UNIVERSIDAD NACIONAL DE ASUNCIÓN

Facultad Politécnica

"Análisis de Componentes Principales Categorizado aplicado al Aprendizaje de Matemática "

Trabajo Final de Grado presentado por

Emilio Gerardo Sotto Riveros

como requisito
para obtener el título de Licenciado
en Informática

Orientador:
D.Sc. Christian E. Schaerer Serra, Tutor
M.Sc. Santiago Gómez Guerrero

San Lorenzo - Paraguay. Diciembre de 2017

Índice General

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

Las matemáticas constituyen la base fundamental de áreas como la física, la ingeniería, la informática, la arquitectura, pero se la conoce como una asignatura difícil para estudiantes de diferentes niveles. Identificar los factores que afectan el proceso de aprendizaje de los estudiantes es clave para ayudar a promover mejores niveles en su desempeño en el campo de las matemáticas en la escuela, la universidad y como profesionales. Por lo tanto, descubrir los factores reales en el proceso de aprendizaje de los estudiantes es clave para ayudar a promover mayores niveles de éxito en su vida futura en la universidad y como profesionales.

Una gran cantidad de datos complejos y heterogéneos disponibles puede representar un reto en el momento de identificar cuales son las variables mas influyentes en el rendimiento estudiantil, considerando esto, el análisis de componentes principales como una herramienta importante que permite encontrar múltiples correlaciones en el conjunto de datos, mientras que se reducen muchas variables a unos pocos factores relevantes.

Algoritmos de minería de datos utilizados para predecir el desempeño de estudiantes en cursos de programación de nivel univesitario han demostrado un buenos resultados para identificar variables y extraer reglas que ayuden a identificar varia-

bles en el rendimiento estudiantil en temas específicos, analizando datos de antecedentes matemáticos de los estudiantes, aptitud de programación, habilidades para resolver problemas, experiencia previa en programación de computadoras, nivel de matemáticas de escuela secundaria, localidad y uso de plataformas e-learning . Esto beneficia el trabajo de educadores y expertos en currículo [1].

Estudios sobre programas educativos con orientación científica como STEM [10], demostraron que los estudiantes de bajo rendimiento tenían tasas de mejora significativamente más altas en los puntajes de matemáticas que los estudiantes de alto y medio rendimiento, y además, el origen étnico y social-económico eran buenos predictores del rendimiento académico [13]. En América Latina, este tema es aún más dramático para las carreras de ingeniería y ciencias [8] [7] pero los factores pueden ser diferentes.

Se ha visto que la inclusión de información de los estudiantes del tipo económica, demográfica, social y cultural ha ayudado a la clasificación y determinación de patrones para la creación de perfiles de rendimiento académico con el objetivo principal de utilizar aquellos tendientes al fracaso o deserción estudiantil como base para la determinación de futuras políticas de gestión académica[2]. Dada la naturaleza de los datos categóricos de la encuesta, se opto por el análisis de componentes principales categorizado, cuya utlidad se demostró al determinar el impacto de la emigración de los familiares en el rendimiento estudiantil de adolescentes [5]

Este trabajo de tesis trata sobre la identificación de los factores que afectan el correcto desempeño y la calidad del razonamiento en las matemáticas, de los estudiantes del 3er. ciclo de la Educación Escolar Básica y la Educación Media, procesando y analizando datos de sus rendimientos en pruebas matemáticas y datos relacionados a su vida escolar y familiar, usando como herramienta el análisis de

componentes principales categórico.

1.1. Relevancia y Originalidad

Dada la creciente necesidad de fomentar el interes por las matemáticas en estudiantes del nivel medio, es de vital importancia la identificando de los patrones y correlaciones entre las variables que influyen en sus desempeños, teniendo en cuenta el bajo nivel presentado por nuestro país en comparación a los demas países de la región [8], y dando asi el primer paso hacia investigaciones más profundas que busquen determinar causa y efecto.

El objetivo de esta tesis es la identificación de los factores que afectan el correcto desempeño y la calidad del razonamiento en las matemáticas, de los estudiantes en etapa de educación escolar básica, procesando y analizando datos de sus rendimientos en pruebas matemáticas y datos relacionados a su vida escolar y familiar, usando como herramienta el análisis de componentes principales categórico.

Por tanto, el planteamiento de los objetivos específicos se presenta de la siguiente manera:

- Caracterizar el comportamiento de los estudiantes con mayores rendimientos, para la elaboración de un perfil de los alumnos de alto desempeño.
- Determinar los factores socio-económicos que influyen en el aprendizaje de matemáticas.
- Proponer condiciones para establecer un escenario ideal para el estudio exploratorio de rendimiento de estudiantes.

Finalmente, la contribución del presente trabajo, se pueden resumir a grandes rasgos como sigue:

Si bien existen estudios que analizan el rendimiento estudiantil hechos de forma regional, hasta el momento no se ha realizado un estudio en Paraguay sobre estudiantes de alto rendimiento de la EEB enfocado en sus desempeños en las matemáticas, este trabajo busca ser el primero de varias investigaciones relacionadas al área del aprendizaje de matemáticas y por consiguiente ser una guía de referencias.

1.2. Estructura del Trabajo

Este trabajo está organizado de la siguiente manera.

- En la Sección 2 se presenta una fundamentos teóricos del Análisis de Componentes Principales .
- La Sección 3 informa detalles sobre la recopilación de los datos mediante el test de matemáticas, la encuesta, la limpieza de los datos, la formación de grupos de variables y el procesamiento de los datos.
- En la Sección 4, se muestra resultados de cada grupo analizado detallando los grupos con mejores resultados. También la formación de un perfil de los mejores estudiantes y un perfil de la mayoría de los alumnos.
- Finaliza con conclusiones y futuros trabajos en la Sección 5.

Capítulo 2

Fundamentos Teóricos

En este capítulo se hará una revisión de las nociones de la teoría de la mineria de datos (Data Mining) y la técnica de análisis de componentes principales que son mencionadas a lo largo del presente trabajo y cuyo interés radica en que pueden ser utilizados con el objeto de la reducción de variables e indentificación de patrones.

La minería de datos se define como un proceso que tiene como propósito descubrir, extraer y almacenar información relevante de amplias bases de datos, a través de programas de búsqueda e identificación de patrones y relaciones globales,tendencias, desviaciones y otros indicadores aparentemente caóticos que tienen una explicación que pueden descubrirse mediante diversas técnicas de esta herramienta [12].

2.1. Análisis de Componentes Principales - ACP

El Análisis de Componentes Principales (ACP) pertenece a un grupo de técnicas estadísticas multivariadas. Es una técnica descriptiva que se ha vuelto atractiva para una variedad de campos, porque ayuda a reducir la complejidad inherente a tener múltiples variables, permitiendo descubrir interrelaciones entre los datos y de acuerdo a los resultados, proponer los análisis estadísticos más apropiados. Igualmente se utiliza para reducir la dimensionalidad de la matriz de datos, construyendo las variables no observables llamadas componentes, con el fin de evitar redundancias. En

la mayoría de los casos, tomando sólo los primeros componentes, se puede explicar la mayor parte de la variación total contenida en los datos originales.

