Clasificador Bayesiano usando Distribución Normal Multivariada para predecir el riesgo académico de los estudiantes de Pregrado de la Universidad Nacional de Colombia

Grupo Bayesianos:
Mónica Patricia Pineda Vargas
Raúl Ricardo Orcasitas Hernández
Ingeniería de Sistemas. Universidad Nacional de Colombia

Resumen—Este proyecto tiene como propósito usar un enfoque de aprendizaje de máquina implementando un clasificador de Bayes que permita predecir el riesgo académico de los estudiantes, mediante una clasificación de riesgo en tres posibles categorías: Bajo riesgo, riesgo moderado y alto riesgo, basado en ciertos factores socioeconómicos que influyen en el bajo rendimiento académico.

Palabras Clave—Probabilidad Condicionada, Bayes, Verosimilitud, Zona de Rechazo, Matriz de covarianza, Matriz de confusión.

1. Planteamiento del problema

Actualmente la Universidad Nacional de Colombia, desde el área de Bienestar Universitario junto con el Área de Acompañamiento integral está desarrollando un proyecto de excelencia y seguimiento académico a los estudiantes desde primeros semestres para evitar que factores .ªjenos.ªl ámbito académico puedan influir severamente sobre el rendimiento, aumentando la probabilidad de deserción y pérdida de la calidad de estudiante. Es un proyecto piloto que en este momento se está llevando a cabo con los estudiantes de primeros semestres de la facultad de Ingeniería y dependiendo de los resultados, poder ampliarlo a todos los estudiantes de la universidad. Teniendo en cuenta el gran número de estudiantes que tiene la Universidad, llevar a cabo un seguimiento a cada estudiante podría resultar una tarea difícil requiriendo una gran cantidad de personas para hacerlo.

Tomando como punto de partida esta iniciativa de Bienestar Universitario, con este proyecto se busca implementar una solución que permita clasificar a un estudiante según el riesgo académico en el que se encuentre (alto riesgo, riesgo moderado o bajo riesgo), analizando un conjunto de variables socioeconómicas y así, dependiendo de los resultados, el área de Acompañamiento Integral podría centrarse en aquellos estudiantes que presenten riesgo, apoyando y brindando asesoría a tiempo para disminuír la deserción académica.

2. Introducción

Basados en los resultados obtenidos anteriormente por estudiantes de la Escuela de Estadística de la Universidad Nacional de Colombia, sede Medellín en [1] y por la investigación realizada en la Sede Bogotá [5] donde se identificaron los factores principales que afectan el rendimiento académico y tomando como referencia los datos obtenidos de la encuesta socioeconómica de admisión se implementó un modelo de clasificación Bayesiano para distribución normal multivariada que permite clasificar a los estudiantes en tres grupos según el riesgo académico representado por esos factores. Para lo anterior, el modelo recibe el total de datos de los cuales el 80 % son usados para entrenar el modelo (datos de aprendizaje) y el restante es usado para realizar las pruebas y determinar la precisión del clasificador.

Inicialmente se aborda el fundamento teórico necesario, basado principalmente en la *Teoría de Decisión Bayesiana*, luego se explica la implementación del modelo y finalmente se realiza un análisis de los resultados obtenidos mediante pruebas de cien iteraciones diferentes que se resumen en una matriz de confusión porcentual, con el porcentaje de aciertos, errores y rechazos para cada una de las tres clases: Riesgo moderado, Alto riesgo y Bajo riesgo.

3. Teoría de Decisión de Bayesiana

El proceso de clasificación supervisada consiste en asignar un objeto o instancia **X** representado por un vector de atributos **X**=$[X_1, X_2, ..., X_d]^T$ a una clase C_i que pertenece a un conjunto total de clases C_k.

Un clasificador basado en la teoría de Bayes es un método de aprendizaje supervisado en el que las clases se conocen a priori, que permite deducir una clasificación de **X** en C_i a partir de un conjunto de datos de entrenamiento D que posee n ejemplos, cada uno de ellos compuesto a su vez por un vector de atributos y la clase a la que corresponde ($\mathbf{X}_1$, C_1), ($\mathbf{X}_2$, C_2)...($\mathbf{X}_n$, C_k).

