (14), por el contrario los porcentajes decaen significativa, sin embargo, estos porcentajes no están lejos de aquellos descritos en [2], en donde la presencia de estas dos variables no es en ningún caso estadísticamente significante para describir la presencia o no de la enfermedad, de forma que la clasificación por medio de un modelo en el cual sólo se tienen presentas variables genéticas, no es suficientemente confiable para la predicción sobre nuevos datos.

4. Análisis y resultados utilizando solo las variables demográficas y sociales.

Las variables seleccionadas por los métodos de analisis estadístico en un conjunto de datos con la sola presencia de variables sociales y demográficas, se muestran en la tabla (18). Los espacios vacios indican que la variable no fue seleccionada en ese método.
Los porcentajes de clasificación para los modelos obtenidos con las variables descritas en la tabla (18) se muestran en la tabla (17)

Clasificación \ método	Regresión Logística	Regresión Logística Regularizada	Forward Stepwise	Lasso	Elastic Net
Correcta	26.3 ± 5.8 %	68.8 ± 3.8 %	42.6 ± 5.6 %	67.5 ± 2.4 %	68.1 ± 2.3 %

Tabla 17: Porcentajes de clasificación para las variables demográficas

En la tabla(17) se puede observar que la clasificación mejoró con respecto a los modelos en los cuales solo se tenian en cuenta las variables genéticas. Estos porcentajes tienden

método / Variable	Regresión Logística Regularizada	Lasso	Elastic Net
GEN	-0.3912	-0.1379	-0.1441
EDA	0.0450	0.0265	0.0405
LND	-0.1268	-0.0371	-0.0451
DEP			0.0024
CON		-0.0025	0.0122
TRA	-0.0220	-0.0091	-0.0151
ESN	-1.0950	-0.2164	-0.2085

Tabla 18: Coeficientes Estimados

a ser consistentes con los descritos en la tabla (14), en donde la mayoría de los modelos clasifican mas del 50 % de los datos nuevos.No obstante, es de notar que los modelos encontrados a partir de solo variables demográficas seleccionaron 2 variables nuevas, si bien los porcentajes de acierto de la tabla (14), tienden a ser mayores, no representan una diferencia estadísticamente significante.

A partir de los porcentajes de las tablas (14) y (17) se puede concluir que el incluir un mayor número de variables no garantiza una mejoría al momento de clasificar nuevos datos. Es posible también deducir, que las variables más significativas para determinar si una persona padece o no el TAB-I fueron *GEN*, *EDA*, *LDN*, *TRA* y *ESC*. Sin embargo, no es estadísticamente signicante dado el alto porcentaje de error presentado por todos los métodos para clasificar. Con relación a las variables genéticas, los pobres resultados entregados por los modelo hallados al momento de clasificar nuevos datos, muestran que estas están poco relacionadas con padecer o no el TAB-I, siendo este comportamiento descrito en otros estudios realizados en la región [2].

5. Análisis de significancia estadística

Al aplicar el procedimiento para la validación de la significancia estadística se obtiene la figura (9) en donde se muestra la distribución de las medias para cada uno de los métodos utilizados en la clasificación. La validación de la significancia estadística estimo que existen tres grupos con medias estadísticamente distintas. Los métodos que son estadísticamente iguales son regresión logística regularizada, lasso y elastic net, estos tres métodos conforman el grupo N°1. El siguiente grupo estadísticamente diferente es el conformado únicamente por el método Forward Stepwise, que sería el grupo N°2. De forma que el último grupo, es decir, grupo N°3, es el formado por el método de regresión logística. Los métodos del grupo N°1 poseen medias similares por cuanto las variables que seleccionaron son las mismas, de forma que la clasificación se realizó con las contribuciones de las mismas variables. El grupo N°1 es estadísticamente diferente

al grupo N°2 y al grupo N°3, y a su vez, los grupos N°2 y grupo N°3 son diferentes entre si. Esto se puede deber a que el grupo de variables seleccionado por cada uno de los métodos del grupo N°2 y grupo N°3 son considerablemente distintos al grupo de variables del grupo N°1.