Para aplicar esta técnica es necesario que las varibles sean contínuas y el número n de individuos o elementos observados debe ser mayor que el número p de variables originales. El ACP permite reducir la dimensionalidad de los datos; transformando un conjunto original de variables p en otro conjunto de q variables no correlacionadas donde $p \geq q$, llamados componentes principales.Las variables p se miden en cada uno de n individuos u objetos, obteniendo una matriz de datos de orden np, donde $p < n$.

2.1.1. Aplicando el Análisis de Componentes Principales Real

Cuando todas las variables son numéricas, se puede aplicar ACP real (también llamada estándar). El análisis se realiza en el espacio de las variables y en el espacio de los individuos. Los puntos variables y los puntos individuales se representan gráficamente utilizando los componentes principales calculados como nuevos ejes de coordenadas; es una práctica bastante común usar solo los primeros dos componentes [9].

A menudo, puede facilitar la interpretación de los resultados para observar la similitud de ubicación de los puntos en este nuevo sistema de coordenadas. Aunque el plano de puntos variables no se superpone al plano de los puntos individuales, es útil interpretar la proximidad de un grupo de puntos individuales a ciertas variables.

Si X es un vector aleatorio de dimensión p con matriz de varianza-covarianza finita $p \times p$ $V[X] = \sum$, entonces el ACP resuelve el problema de encontrar las direcciones de la mayor varianza de las combinaciones lineales de observaciones. En

otras palabras, busca el conjunto ortonormal de vectores de coeficientes $a_1, \ldots, a_k$ tal que

$$\mathbf{a}_1 = \underset{\mathbf{a}:\|a\|=1}{\arg\max}\ \mathbb{V}[\mathbf{a}'\mathbf{x}]$$

$$\vdots$$

$$\mathbf{a}_k = \underset{\substack{\mathbf{a}:\|a\|=1 \\ a\perp a_1,\ldots,a_{k-1}}}{\arg\max}\ \mathbb{V}[\mathbf{a}'\mathbf{x}]$$

$$\vdots$$

donde $\|a\|$ es la norma de a. Los máximos son los de una función convexa en un conjunto compacto, y así existen, y son únicos si no existe una colinealidad perfecta en los datos, hasta el cambio del signo de todos los elementos de $\mathbf{a}_k$. La combinación lineal $a_k'X$ se conoce como el componente principal k-ésimo (PC).

La motivación detrás de este problema es que las direcciones de mayor variabilidad dan "más información"sobre la configuración de los datos en el espacio multidimensional.

Sea sea $X^t = [X_1 X_2 \ldots X_p]$ el vector aleatorio de las variables p con la matriz de varianza-covarianza P donde$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ y $a_1, a_2, ..., a_p$ son los valores propios y vectores propios correspondientes a P. Considere las siguientes combinaciones lineales:

$$Y_1 = a_1^t X = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$

$$Y_2 = a_2^t X = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p$$

$$\vdots$$

$$Y_p = a_p^t X = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p$$

Se puede probar que

$$Var(Y_i) = a_i^t \sum a_i \qquad i = 1, 2, ..., p$$
$$Cov(Y_i, Y_j) = a_i^t \sum a_j \qquad i, j = 1, 2, ..., p$$

Los componentes principales son las combinaciones lineales $Y_1, Y_2, \ldots, Y_p$ que no están correlacionados entre sí y cuyas varianzas satisfacen $Var(Y_1) \geq Var(Y_2) \geq ... \geq Var(Y_p) \geq 0$. Los componentes principales se definen a continuación:

- El primer componente principal es la combinación lineal $Y_1 = a_1^t X$ que maximiza $Var(a_1^t X)$ sujeto $\langle a_1, a_1 \rangle = 1$.
- El segundo componente principal es la combinación lineal $Y_2 = a_2^t X$ que maximiza $Var(a_2^t X)$ sujeto a $\langle a_2, a_2 \rangle = 1$ y $Cov(Y_1, Y_2) = 0$.
- En general, el componente principal i-ésimo es la combinación lineal $Y_i = a_i^t X$ que maximiza $Var(a_i^t X)$ sujeto a $\langle a_i, a_i \rangle = 1$ y $Cov(Y_i, Y_k) = 0$ para $k < i$.

Por lo tanto, bajo este método tenemos el siguiente resultado: considere la matriz de varianza-covarianza $\sum$ asociada con el $X^t = [X_1 X_2 \ldots X_p] \in R^p$, y sean $(\lambda_1, a_1), (\lambda_2, a_2)...(\lambda_p, a_p)$ valores propios y vectores propios correspondientes a la matriz $\sum$ donde$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$.

Entonces el i-ésimo componente principal está dado por:

$$Y_i = a_i^t X = a_{i1} X_1 + a_{i2} X_2 + ... + a_{ip} X_p \qquad i = 1, 2, ..., p$$

sujeto a las siguientes condiciones:

$$Var(Y_i) = a_i^t \sum a_i = \lambda_i \qquad i = 1, 2, ..., p$$
$$Cov(Y_i, Y_j) = a_i^t \sum a_j = 0 \qquad i \neq j$$

La mayor proporción de la varianza total de la población explicada por los componentes principales está dada por

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_k} \qquad k = 1, 2, ...p \tag{2.1}$$

El primer componente principal tendrá la mayor varianza y extraerá la mayor cantidad de información de los datos; el segundo componente será ortogonal al primero, y tendrá la mayor variación en el subespacio ortogonal al primer componente, y extraerá la mayor información en ese subespacio; y así. Además, los componentes principales minimizan la norma L_2 (suma de las desviaciones cuadradas) de los residuos de la proyección en subespacios lineales de dimensiones 1, 2, etc.

La primera CP da una línea tal que las proyecciones de los datos en esta línea tienen la suma más pequeña de desviaciones al cuadrado entre todas las líneas posibles. Las dos primeras CP definen un plano que minimiza la suma de las desviaciones al cuadrado de los residuos, y así sucesivamente.

El análisis de componentes principales se puede llevar a cabo tanto para las distribuciones teóricas como para los datos reales. En este último caso, se analizaría la matriz de covarianza empírica.Trazar varios primeros componentes unos contra otros a menudo puede dar una buena idea de la estructura de los datos, la presencia de clústeres, no linealidades, valores atípicos, etc.

Hay una serie de opciones prácticas que los investigadores deben tomar al realizar el análisis de componentes principales. El primero es elegir qué variables incluir en el análisis. La opción deseable es que todas las variables describan un fenómeno común. En cuanto a ACP se desarrolló originalmente para la distribución normal multivariada y muestras de la misma, El ACP funcionará mejor en las variables que

son continuas y al menos aproximadamente normales.

2.1.2. Análisis de Componentes Principales Categorizado

Con frecuencia encontramos datos discretos en el análisis de observaciones de estudios educativos, sociales o de salud. A menudo, los datos discretos son binarios, es decir, una variable que solo puede tomar uno de dos valores, como el género (masculino / femenino) o la propiedad de una casa (sí / no). Otras veces, las variables discretas muestran varias categorías, como el tipo de asiento en un avión, la industria de una empresa o región geográfica.