3.1. Teorema de Bayes

(Bayes, 1764) [2] Sean A y B dos sucesos aleatorios cuyas probabiliades se denotan por $p(A)$ y $p(B)$, donde $p(B) > 0$. Suponiendo conocidas las probabilidades a priori de los sucesos A y B, es decir, $p(A)$ y $p(B)$, así como la probabilidad condicionada del suceso B dado el suceso A, es decir $p(B|A)$. La probabilidad a posteriori del suceso A conocido que se verifica del suceso B, es decir $p(A|B)$, puede

•

Texto desarrollado en Noviembre 25, 2014.

calcularse a partir de la siguiente fórmula:

$$p(A|B) = \frac{p(A)p(B|A)}{p(B)} \tag{1}$$

Para este caso, se debe calcular $P(C|\mathbf{x})$, donde $\mathbf{x} \in R^d$ (espacio Euclideano de d dimensiones) es el vector de atributos relacionados a un estudiante y C representa la categoría en la que puede ser clasificado: $C = [0, 1, 2]$ bajo riesgo, riesgo moderado o Alto riesgo respectivamente. Usando la regla de Bayes (1) se tiene que:

$$P(C|\mathbf{x}) = \frac{P(C)p(\mathbf{x}|C)}{P(\mathbf{x})} \tag{2}$$

Como el clasificador aprende de un conocimiento previo, se tiene la probabilidad a priori P(C), para este modelo P(C=0), P(C=1) y P(C=2) que son las probabilidades ya conocidas para cada una de las categorías, es decir, la probabilidad de que un estudiante pertenezca a una categoría independientemente del valor de **x**,

$$P(C_i) = \frac{\sum_j^N I(x_j)_i}{N} \tag{3}$$

Donde $I(x_j)_i$ es la función indicadora para la clase i definida de la siguiente manera:

$$I(x_j)i = \begin{cases} 1 & \text{si } x_j \in C_i \\ 0 & \text{en otro caso} \end{cases} \tag{4}$$

Y N es el número total de datos, y a su vez se cumple que:

$$\sum_{i=1}^{n} P(C_i) = 1 \tag{5}$$

$P(\mathbf{x}|C)$ denominada *"verosimilitud"*, es la probabilidad condicional de que un evento perteneciente a C tenga asociación con la variable **x**.

$P(\mathbf{x})$ es la probabilidad marginal de **x** que está denotada por:

$$P(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}|C_k)P(C_k) \tag{6}$$

Basado en lo anterior, el clasificador de Bayes calcula la probabilidad posterior $P(C_i|\mathbf{x})$ (2) para cada clase y selecciona la clase que tiene la probabilidad posterior más alta, es decir:

$$Selecciona \begin{cases} 0 \text{ (bajo riesgo)} & \text{si } P(0|\mathbf{x}) \text{ es } maxP(C_k|\mathbf{x}) \\ 1 \text{ (riesgo moderado)} & \text{si } P(1|\mathbf{x}) \text{ es } maxP(C_k|\mathbf{x}) \\ 2 \text{ (alto riesgo)} & \text{en otro caso} \end{cases}$$

3.2. Riesgos y pérdidas

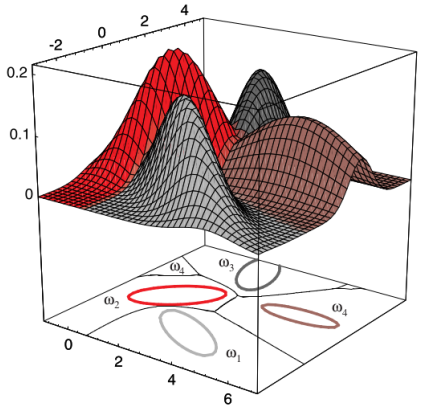


Figura 3.2.1: Regiones para cuatro categorías. Tomado de Duda [4]. Pág. 28

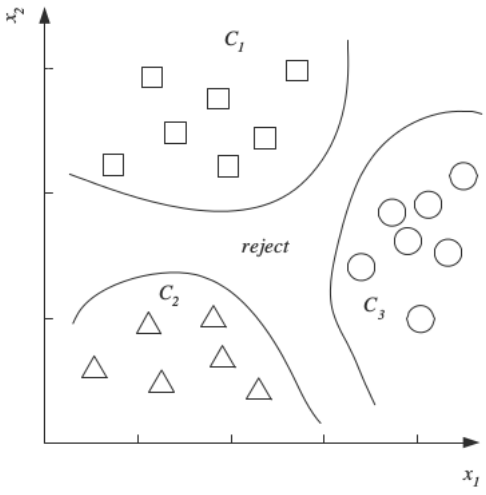
Un clasificador de K categorías divide el espacio en k regiones como se puede ver en la Figura 3.2.1 En este caso R_0 para la categoría 0, R_1 para la categoría 1 y R_2 para la categoría 2. Existe la posibilidad de que un elemento **x** que pertenece a la categoría 2 caiga en R_0, uno de la categoría 1 caiga en R_2, etc. Cuando se trata de estudiantes en riesgo, es preferible que el clasificador informe que no puede categorizar ciertos estudiantes a que estos sean mal clasificados; es por esta razón que se define el parámetro λ_{ik} para minimizar el riesgo de un error de clasificación. Como se puede observar en la gráfica 3.2.2, se genera una nueva región R_{k+1} (de rechazo) que a diferencia de la Figura 3.2.1 en donde las R_k regiones forman todo el espacio, en esta, las regiones R_k no hacen uso de todo el espacio, sino que se torna limitado, reduciendo así la probabilidad de error de clasificación. El tamaño de esta región de rechazo está determinado por el factor λ y el parámetro de minimización de error λ_{ik} está determinado de la siguiente manera:

$$\lambda_{ik} \begin{cases} 0 & \text{si } i = k \\ \lambda & \text{si } i = K + 1 \\ 1 & \text{en otro caso} \end{cases}$$

Donde $0 < \lambda < 1$. Por lo tanto, la regla de decisión queda definida como:

- Seleccionar C_i si $P(C_i|\mathbf{x}) > P(C_k|\mathbf{x}) \forall k \neq i$ y además $P(C_i|\mathbf{x}) > 1 - \lambda$
- De lo contrario selecciona la Región de rechazo.

Es importante tener en cuenta que cualquier pequeño cambio realizado en el valor de λ puede alterar los resultados; éste es un valor arbitrario, definido como más convenga para el modelo. Si λ es muy cercano a 1, la región de rechazo se hará muy pequeña, forzando al clasificador a ubicar a todos los datos en una clase, sin importar que queden mal clasificados; si por el contrario λ es muy cercano a cero, la región de rechazo se hará muy grande, lo que implica que la mayoría de datos no se clasificarán, por lo tanto es importante encontrar un punto de equilibrio que arroje los mejores resultados.



3.2.2: Región de rechazo.
Tomado de Alpaydin [10] Pág 54.

3.3. Función discriminante y fronteras de decisión

Como el denominador $P(\mathbf{x})$ en la ecuación (2) no varía para las C_i clases, se puede considerar como una constante, ya que el interés es calcular la máxima probabilidad entre todas las clases. Por lo tanto se tiene para cada clase:

$$g(\mathbf{x}) = P(C_i|\mathbf{x}) = \alpha P(C_i)p(\mathbf{x}|C_i) \tag{7}$$

Ignorando el término común para todas las clases (α), se tiene que:

$$g(\mathbf{x}) = P(C_i)p(\mathbf{x}|C_i) \tag{8}$$

La función $g(x)$ es conocida como función discriminante, como su nombre lo indica, será la encargada en ultima instancia de decidir a qué categoría o clase pertenece un dato x, como se observa, dicha función está completamente basada en la teoría de deción bayesiana, pero no calcula directamente la probabilidad sino un valor proporcional dado que se garantiza que $P(x)$ es el mismo valor para el cálculo de $P(C_i|x)$, lo que hace el contexto del algoritmo mucho más eficiente.

Como se vio anteriormente, el espacio queda divido en Regiones; es posible que estas regiones se crucen entre sí Figura 3.3.1 y al caer un dato en ese lugar, puede quedar mal clasificado. Por esta razón es importante determinar una frontera entre dos regiones, buscando siempre que se minimice el error de clasificación.

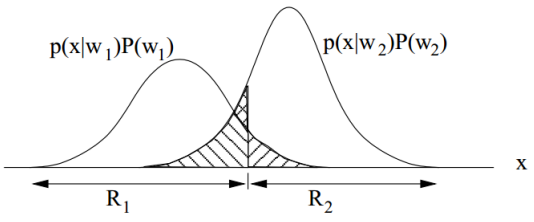


Figura 3.3.1: Cruce de Regiones
Tomado de Fernández [11]

3.4. Clasificación Multivariable

Cuando $\mathbf{x} \in \mathbb{R}^d$, es decir se tienen d variables, $p(\mathbf{x}|C_i)$ se puede tomar como la densidad para una distribución normal $\mathcal{N}_d(\boldsymbol{\mu}_i, \Sigma_i)$ ya que éste es un modelo para muchos fenómenos que ocurren naturalmente en la mayoría de ejemplos para clases, como en nuestro caso un muestreo de información socioeconómica y académica que son datos que no se presentan exactamente como una normal multivariable, pero si puede ser una aproximación útil para facilitar ciertos cálculos; por lo tanto la *verosimilitud* $P(\mathbf{x}|C_i)$ está dada por:

$$P(\mathbf{x}|C_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\left[\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right] \tag{9}$$

Donde:

- $\boldsymbol{\mu}$ es el vector de medias de tamaño d, formado por la media de cada variable. Es decir:

$$\mu = E(\mathbf{x}) = \begin{bmatrix} E[x_1] \\ E[x_2] \\ \cdot \\ \cdot \\ \cdot \\ E[x_d] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \cdot \\ \mu_d \end{bmatrix} = \sum_{\mathbf{x}} \mathbf{x}P(\mathbf{x}) \tag{10}$$

- **X** representa la matriz con la información de todos los estudiantes.