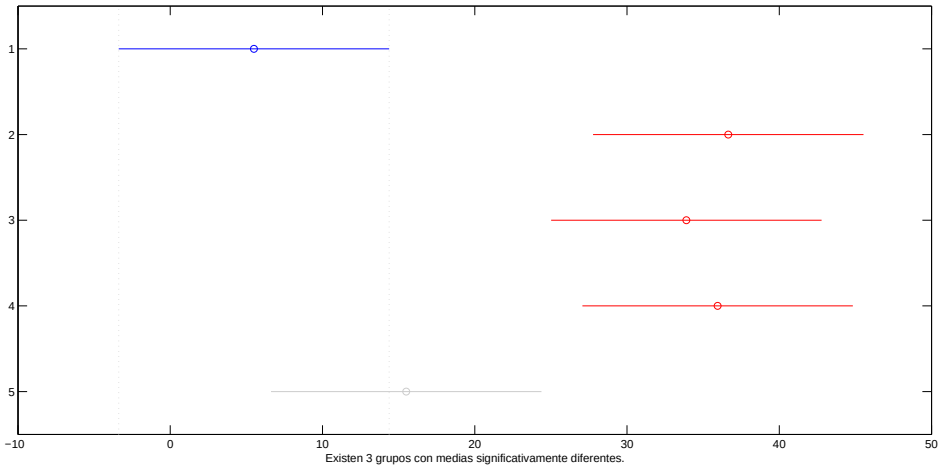

Figura 9: Gráfica de los grupos de medias detectados mediante el análisis de significancia estadística.

VI. CONCLUSIONES

- Las variables biológicas, Promotor e Intrón, no son estadísticamente significativas y no ayudan en la clasificación de nuevos datos. La heredabilidad descrita por diversos estudios [2], puede no deberse al componente genético de la persona, o las variables medidas podrían no ser las adecuadas para esta estimación. Como se mencionó previamente, algunos estudios describen que el componente hereditable de la enfermedad puede deberse a la interacción de diversos biomarcadores, de forma que las variables genéticas estudiadas pudieran ser sólo una parte de un conjunto de biomarcadores mayor, por lo cual estas variables aisladas no puedan describir a plenitud si se padece o no la enfermedad. Esto se puede observar en el estudio pues este estimó que a partir de estas dos variables la probabilidad de padecer la enfermedad es de aproximadamente del 50 %.

- Las variables demográficas son suficientes para determinar si una persona padece o no TAB-I, esto con un aproximado 68 % de acierto. De variables como el nivel de escolaridad o el trabajo, se puede concluir que el desarrollo de las personas en su ambiente social, laboral, familiar y personal, influye considerablemente en el desarrollo o manifestación de la enfermedad.

- El ambiente familiar o social que rodea a una persona puede influenciar en el comportamiento y en el desarrollo de la personalidad, de tal forma, que la persona se ve afectada por los comportamientos de los demás que lo rodean, consiguiendo que una persona en edad de desarrollo tome como ejemplo estos comportamientos y eventualmente desarrolle TAB-I; esto es evidenciado ya que las variables más significativas para la investigación fueron netamente demográficas y relacionadas a diferentes campos del desarrollo personal.

- Los métodos que mejores resultados entregaron fueron la Regresión Logística Regularizada, el Lasso y Elastic Net. Como previamente se mencionó estos tres métodos son estadísticamente iguales, por lo cual sus porcentajes de acierto fueron similares. Esto se puede deber a la forma en la que los tres métodos realizan la estimación de los coeficientes y la selección de las variables.