Si hay varias categorías de una variable discreta, pueden tener un orden natural y la variable se denomina ordinal, porque las categorías se pueden enumerar utilizando una relación monótona entre ellas. Dos ejemplos podrían ser: el nivel de educación de una persona (primaria, secundaria, técnica, profesional o de grado avanzado) o la opinión sobre cierto tema debatido (totalmente de acuerdo, parcialmente de acuerdo, indeciso, algo en desacuerdo, completamente en desacuerdo). Quizás en el último ejemplo podríamos usar códigos como (2, 1, 0, -1, -2); pero observe que estas etiquetas numéricas son solo una elección arbitraria. Sin embargo, se debe tener cuidado para evitar el uso de códigos como una variable continua.

Otro tipo más de datos discretos son los recuentos de datos, como el número de tareas asignadas a los estudiantes en una semana o la cantidad de alumnos exitosos en un tema determinado.

Cuando los datos x_k discretos observados o las etiquetas se utilizan tal como vienen, analizarlos con el análisis de componentes principales estándar tiene al menos dos implicaciones.

En primer lugar, los datos discretos, especialmente aquellos que tienen pocas categorías, solo tienen una función de masa de probabilidad (sin densidad), por lo

tanto, se violan los supuestos de distribución que incluyen la normalidad. Además, incluso con su rango finito, los datos discretos pueden mostrar una gran asimetría y curtosis, especialmente si la mayoría de los puntos de datos se concentran en una sola categoría.

En segundo lugar y más importante, una consecuencia de la discreción es que las covarianzas o correlaciones calculadas entre las versiones discretizadas X_1^*, X_2^* de cualquiera de las dos variables de interés no reflejan las covarianzas "verdaderas" o las correlaciones de las variables subyacentes (no observadas o desconocidas) X_1, X_2. Pero en rigor, las covarianzas de versiones discretizadas no se pueden calcular, a menos que usemos etiquetas de categoría como si fueran valores de variables continuas.

Un desafío adicional es que el ordenamiento natural de categorías generalmente no se mantiene mediante el análisis del componente principal, por lo que la única forma de identificar dicho ordenamiento sería el uso de variables ordinales para las cuales los valores más altos realmente significan puntajes más altos en la variable de respuesta. Pero muchas veces, puede que no sepamos cuál es el "orden correcto" antes de las primeras analisis.

El análisis de componentes principales categóricos (CatPCA) es una técnica estadística multivariante utilizada para la reducción de datos, que no hace suposiciones de normalidad, en las cuales estudiamos p variables observables, $x_1, x_2, ..., x_p$, a través del cual generaremos otras k variables no observables, $k < p$.

El ACP estándar asume relaciones lineales entre las variables numéricas, mientras que el ACP categórico permite escalar las variables a diferentes niveles. Las variables categóricas se *cuantifican* de forma óptima en la dimensionalidad especificada. Como resultado, se pueden modelar relaciones no lineales entre variables.

La técnica estadística CatPCA se utiliza para la reducción de cualquier combinación de datos nominales, ordinales y numéricos, sin hacer suposiciones de normalidad. El CatPCA realiza la escala óptima de variables categóricas, asignando valores

numéricos apropiados a las diferentes respuestas. También reduce la dimensionalidad a q nuevas dimensiones, donde $p \geq q$, aplicando ACP estándar a todas las variables transformadas a escala numérica.

Para variables numéricas continuas, el proceso de escalado óptimo es como el caso tradicional. Supongamos que tenemos medidas de n individuos en m variables dadas con un $n \times m$ la matriz de puntuaciones observadas H donde cada variable se denota por $X_j, j = 1, ..., m$ que es la j-ésima columna de H. Si las variables X_j son de nivel de medición nominal u ordinal, entonces se requiere una transformación no lineal llamada escalamiento óptimo donde cada puntaje observado se transforma en una cuantificación de categoría dada por:

$$q_j = \varphi_j(X_j)$$

Donde Q es la matriz de cuantificaciones de categoría. Sea S la matriz $n \times p$ de puntuaciones de objeto, que son los puntajes de los individuos en los componentes principales, obtenidos por CatPCA. Los puntajes de objetos se multiplican por un conjunto de pesos óptimos que se denominan cargas de componentes. Sea A una matriz $m \times p$ de las cargas componentes donde la j-ésima columna se denota por a_j. A continuación, la función de pérdida para minimizar la diferencia entre los datos originales y los componentes principales se puede dar de la siguiente manera:

$$L(Q, A, S) = n^{-1} \sum_{j=1}^{m} tr(q_j a_j^T - S)^T (q_j a_j^T - S)^T \tag{2.2}$$

donde tr vendría a ser la función de seguimiento, es decir, para cualquier matriz A, $tr(A^T A) = \sum_i \sum_j a_{ij}^2$. En consecuencia, el CatPCA se realiza minimizando la función de pérdida de mínimos cuadrados dada en la ecuación en la que la matriz X se reemplaza por la matriz Q.

Al igual que con ACP estándar, que tiene un vector aleatorio p, obtenemos la matriz de varianzas y covarianzas a partir de la cual se extraen los componentes

principales. Estos componentes son combinaciones lineales que no están correlacionadas entre sí. Entonces, el primer componente se designa como la combinación lineal con la varianza máxima, el segundo componente es la combinación lineal con la segunda varianza máxima, y así sucesivamente, procediendo de la misma manera para encontrar los otros componentes.

Nuevamente, la cantidad de componentes principales elegidos dependerá del porcentaje de la varianza total que se necesita explicar. La técnica es más útil cuando un número extenso de variables impide una interpretación efectiva de las relaciones entre objetos (casos o unidades). Bajo una dimensionalidad reducida, se interpreta un pequeño número de componentes en lugar de un número extenso de variables [5]. En este trabajo utilizamos la implementación de IBM SPSS [6] de PCA categórico.

2.1.3. Referencia de sintaxis de comandos de IBM SPSS

```
CATPCA VARIABLES = varlist

 /ANALYSIS = varlist
    [[(WEIGHT={1**}] [LEVEL={SPORD**}] [DEGREE={2}] [INKNOT={2}]]
              {n   }                          {n}          {n}
                              {SPNOM   }  [DEGREE={2}] [INKNOT={2}]
                                                  {n}          {n}
                              {ORDI    }
                              {NOMI    }
                              {MNOM    }
                              {NUME    }

 [/DISCRETIZATION = [varlist [([{GROUPING    }] [{NCAT={7}   }] [DISTR={NORMAL }])]]]
                                                {n}              {UNIFORM}
                                 {EQINTV={n} }
                                 {RANKING    }
                                 {MULTIPLYING}

 [/MISSING = [varlist [([{PASSIVE**}] [{MODEIMPU}])]]]
```

```
                                          {EXTRACAT}
                              {ACTIVE     } {MODEIMPU}
                                          {EXTRACAT}
                                          {LISTWISE}

[/SUPPLEMENTARY = [OBJECT(varlist)] [VARIABLE(varlist)]]

[/CONFIGURATION = [{INITIAL}] (file)]
                   {FIXED   }

[/DIMENSION = {2**}]
              {n   }

[/NORMALIZATION = {VPRINCIPAL**}]
                  {OPRINCIPAL   }
                  {SYMMETRICAL  }
                  {INDEPENDENT  }
                  {n            }

[/PRINT = [DESCRIP**[(varlist)]] [VAF] [LOADING**][QUANT[(varlist)]][HISTORY]
          [CORR**] [OCORR] [OBJECT[([(varname)]varlist)]] [NONE]]

[/PLOT = [OBJECT**[(varlist)][(n)]]
         [LOADING**[(varlist [(CENTR[(varlist)])])][(n)]]
```

A continuacion se da una breve reseña de los comandos utilizados para el analisis CatPCA.