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ x_{N1} & x_{N2} & \dots & x_{Nd} \end{bmatrix} \tag{11}$$

  Para la distribución normal multivariable, **x** es cada fila de la matriz, es decir, cada estudiante.

- d es la dimensión del vector **x**, que representa el número de variables a analizar.

- Σ es la matriz de covarianza. La varianza entre de $\mathbf{x}_i$ está representada por σ^2 y la covarianza entre dos variables $\mathbf{x}_i$ y $\mathbf{x}_j$ está definida como:

$$\sigma_{ij} = COV(\mathbf{x}_i, \mathbf{x}_j) = E[(\mathbf{x}_i - \mu_i)(\mathbf{x}_j - \mu_j)] \tag{12}$$

  La matriz de covarianza es una matriz cuadrada de tamaño d, cuyas diagonales son las varianzas de $\mathbf{x}_i$.

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2d} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_d^2 \end{bmatrix} \tag{13}$$

- $|\boldsymbol{\Sigma}|$ y $\boldsymbol{\Sigma}^{-1}$ Son el determinante y la inversa de la matriz de Covarianza, respectivamente.

4. IMPLEMENTACIÓN

Para la recolección de datos, se buscó el apoyo por parte de La Dirección Académida de Sede, Admisiones y la División de Registro de la Universidad Nacional De Colombia, pero por ciertos inconvenientes no se logró contar con los datos necesarios; por esta razón, se procedió a la realización de un muestreo de datos entre algunos estudiantes de las diferentes Facultades de la Universidad, mediante una encuesta. Esta encuesta constaba de 15 preguntas:

- Semestre Actual.
- Promedio PAPA.
- Carrera.
- Género.
- Rango en el puntaje de admisión.
- Colegio del que proviene, público o privado
- Edad de ingreso a la universidad
- Tipo de vivienda: propia, arriendo o residencia universitaria
- Estrato.
- Tiempo de espera en semestres desde que se graduó de bachiller hasta el Ingreso a la Universidad.
- Si actualmente trabaja
- Si tiene hijos
- Lugar de procedencia
- Tipo de admisión (Regular, mejores bachilleres o programa especial).
- Ingresos mensuales propios o de la persona a cargo.

Las anteriores preguntas, fueron basadas en las investigaciones realizadas por la Escuela de Estadística de la Sede Medellín [1] y el estudio llevado a cabo en la Sede Bogotá mediante Minería de Datos [5], tomando como parámetros los resultados obtenidos en cuanto a las variables que más influyen en la deserción académica y pérdida de calidad de estudiante por bajo rendimiento académico. Estas encuestas se realizaron de dos formas: encuestas virtuales y encuestas físicas logrando un total de 686 datos. Como este clasificador funciona con aprendizaje supervisado, necesita de una información a priori definida para las clases. Tomando como referencia, información brindada por el Área de acompañamiento de la Universidad, se decidió definir para el conocimiento a priori, los rangos para las clases de la siguiente forma:

- Clase 0, Bajo riesgo: Promedio Aritmético Ponderado Académico (PAPA) superior a 4.0
- Clase 1, Riesgo Moderado: PAPA mayor a 3.4 y menor o igual a 4.0.
- Clase 2, Alto Riesgo: PAPA menor o igual a 3.4.

De la cantidad de datos obtenidos, clasificados con anterioridad, el 80 % "trainingSet"se usó para entrenar el modelo. El 20 % restante "testSet"fue usado para realizar las pruebas; a diferencia de los datos de entrenamiento, el "testSet"no se encuentra clasificado; es el modelo el que debe realizar esa labor con la información .ªprendidá"de los datos de de entrenamiento.

La implementación del software se realizó en Python (el repositorio se encuentra en github.com/MonicaPineda/ClasificadorBayesiano.git), basada en la teoría expuesta anteriormente, mostrando como método de análisis de resultados, una matriz de confusión porcentual para así poder observar el porcentaje de aciertos, errores y datos sin clasificar. La matriz de confusión porcentual, es el resultado de calcular la media aritmética de todas las matrices de confusión para cada prueba; en este caso fueron realizadas cien pruebas y sobre las 100 se calculó la media aritmética para cada elemento ij de la matriz de confusión.
La matriz de confusión es una matriz con k filas correspondientes a las C_k clases y con $k+1$ columnas que hacen referencia a las C_k clases y a la Región de rechazo.

$$\begin{bmatrix} \alpha & \beta & \beta & \gamma \\ \beta & \alpha & \beta & \gamma \\ \beta & \beta & \alpha & \gamma \end{bmatrix} \tag{14}$$

Donde α representa el porcentaje de datos que fueron clasificados correctamente para cada una de las clases. β representa los datos que fueron mal clasificados y γ los datos que el modelo no pudo clasificar.