ANEXOS

Variable	Nombre	Identificador
x_1	PROMOTOR	PRO
x_2	INTRON	INT
x_3	GENERO	GEN
x_4	EDAD	EDA
x_5	LUGAR DE NACIMIENTO	LDN
x_6	DEPARTAMENTO	DEP
x_7	ESTADO CIVIL	ESC
x_8	NÚMERO DE HIJOS	NUH
x_9	CONVIVE CON	CON
x_{10}	OCUPACIÓN ACTUAL	OCU
x_{11}	TRABAJO DE MAYOR RESPONSABILIDAD	TRA
x_{12}	ESCOLARIDAD NIVEL	ESN
x_{13}	DIABETES	DIA
x_{14}	HIPERTIROIDISMO	HIE
x_{15}	HIPOTIROIDISMO	HIO
x_{16}	CUSHING	CUS
x_{17}	MIGRAÑA	MIG
x_{18}	ULCERAS	ULC
x_{19}	HTA	HTA
x_{20}	EPILEPSIAS	EPI
x_{21}	TRASTORNO DE DESARROLLO	TRD
x_{22}	ALTERACIÓN EMOCIONAL POSTPARTO	ALT

Tabla 19: Variables de estudio.

BIBLIOGRAFÍA

[1] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed. American Psychiatric Association, 1994.

[2] L. Rengifo Ramos, D. Gaviria Arias, L. Salazar Salazar, J. P. Vélez, and S. Lozano Pardo, "Polimorfismo en el gen del transportador de seratonina(slc6a4) y el trastorno afectivo bipolar en dos centros regionales de salud mental del eje cafetero," *Revista Colombiana de Psiquiatria*, vol. 41, no. 1, pp. 86–100, 2012.

[3] J. Ospina Duque, C. Duque, L. Carvajal Carmona, D. Ortiz Barrientos, I. Soto, N. Pineda, M. Cuartas, J. Calle, C. Lopez, L. Ochoa, J. Garcia, J. Gomez, A. Agudelo, M. Lozano, G. Montoya, A. Ospina, M. Lopez, A. Gallo, A. Miranda, L. Serna, P. Montoya, C. Palacio, G. Bedoya, M. McCarthy, V. Reus, N. Freimer, and A. Ruiz Linares, "An association study of bipolar mood disorder (type i) with the 5-httlpr serotonin transporter polymorphism in a human population isolate from colombia," *Neuroscience Letters*, vol. 292, no. 3, pp. 199–202, 2000.

[4] J. W. Smoller and C. T. Finn, "Family, twin and adoption studies of bipolar disorder," *American Journal of Medical Genetics*, vol. Part C, pp. 48–58, 2003.

[5] J. H. Barnett and J. W. Smoller, "The genetics of bipolar disorder," *Neuroscience*, vol. 164, no. 1, pp. 331–343, 2009.

[6] M. Logotheti, O. Papadodima, N. Venizelos, A. Chatziioannou, and F. Kolisis, "A comparative genomic study in schizophrenic and in bipolar disorder patiens, based on microarray expression profiling meta - analysis," *The Scientific World Journal*, vol. 2013, p. 14, 2013.

[7] M. T. Tsuang, L. Taylor, and S. V. Faraone, "An overview of the genetics of psychotic mood disorders," *Journal of the Psychiatric Research*, vol. 38, pp. 3–15, 2004.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.

[10] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[11] D. Ghosh and A. M. Chinnaiyan, "Classification and selection of biomarkers in genomic data using lasso," *Journal of Biomedicine and Biotechnology*, vol. 2005, no. 2, pp. 147–154, 2005.

[12] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[13] D. A. Blandón Salazar, "Comparación de maquinas de soporte vectorial vs. regresión logística ? cuál es más recomendable para discriminar?" Master's thesis, Universidad Nacional de Colombia, 2012.

[14] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society*, vol. 67, pp. 301–320, 2005.

[15] P. L. G. Joaquín Pizarro, Elsa Guerrero, "Muliple comparison procedures applied to model selection," *Neurocomputing*, vol. 48, pp. 155–173, 2002.