ANALYSIS: especifica las variables que se utilizarán en los cálculos, el nivel de escala óptimo y el peso variable para cada lista de variables o variables. ANÁLISIS también especifica variables suplementarias y su nivel óptimo de escala. No se puede especificar peso para las variables suplementarias.

DISCRETIZATION: especifica las variables de valores fraccionarios que desea discretizar. Además, puede usar DISCRETIZATION para clasificar o para dos formas de recodificar variables categóricas.

MISSING: En CATPCA, se considera un valor perdido del sistema, valores per-

didos definidos por el usuario y valores que son menores a 1 como valores perdidos. El subcomando MISSING permite indicar cómo manejar los valores perdidos para cada variable.

SUPPLEMENTARY: especifica los objetos y / o variables que se desea tratar como suplementarios. Las variables suplementarias se deben encontrar en el subcomando ANALYSIS. No puede ponderar objetos y variables adicionales (los pesos especificados se ignoran). Para variables suplementarias, todas las opciones en el subcomando MISSING se pueden especificar excepto LISTWISE.

CONFIGURATION: permite leer datos de un archivo que contiene las coordenadas de una configuración. La primera variable en este archivo debe contener las coordenadas para la primera dimensión, la segunda variable debe contener las coordenadas para la segunda dimensión, y así sucesivamente.

DIMENSION: especifica la cantidad de dimensiones (componentes) que desea que CatPCA calcule.

NORMALIZATION: especifica una de las cinco opciones para normalizar los puntajes de objeto y las variables. Solo se puede usar un método de normalización en un análisis dado.

PRINT: Muestra el resumen del modelo (alfa de Cronbach y Variación contabilizada) y las estadísticas de HISTORIA (la varianza contabilizada, la pérdida y el aumento en la varianza contabilizada) para la solución inicial (si corresponde) y la última iteración siempre se muestran.

Capítulo 3

La Propuesta

En este capítulo se describe las características de la población estudiantil objeto del estudio, asi como el contexto en el que fueron seleccionados los estudiantes para tomar parte de las pruebas y la encuesta. Igualmente se presentan los conceptos y las áreas que fueron abarcados en las pruebas de matemáticas.

Tambien se describe el proceso de la recopilación de los datos mediante el test de matemáticas, se describe el contenido de la encuesta, los pasos seguidos para la limpieza de los datos, la formación de grupos de variables y el procesamiento de los datos.

3.1. Población Estudiantil Analizada

Los estudiantes que fueron objeto de este estudio son 110 alumnos en el rango de edades de 11 a 17 años, de 26 instituciones educativas públicas y privadas, que forman parte del Programa "Iniciación Científica, con Énfasis en Matemáticas para Jóvenes Talentos" que es desarrollado por OMAPA [11] , buscando potenciar a los estudiantes del 3er. ciclo de la Educación Escolar Básica y la Educación Media que tienen aptitudes para las ciencias exactas. Es el primer y único programa con esas características en Paraguay con alcance nacional y demostrando su efectividad

a nivel internacional ya que numerosos alumnos capacitados en este programa se encuentran becados en renombradas universidades de diferentes países.

Solo estudiantes que hayan sido convocados pueden formar parte de "Jóvenes Talentos", por estar entre los estudiantes con los mejores puntajes de la Olimpiada Nacional de Matemáticas. Esta competencia se realiza cada año entre los alumnos de colegios públicos, subvencionados y privados de todo el país con una cantidad aproximada de 90.000 estudiantes.

Este Programa posibilita que jóvenes con aptitudes para las ciencias exactas encuentran un espacio óptimo para el máximo desarrollo de sus capacidades lógico-matemáticas. De esta manera se logra formar estudiantes interesados, desde temprana edad, en la investigación científica y en el estudio de las ciencias, por el carácter lúdico que presentan las matemáticas, su posibilidad de integración interdisciplinaria y su permeabilidad en todas las áreas de actuación del ser humano.

El programa también tiene como objetivo la igualdad de oportunidades y la inclusión social y académica de los estudiantes talentosos pertenecientes a sectores sociales desfavorecidos. Estudiantes de colegios de escasos recursos que se incorporados al programa ven mayores oportunidades de exito en el examenes de ingreso a las universidades, gracias a la preparación matemática que reciben, a las experiencias desafiantes que viven y el roce y compañerismo con sus pares. El programa se desarrolla en tres etapas principales Captación y Motivación a través de las Olimpiadas Nacionales de Matemáticas; Formación Académica y Científica; y Estímulo a la Excelencia.

3.2. El test de Matemáticas

Se administró una prueba de 15 ejercicios en matemáticas estableciendo el nivel de dificultad para cada ítem de acuerdo con la taxonomía de Bloom [3]. Cada ejercicio en la prueba fue pre-clasificada en una de las siguientes áreas:

- ***Aritmética:*** aplicando conceptos de operatoria en N, múltiplos de 3, cuadrados y cubos de un número, dígitos, cálculo de potencias, análisis de posibilidades, identificación de datos innecesario en un problema, problema de planteo de una división: criterio de divisibilidad por 3, concepto de divisibilidad y planteo de posibilidades dadas ciertas condiciones.

- ***Álgebra:*** con procedimientos de traducción de lenguaje coloquial a algebraico, concepto de triple, resolución de ecuaciones, secuencia, anterior, planteo de sumas y restas en naturales, respetando una regla de formación y definición de una operación con número naturales.

- ***Geometría:*** buscando la aplicación de conceptos de puntos colineales, identificación de rectas dados dos puntos, áreas de triángulos, círculos, ángulo central,propiedades de los ángulos interiores de un triángulo, identificación de la parte de superposición de círculo y triángulo como medio círculo.

- ***Lógica:*** con ejercicios que requieren el manejo de conceptos de planteo de posibilidades, organización de datos y valor de verdad.

Igualmente se presentaron ejercicios que combinaban 2 áreas, como las siguientes:

- ***Geometría / Álgebra:*** con conceptos de triángulos, ángulos, bisectriz, propiedad de suma de ángulos interiores de un triángulo, planteo de ecuaciones, traducción de lenguaje coloquial y geométrico a operaciones con naturales, cuerpos, caras, vértices de una cara, traducción de lenguajes coloquial y gráfico a operaciones con naturales.

- ***Estadística / Lógica:*** con conceptos de promedio, características de un promedio y valores de verdad.

- ***Aritmética / Geometría:*** con procedimientos en rectángulo, área, descomposición de un número en producto de otros dos.

La calificación global de matemáticas de cada alumno se calculó como la suma de todas las calificaciones individuales.

3.3. La Encuesta

Además, se le dio un segundo instrumento a los mismos estudiantes consistente en una encuesta, elaborada por Daniel Bogoya M.[14] y su equipo; para determinar las líneas de base del proyecto de Actualización en Educación Matemática, para docentes de 1° a 6 ° grado de la EEB; llevado a cabo por OMAPA.