Como se mencionó en la sección 3.2, es necesario definir un valor λ que se encuentra entre 0 y 1, para determinar qué tan grande es la región de rechazo en donde caerán los datos que el clasificador no pueda categorizar y así poder minimizar el error en la clasificación. Para determinar el rango con los posibles mejores valores, se realizaron mil variaciones de λ obteniendo los resultados presentados en la gráfica 6.1 para las tres clases.

Ya determinado el mejor rango, es decir aquel en el que el porcentaje de aciertos aumenta considerablemente y la tasa de fallos disminuye, se realizaron mil pruebas nuevamente variando en ese rango definido, para obtener una mejor visión de cómo se comporta el clasificador para esos posibles λ. Finalmente, como la categoría más importante es la de alto riesgo, se realizó una nueva ejecución con mil iteraciones desde 0.999999 hasta 1; teniendo en cuenta que es preferible no clasificar un dato a clasificarlo mal, se realizó un "score"penalizando el doble el porcentaje de fallos respecto a los datos no clasificados, mediante la siguiente ecuación:

$$sum = \%aciertos*0 + \%fallos*1 + \%noclasificados*0{,}5 \tag{15}$$

$$score = \frac{1}{sum} \tag{16}$$

4.1. Pseudocódigo

Data: Encuesta de estudiantes ($dataSet$)
Result: Matriz de confusión porcentual basada en 100 clasificaciones de el número total de estudiantes, para cada valor de lambda
$\lambda_k \leftarrow$ CalcularLamda(1000 repeticiones)
while λ_k *not null* **do**
  for $l \leftarrow 1$ **to** 100 **do**
    $trainingSet, testSet \leftarrow$SepararDatosPorClasesAleatoriamente(por
    de entrenamiento)
    $\mu_i, \Sigma_i \leftarrow$CalcularParametrosEstadisticos($trainingSet$)
    for *testSet* **do**
      $clasificacion_i \leftarrow$ Clasificar($\mu_i, \Sigma_i, \lambda_k, testSet_i$)
    end
    $matrizConfusion \leftarrow$
    CompararPrediccionConRealidad($clasificacion$,$dataSet$)
  end
  $matrizConfusionPorcentual_k \leftarrow$
  mediaAritmetica($matrizConfusion_l$)
end

Algorithm 1: Pseudocódigo para el experimento ejecutado

5. PROBLEMAS PRESENTADOS

Al no obtener los datos reales, la confianza de los resultados disminuye ya que no se tiene en cuenta muchos estudiantes y no se posee la garantía acerca de la veracidad de los datos. Como no se contó con la suficiente información, en las primeras pruebas se presentaron cierto tipo de errores por falta de variación en los datos. Una de las preguntas realizadas fue el tipo de admisión; de los estudiantes encuestados, menos de 10 provenían de un tipo de admisión diferente a la regular; al seleccionar el trainingSet, como se hace de forma aleatoria, existía una alta probabilidad que esos datos quedaran por fuera y al realizar los cálculos se obtenía una matriz singular ya que para esta variable (Tipo de Admisión) los datos no presentaban una variación (varianza = 0), por lo tanto no se podía calcular Σ^{-1} lo que detenía el proceso de clasificación. Al no contar con los suficientes datos y al no contemplar la posibilidad de inventar datos, la única solución que se podía implementar era descartar esta variable.

Cuando se hizo la investigación sobre un método que permitiera representar mejor los resultados obtenidos, encontramos varias medidas de precisión para métodos de aprendizaje multiclase (F1 Score) pero ninguna de ellas admitía cálculos para la región de rechazo que se implementó en el modelo. La decisión llevada a cabo para solucionar este problema fue realizar nuestra propia medida. Se realizaron cien pruebas y se calculó la media aritmética para la matriz de confusión, obteniendo como resultado una matriz de confusión porcentual con los procentajes de aciertos, fallos y rechazos para cada clase.

Otro de los problemas presentados fue el tener que tratar con decimales de órdenes de 10^{-17} para el valor de λ. Al hacer las mil variaciones de este parámetro, surgían problemas con las operaciones de incremento porque la precisión decimal del flotante de 32 bits en python no era suficiente para representar los incrementos de λ para estos órdenes de magnitud. Por esto, inicialmente se decidió realizar pruebas con aumentos de escala de órdenes de magnitud, sin alterar los resultados. Luego se realizaron pruebas usando la librería "decimal"de Python, que permitió usar la precisión necesaria para los cálculos ya que permite una precisión decimal de 128 bits.