Este fue un cuestionario en un diseño tipo encuesta, con 61 preguntas diseñadas y organizadas para valorar el contexto personal, hábitos y actitudes de los estudiantes considerados frente a diversas situaciones. Cada pregunta fue de elección única en una escala ordinal de Likert[4], que incluian preguntas con dos opciones, otras de cuatro opciones y otras de cinco opciones .

Estas preguntas se clasificaron según sus temas como:

- ***Seguimiento de los padres:*** Agrupaba preguntas respecto al interes y acompañamiento de los padres, con un total de 6 preguntas que incluia preguntas como "Hablo con mis padres sobre la escuela", *"Mi padres me preguntan lo que aprendo en la escuela", "Mis padres revisan mi tarea"*, entre otras. Este grupo también abarcaba preguntas sobre el nivel de educaciona escolar / académica de los padres.

- ***Compromiso de los docentes:*** Consiste en un grupo de 8 preguntas sobre la forma en que el alumno ve que el docente encara las clases y el proceso de enseñanza, entre las que están *"Mi profesor falta mucho a clase", "Mi profesor revisa mis tareas", ."El profesor explica muy bien"*.

- ***Compromiso del alumno:*** Con una cantidad de 8 preguntas, agrupa preguntas relativas a la actitud del alumno frente a las clases de matemáticas, algunas de ellas son *"Falto mucho a clase", "Tengo interes en lo que el profesor dice", "Los estudiantes participamos en clase"*, entre otras.

- ***Trabajo asignado para el hogar:*** Agrupa 3 preguntas relacionadas a la frecuencia de tareas para la casa, el tiempo que le toma al estudiante hacerlas correctamente y el tiempo que el estudiante dedica a estudiar la materia. Las preguntas son "*Cuántas veces tu profesor de matemáticas te deja tareas para la casa*", "*Qué tiempo por dia te toma hacer las tareas de matematica en la casaz* "*Estudio matemática cuando tengo tiempo*".

- ***Instalaciones caseras para el estudio:*** Grupo formado por cuatro preguntas referentes a la infraestructura y elementos que tiene en casa para el estudio. Las preguntas son *"Tenés habitación para ti solo"*, "*Tenés escritorio para estudiar*", "*Tenés instrumentos musicales*", "*Elementos para deportes*".

- ***Libros en Casa:*** Agrupa dos preguntas referentes a libros que tiene para leer, ellas son "*Tenés libros para leer en casa*", "*Cuánto libros hay en tu casa*".

- ***Uso de Computadoras:*** Grupo formado po cinco preguntas relacionadas al acceso a computadoras y conexión a internet del estudiante. Algunas de las preguntas son "*Tienes Computadora en casa*", "*Tienes conexión a Internet en casa*", "*Qué tan frecuente usas la computadora en casa*".

- ***Autoevaluación en matemáticas:*** Reune seis preguntas que tratan sobre el concepto que el estudiante sobre su rendimiento en la materia de matemática,

entre las preguntas estan *"La Matemática es dificil para mi"*, "*No soy muy bueno en matemáticas*", "*Mi profesor dice que soy bueno en matemáticas"*.

- ***Utilidad de las matemáticas:*** Grupo formado por cuatro preguntas que apuntan a averiguar lo que el estudiante piensa sobre la utilidad que las matemáticas tienen para su presente y futuro, entre las preguntas está "*Es útil para el trabajo que quiero*", "*Me servirá para entrar a la universidad*", "*Puede ayudarme en la vida diaria*".

- ***Preferencia por la escuela:*** Grupo que reune tres preguntas referentes a lo que siente el estudiante estando en la escuela. Las preguntas son "*Me siento seguro en la escuela*", "*Me gusta estar en la escuela* ", "*Prefiero la escuela que otro sitio*".

- ***Relación con compañeros de clase:*** Agrupa seis preguntas que buscan averiguar la relacionamiento del estudiante con sus compañeros, incluyendo preguntas sobre situciones de posibles acoso (bullying) de parte de los compañeros, entre las preguntas estan *"Mis compañeros no me dejan jugar con ellos", "Mis compañeros se burlan de mi o ponen apodos", "Soy obligado a hacer cosas que no quiero*".

Las respuestas de cada alumno a este instrumento se unieron a sus puntaje obtenido en la prueba de matemáticas, lo que permite el análisis de todo el conjunto de datos o porciones seleccionadas en estos.

3.4. Preprocesamiento de datos

Previo al análisis de componentes principales, se hizo un pre-procesamiento de los datos, para la identificación de valores nulos o sucios en las respuesta de la encuesta.

Se consideraron como datos sucios respuestas múltiples a una misma pregunta de la encuesta, donde solo corresponde una respuesta. Igualmente los en los casos de respuestas en blanco los valores faltantes fueron reemplazados por la moda (el valor con mayor frecuencia en una distribución de datos) de la variable correspondiente, justo antes de la ejecución. Una excepción a esta regla fue cuando faltaban todas las respuestas de un encuestado, en cuyo caso el registro se eliminó. A los datos sobre las escalas Likert ordinales se les asignaron valores numéricos crecientes que van desde las percepciones "malas" asignandole el valor 1 (uno) a las "buenas" asignandoles el valor 4 ó 5 (cuatro ó cinco) segun corresponda a la cantidad de respuestas; esto fue posible debido a la forma en que se presentaron todas las preguntas.

Los resultados de las pruebas de matemáticas, solo la variable numérica en el estudio correspondiente al puntaje global del examen, se transformaron en categorías del 1 al 7 el análisis y para una interpretación más sencilla de los resultados.

La escala de puntaje del exámen fue categorizada de la siguiente manera:

- Del 0 % al 20 % del examen = 1
- Del 21 % al 35 % del examen = 2
- Del 36 % al 50 % del examen = 3
- Del 51 % al 60 % del examen = 4
- Del 61 % al 75 % del examen = 5
- Del 76 % al 90 % del examen = 6
- Del 91 % al 100 % del examen = 7

También, para un análisis alternativo, se realizó una evalución por separado para cada tipo de ejercicio del examen, es decir se asigno una nota a cada estudiante

por área, como si se trataran de varios exámenes tomados al mismo estudiantes , calificandolo en álgebra con una nota, en artimética con otra nota, procediendo de la misma manera con todas las demás áreas ya citadas.

No existen valores perdidos en el instrumento de examen de matemáticas. La estandarización se considera una necesidad cuando no todas las variables se miden en las mismas unidades, esto se aplica a este estudio, por lo tanto, todos los datos se estandarizaron al inicio de cada ejecución.

Capítulo 4

Resultados

En el presente capítulo se presentan los resultados obtenidos a través de los analisis realizadas sobre los datos agrupados en matrices a partir del preprocesamiento hecho al filtrar datos nulos o inexistentes.

Como análisis inicial exploratorio se realizaron algunas corridas de PCA sobre las matrices formadas por todas las variables de la encuesta junto con cada uno de las calificaciones por área de cada alumno, formandose las siguientes matrices:

- Calificación en aritmética y la encuesta.
- Calificación en álgebra y la encuesta.
- Calificación en geometría y la encuesta.
- Calificación en lógica - estadística y la encuesta.