Al trabajar con un total de 14 variables y debido a la brecha de conocimiento a la hora de realizar una reducción de dimensionalidad, no fue posible generar un diagrama de los clusters de datos para cada categoría.

6. ANÁLISIS DE RESULTADOS

Para la primera variación general de λ, entre cero y uno con mil iteraciones, se obtuvieron los siguientes resultados:

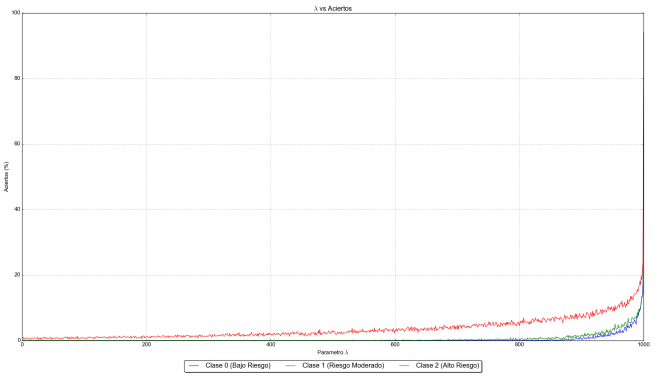


Figura 6.1: Porcentaje de aciertos para mil variaciones de λ entre 0 y 1

Se puede observar que a partir del 80 %, es decir cuando λ se acerca cada vez más a 1, el porcentaje de aciertos aumenta notablemente. Es importante resaltar que la clase con más aciertos respecto a las otras es la Clase 2 que representa la categoría de alto riesgo, lo que da a entender que el porcentaje de fallos y datos sin clasificar para esta categoría es mucho menor que para el resto.

Para las siguientes mil iteraciones y para poder ampliar la visión de los resultados, se tuvo en cuenta el rango del 20 % final de las iteraciones presentadas en la gráfica 6.1, obteniendo como resultados los mostrados en las siguientes imágenes:

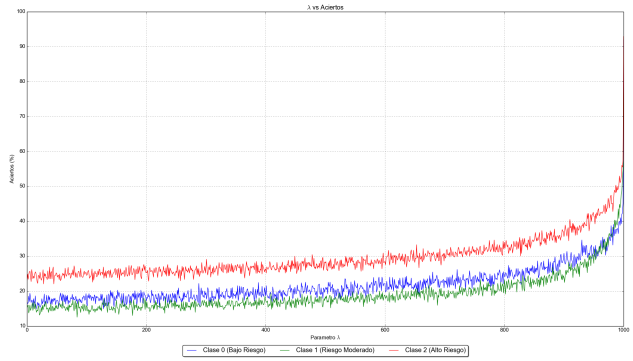


Figura 6.2: Porcentaje de aciertos para mil variaciones de λ entre 0,999 y 1

Se puede notar que aún es posible hacer un enfoque en los últimos datos, lo que conlleva a un aumento en la precisión de λ en 6 órdenes de magnitud menos; es decir, realizar una variación desde 0.999999999 a 1, obteniendo los siguientes resultados:

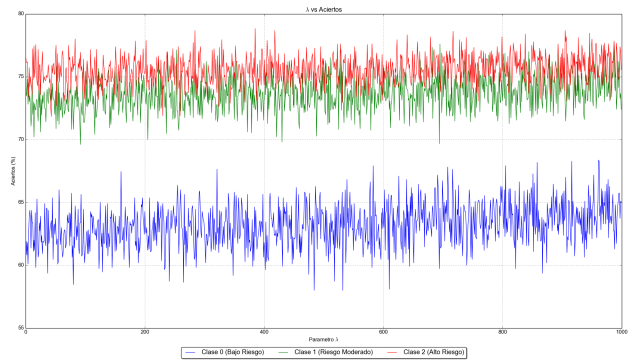


Figura 6.3: Porcentaje de aciertos para mil variaciones de λ desde 0,999999999 *hasta* 1

Como se puede ver en la Figura 6.3, el porcentaje de aciertos en los datos clasificados para la Clase 2 (categoría de alto riesgo) es el más alto, superando un 70 % del total de los datos para esta clase, seguido por la clase de riesgo moderado, que dificilmente y en contadas ocasiones, acierta menos del 70 %. La clase que presentó menos aciertos, fue la correspondiente a la categoría de estudiantes con bajo riesgo académico, pero se puede notar que su total de aciertos no baja del 55 %. El otro porcentaje de datos que no fueron clasificados correctamente, se encuentran distribuidos entre los datos mal clasificados y los datos que no se pudieron clasificar como se presenta a continuación en las gráficas 6.4 y 6.5.