Sin embargo, la varianza total explicada por los primeros 2, e incluso 3, componentes principales permaneció por debajo del 50 % en todas las matrices, no consiguiendose el objetivo de reducir la dimensionalidad, debido a que la cantidad de variables de cada matriz era grande en relación a la cantidad de registros (estudiantes) de las tablas.

En una serie de nuevas corridas, solo se tomó uno de los grupos formados a partir de las preguntas de la encuesta junto a la calificación global de matemáticas de cada estudiante, formandose los siguientes grupos de matrices.

1. Seguimiento de los padres y calificación global del examen.
2. Compromiso de los docentes y calificación global del examen.
3. Compromiso del alumno y calificación global del examen.
4. Trabajo asignado para el hogar y calificación global del examen.
5. Instalaciones caseras para el estudio y calificación global del examen.
6. Libros en Casa y calificación global del examen.
7. Uso de Computadoras y calificación global del examen.
8. Autoevaluación en matemáticas y calificación global del examen.
9. Utilidad de las matemáticas y calificación global del examen.
10. Preferencia por la escuela y calificación global del examen.
11. Relación con compañeros de clase y calificación global del examen.

Con esta segunda serie de corridas se lograron mejorar notablemente los resultados en cuanto a porcentaje de varianza concentrada en las dos primeras dimesiones.

Como resultados preliminares la siguiente tabla muestra las matrices cuyos porcentajes de la varianza total explicada por las dos primeras dimensiones superan el 70 % , en esta serie de ejecuciones.

Tabla 4.1: Grupos de Variables y Porcentaje total de varianza explicada por dos dimensiones

Grupos de Variables	*Dimensión 1*	*Dimensión 2*	*Total*
Tareas Casa	50,985 %	37,437 %	88, 421 %
Libros en casa	45,622 %	33,577 %	79, 198 %
Motivación Matemáticas	46,696 %	25,676 %	72,372 %
Preferencia por la Escuela	40,269 %	29,553 %	69,823 %

A continuación los detalle de cada matriz, presentadas en orden descendente de acuerdo a la cantidad de varianza mostrada por cada una. El criterio de selección de utilizado para seleccionar las matrices fue que la varianza acumulada en las dos primeras dimesiones sea igual o mayor al 70 %.

4.1. Tareas para la Casa vs. Puntaje en el test

- Varianza acumulada en dos dimensiones: 88, 421 %.
- Cantidad de estudiantes analizados tras eliminacion de datos nulos: 110.

Los vectores propios; es decir, los coeficientes de las dos primeras dimensiones, tienen los siguientes valores:

Tabla 4.2: Saturaciones en Componentes - Tareas para la Casa

Variables	*Dimensión 1*	*Dimensión 2*
Frecuencia de tareas para la casa	-0,261	0,837
Tiempo que lleva para resolver tarea	0,024	0,881
Estudio Matemática cuando tengo tiempo	0,993	0,100
MAT	0,993	0,100

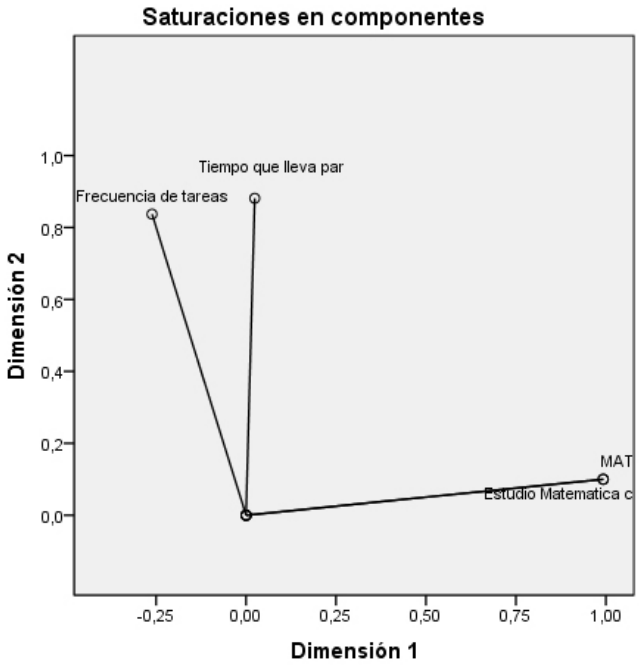

Figura 4.1: Grafico de Saturación de Componentes Tareas Casa vs. Puntaje en el test

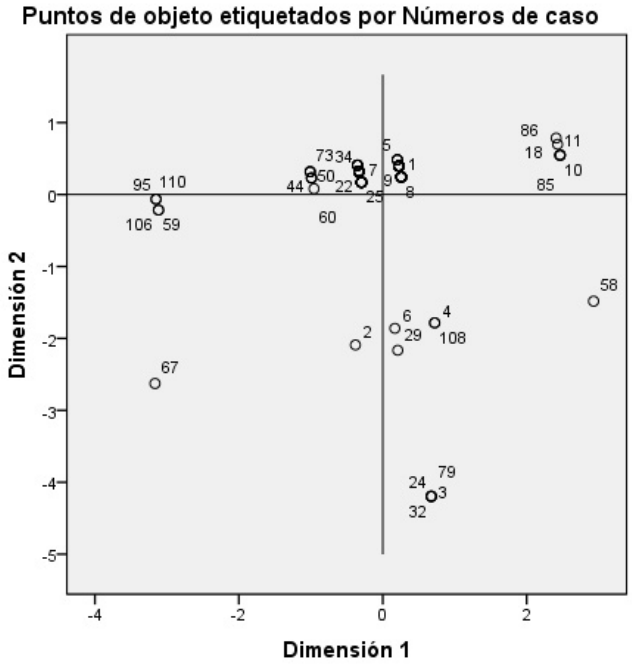

Figura 4.2: Puntos por objeto Tareas Casa vs.Puntaje en el test

Para la primera dimensión se observa mayor peso en las variables ***Estudio Matemática cuando tengo tiempo*** y ***MAT*** ambas con signo positivo y con el mismo peso, que muestra una relación directa del rendimiento en el test con la frecuencia de estudio en casa. Grandes valores en esta dimensión indican estudiantes que estudian con mucha frecuencia matemáticas en casa con rendimiento alto en el test.

Para la segunda dimensión las variables con más peso son ***Frecuencia de tareas para la casa*** y ***Tiempo que lleva para resolver tarea***, ambos con signos positivos, valores altos indican muchas tareas para la casa y rapidez para resolverlas.

La ortogonalidad de ambas dimensiones obsevadas en el gráfico de Saturación de Componentes sugiere que la cantidad de tarea asignada no está relacionada con el puntaje global del examen. Esto se verifica claramente al observar la simetría de la figura anterior con respecto al eje de la Dimensión 1, y también al comparar los dos grupos de estudiantes (10,11,18,83) y (58,92,103,107) ambos estan en el lado de *"mucho tiempo para matemáticas en casa"*; el primer grupo obtiene altas calificaciones de matemáticas, mientras que el segundo obtiene calificaciones bajas.