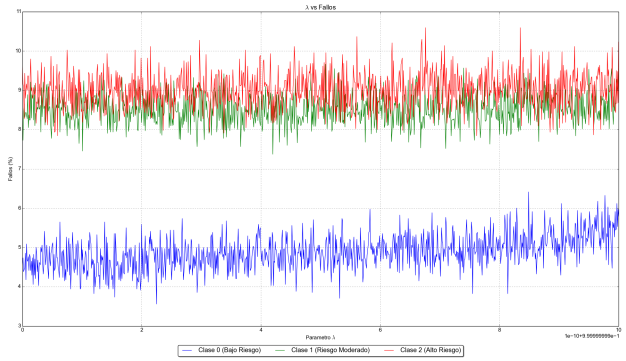


Figura 6.4: Porcentaje de fallos para mil variaciones de λ desde 0,999999999 *hasta* 1

Con los datos obtenidos de la gráfica anterior, se puede observar que el porcentaje de fallos de clasificación para ninguna de las clases superó el 11 %, la categoría con menos fallos fue la categoría cero, con un rango de fallo entre el 3 y 7 %. Las categorías 1 y 2, se mostraron parejas en el rango de fallo, ubicándose entre el 7 % y el 11 %.

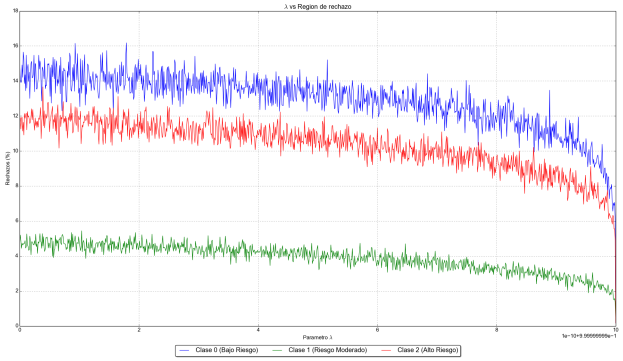


Figura 6.5: Porcentaje de datos no clasificados para mil variaciones de λ desde 0,999999999 *hasta* 1

Finalmente, se tiene el porcentaje de datos que el clasificador no pudo categorizar en ninguna de las 3 clases y fueron ubicados en la categoría de rechazo que no superó el 17 % de los datos para ninguna de las clases. Se puede notar que como era de esperarse, este porcentaje de rechazo se ve reducido a medida que el valor de λ se acerca más a 1.

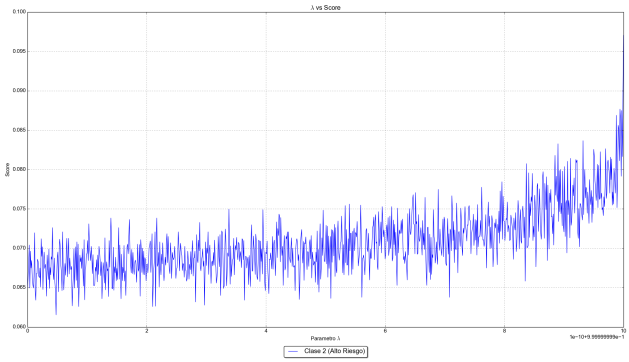


Figura 6.6: Score para la Categoría 2

Analizando individualmente el comportamiento de la Clase 2 (alto riesgo) frente a los diferentes valores de λ y los resultados del score presentado en la ecuación (14) que se muestran en la gráfica 6.5, se determinó que el mejor λ para el clasificador en general es

$$\lambda = 0{,}99999999999900002212172 \tag{17}$$

Que mostró un score total de $0{,}08988764044943819975231$, el más alto para las mil iteraciones. Al realizar la prueba de clasificación con ese λ, se genera como resultado la siguiente matriz de confusión porcentual para 100 iteraciones.

	Bajo R	R Moderado	Alto R	Rechazo
Bajo	88.235	3.529	0.0	8.235
Moderado	5.0	87.258	6.129	1.613
Alto	0.0	6.591	88.636	4.773

Como se puede observar, para las tres categorías, más del 85 % de los datos quedaron bien clasificados, 88.235 % de aciertos para la clase 0, 87.258 % para la clase 1 y 88.636 % para la clase 2. En el caso de los fallos, ninguno de ellos superó el 12 % y en la zona de rechazo quedaron clasificados el 8.235 % para la clase 0, el 1.613 % para la clase 1 y el 4.773 % para la clase 2. Es importante resaltar que como se ve en la posición (Bajo, Alto R.) de la tabla, a pesar de las 100 iteraciones, ninguna de las personas en alto riesgo fue clasificada en la categoría 0 (bajo riesgo) lo que brinda una confiabilidad mayor en el modelo y de la misma forma ninguna de las personas en bajo riesgo fue clasificada como si presentara alto riesgo académico (posición (Alto, Bajo R) en la matriz).