En el gráfico de puntos de objetos por número de casos, muestra la distribucion de los estudiantes respecto a las 2 dimensiones, en el que cada punto númerado representa a un estudiante. En el cuadrante de la esquina superior derecha tenemos alumnos que indicaron que se les asigna mucha tarea, que resuelven sus tareas rápidamente y obtienen altas calificaciones en el examen; la mayor concentración de puntos se encuentra allí. Por el contrario, en el cuadrante inferior derecho tenemos estudiantes que dicen tener muchas tareas, resolverlas rápidamente y que obtuvieron bajas calificaciones en la prueba; este cuadrante tiene la menor cantidad de puntos.

4.2. Libros en Casa vs. Puntaje en el test

- Varianza acumulada en dos dimensiones: 79, 198 %.
- Cantidad de estudiantes analizados tras eliminación de datos nulos: 109.

Tabla 4.3: Saturaciones en Componentes - Libros en Casa

Variables	*Dimensión 1*	*Dimensión 2*
Cuanto libros hay en tu casa	-0,830	-0,060
Tenés libros para leer en casa	0,821	0,156
MAT	-0,079	0,990

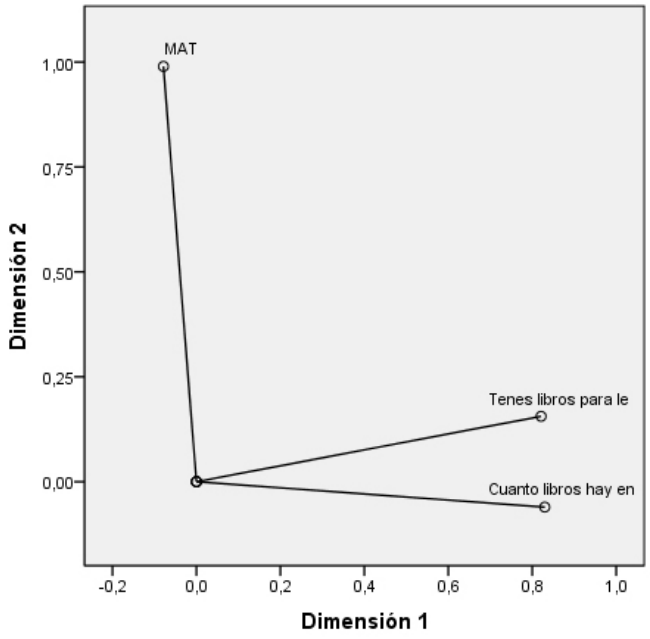

Figura 4.3: Gráfico de Saturación de Componentes Libros en Casa vs. Puntaje en el test

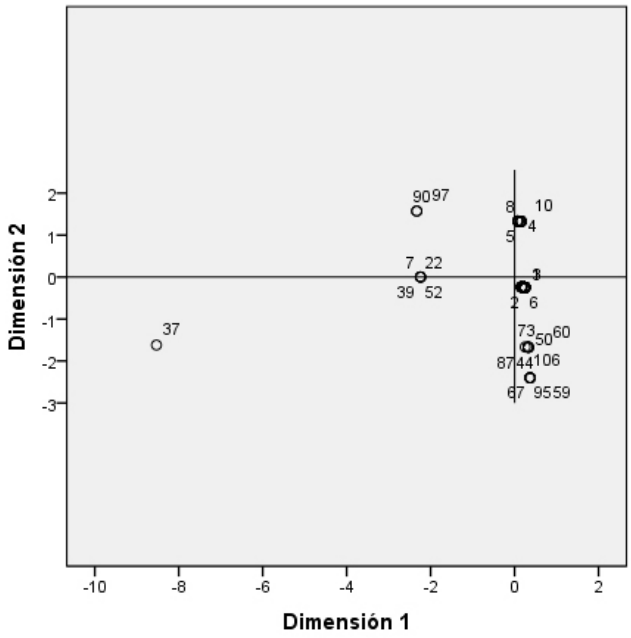

Figura 4.4: Puntos por objeto Libros en Casa vs. puntaje Puntaje en el test

Para la primera dimensión se ve mayor peso en las variables ***Cuánto libros hay en tu casa*** y ***Tenés libros para leer en casa***, ambas con signo positivo, que podríamos englobar como la *Tenencia de libros en la casa disponibles para la lectura*, lo que no implica que el estudiante lea o haya leído tales materiales de lectura. Grandes valores en esta dimensión indican alumnos con mucha cantidad de libros en casa.

En la segunda dimensión la variable con más peso es unicamente ***MAT***, es decir el puntaje del test, con signo positivo, valores altos en esta variable indican altas calificaciones.

Con estos datos podemos inferir que no hay relación entre la cantidad de libros que los estudiantes tengan y el rendimiento en el test de matemáticas, por encontrarse ambas variables en dimensiones separadas.

Observando el gráfico de puntos de objetos por número de casos se ve una mayor concentración de estudiantes en el 4º cuadrante que corresponden a estudiantes que

tienen gran cantidad de libros y rendimiento bajo en el test.

4.3. Motivación por las Matemáticas vs. Puntaje en el test

- Varianza acumulada en dos dimensiones: 72,372 %.
- Cantidad de estudiantes analizados tras eliminación de datos nulos: 109.

Tabla 4.4: Saturaciones en Componentes - Motivación por las Matemáticas

Variables	*Dimensión 1*	*Dimensión 2*
La materia es aburrida	0,285	0,812
No deseo estudiar matemática	0,141	0,805
Disfruto aprender matemáticas	0,941	0,004
Aprendo cosas interesantes	0,821	0,247
Es importante hacerlo bien en matemáticas	0,875	-0,321
Me gusta la matemática	0,911	-0,230
MAT	0,116	-0,522

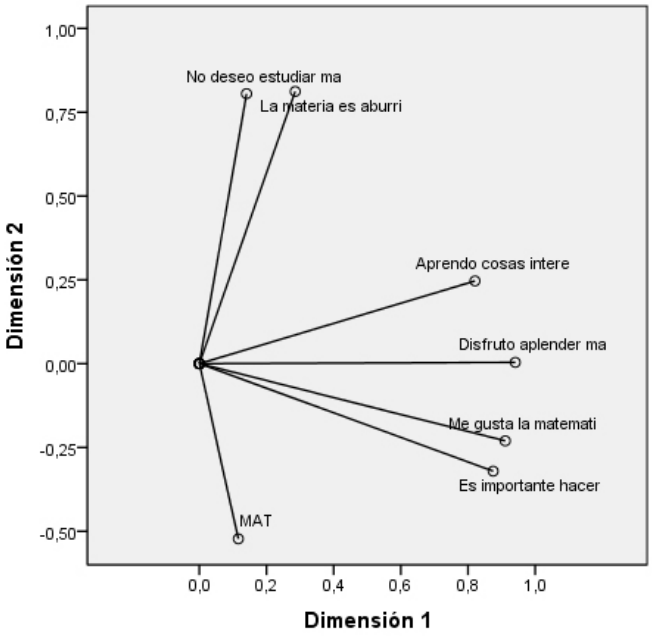

Figura 4.5: Gráfico de Saturación de Componentes Motivación por las Matemáticas vs. Puntaje en el test