7. Conclusiones

Como se pudo ver en los resultados obtenidos, el clasificador de Bayes resultó ser una buena opción para este tipo de proyecto. A pesar de no contar con los datos reales de los estudiantes de Universidad Nacional, obtenidos en la encuesta socioeconómica realizada por Admisiones, lo que resta confianza a la información, los resultados superaron las espectativas mostrando un porcentaje de clasificación fallida muy bajo, en especial para la categoría que representa el alto riesgo académico estudiantil.

Con el porcentaje de acierto obtenido, se puede ver que este es un proyecto viable, que podría ser implementado en la Universidad, permitiendo llevar un control de riesgo académico de todos los estudiantes ya que para las personas que estarían encargadas de remitir la información de riesgo al Área de

Acompañamiento, se reduciría notablemente el volumen de datos, teniendo que analizar solamente los casos en los que los estudiantes no pudieron ser clasificados y cayeron en la región de rechazo. Si el Área de Acompañamiento tiene conocimiento de todos los casos de los estudiantes en riesgo, podrían mirar estos casos con tiempo, para así probablemente reducir la deserción estudiantil y la pérdida de calidad de estudiante por bajo rendimiento académico.

8. Trabajo Futuro

A pesar de los buenos resultados obtenidos con el clasificador de Bayes implementado, sería un interesante proyecto de investigación hacer una comparación de resultados usando otros métodos de aprendizaje como Redes Bayesianas, Naive Bayes, árboles de decisión, Máquina de soporte de vectore ycombinar diferentes tipos de aprendizaje, usando reducción de dimensionalidad para poder graficar los clusters de datos por categorías.

A futuro se podría llevar a cabo una implementación del proyecto en la Universidad; como se pudo observar con los resultados obtenidos, el clasificador podría beneficiar a los estudiantes que en muchos casos por falta de información, no hacen uso de los recursos y ayudas brindadas por el Área de Acompaãmiento Integral de la Universidad, pudiendo aumentar con el paso del tiempo su riesgo académico y de la misma forma aumentar la probabilidad de perder la calidad de estudiante por bajo rendimiento o desertando de la Universidad.

9. Agradecimientos

Agracedemos a Miguel Alexander Chitiva Díaz, estudiante de Ingeniería de Sistemas de la Universidad Nacional de Colombia quien gracias a tutorías, apoyo biliográfico y atención fue una ayuda importante para el desarrollo de este proyecto.

Bibliografia

[1] J. C Salazar, C. M. Lopera y M. C. Jaramillo, Un modelo de supervivencia para datos discretos en la identificación de factores que afectan el rendimiento academico universitario, Escuela de Estadística, Universidad Nacional de Colombia, Sede Medellín, 2010.

[2] T. Bayes (1764). Essay towards solving a problem in the doctrine of chances. The Phylosophical Transactions of the Royal Society of London.

[3] Tom Mitchell. Machine Learning, McGraw-Hill, 1997.

[4] Pattern Classification and Scene Analysis, Duda and Hart.

[5] C. E. López, E. León, F. A. González, Data Mining Model to Predict Academic Performance at the Universidad Nacional de Colombia, 12th Latin American and Caribbean Conference for Engineering and Technology, Guayaquil, Ecuador.

[6] S. Kotsiantis, C. Pierrakeas, and P. Pintelas. (2003). "Preventing student dropout in distance learning systems using machine learning techniques," in Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst., Oxford, U.K.

[7] Kotsiantis. (2009). "Educational data mining: a case study for predicting dropout-prone students," International Journal of Knowledge Engineering and Soft Data Paradigms.

[8] J. F. Superby, J. P. Vandamme, and N. Meskens. (2006). "Determination of factors influencing the achievement of the first-year university students using data mining methods," in Workshop on Educational Data Mining, Boston, USA.

[9] A. Ardila, Predictors of university academic performance in Colombia. International Journal of Educational Research. 2001.

[10] E. Alpaydin, Introduction to Machine Learning, second Edition, The MIT Press Cambridge, Massachusetts London, England

[11] F. Fernández y D. Borrajo, Aprendizaje automático. Grupo de planificación y Aprendizaje (PLG). Universidad Carlos III de Madrid. 2009.