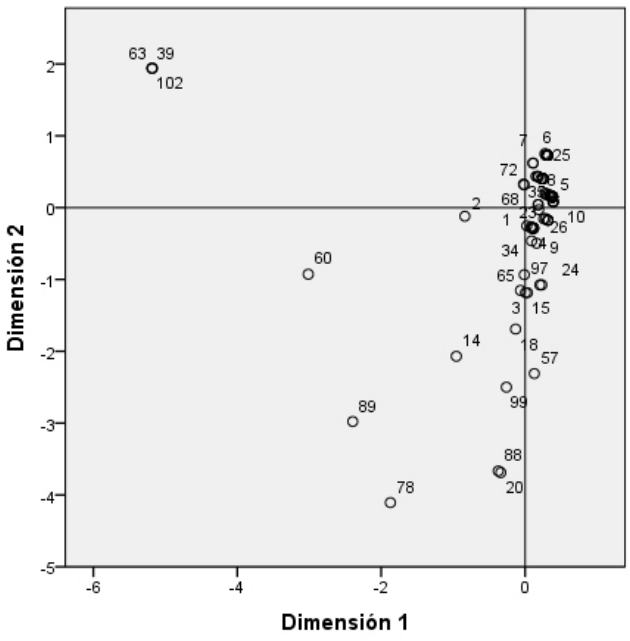

Figura 4.6: Puntos por objeto Motivación por las Matemáticas vs. puntaje Puntaje en el test

En la primera dimensión muestran mayor peso las variables ***Disfruto aprender matemáticas***, ***Me gusta la matemática***, y ***Es importante hacerlo bien en matemáticas*** todas con signo positivo, valores grandes en esta dimensión muestra afinidad por las matemáticas y revelan una motivación para la dimensión matemática.

Para la segunda dimensión es opuesta a la primera, para ingresar los de datos se invirtió la escala ordinal de respuestas, alineándolas con el resto de las preguntas de la encuesta, por lo tanto, las opiniones deben interpretarse como ***"La materia de matemáticas no es aburrida"*** y ***"Quiero estudiar matemáticas"***, estas opiniones aparecen con signo positivo y conforman la segunda dimensión junto con el puntaje global del examen que tiene signo positivo, observándose una relación inversa entre estas variables. Valores grandes de esta componente demuestran poco afinidad con las matemáticas con alta calificación en el examen.

El gráfico de puntos de objetos por número de casos indica una mayor concentración de estudiantes en el 1º cuadrante que corresponden a estudiantes con rendimiento bajo en el test, que expresaron su gusto por las matemáticas pero que están disconformes con la forma en que las estudian.

4.4. Preferencia por la Escuela vs. Puntaje en el test

- Varianza acumulada en 2 dimensiones: 69,823 %.
- Cantidad de estudiantes analizados tras eliminación de datos nulos: 109.

Como puede observarse, la varianza acumulada es ligeramente inferior, por unos puntos decimales, al 70 % designado como criterio de selección, pero por tratarse de

un grupo que reune variables muy interesante para el objetivo de este estudio, se decidió incluir entre los resultados.

Tabla 4.5: Saturaciones en Componentes - Preferencia por la Escuela

Variables	*Dimensión 1*	*Dimensión 2*
Me siento seguro en la escuela	0,359	0,714
Me gusta estar en la escuela	0,743	-0,498
Prefiero la escuela que otro sitio	0,855	-0,193
MAT	-0,448	-0,622

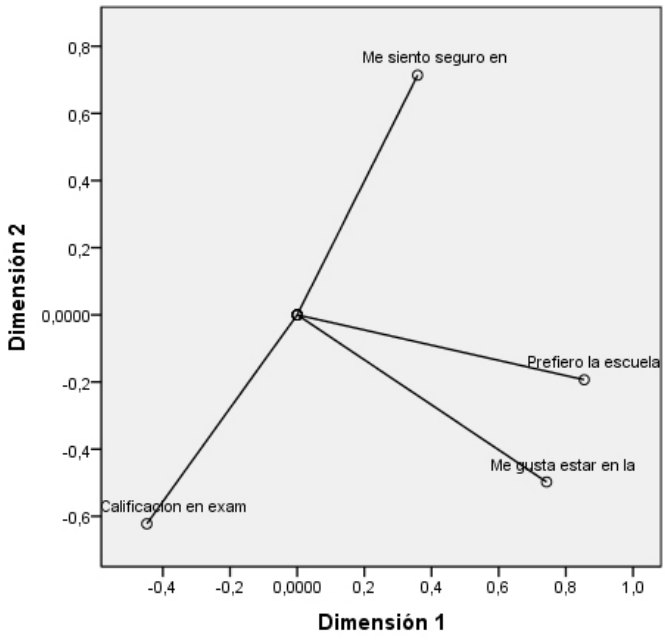

Figura 4.7: Gráfico de Saturación de Componentes Preferencia por la Escuela vs. Puntaje en el test

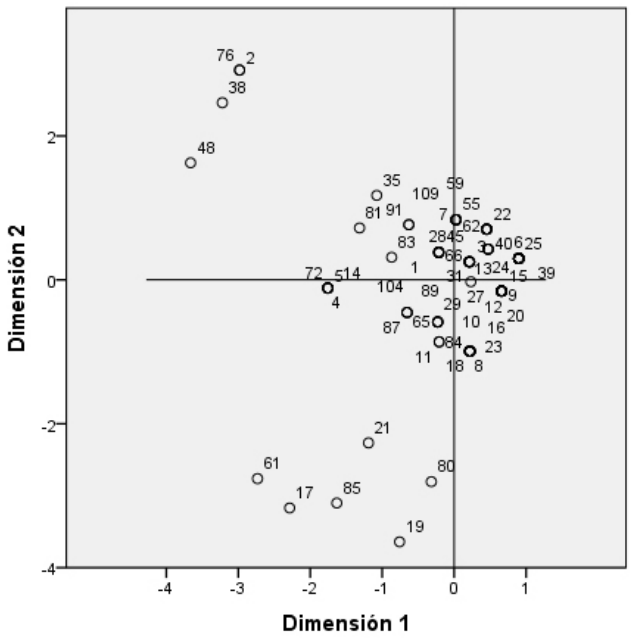

Figura 4.8: Puntos por objeto - Preferencia por la Escuela vs. puntaje Puntaje en el test

En la primera dimensión se ve que las variables con mayores pesos son ***Prefiero la escuela que otro sitio*** y ***Me gusta estar en la escuela*** ambas con signo positivo, lo que se podría consideraser como *Gusto por la escuela*. Valores altos en esta dimensión señalan gusto por la escuela.

En la segunda dimensión la variable con más peso es ***Me siento seguro en la escuela*** presentando signo positivo y la calificación del examen con signo negativo, lo que indica una relación inversa entre la sensación de seguridad en la escuela y el rendimiento en el examen. Valores altos en esta dimensión muestran sentimiento de mucha seguridad en la escuela y baja calificación.

En el gráfico de puntos de objetos por número de casos la mayoría de los puntos permanecen relativamente cerca del nuevo origen de coordenadas. Se ve una mayor concentración de estudiantes en el primer cuadrante que corresponden a estudiantes a quienes les gusta estar en la escuela, sintiéndose seguros en ella y bajo rendimiento

en el test. En el cuadrante superior izquierdo vemos a estudiantes a los que no les gusta la escuela pero que se sienten seguros allí y obtienen calificaciones relativamente más bajas; este es el cuadrante menos poblado.

La siguiente tabla muestra las matrices cuyos porcentajes de varianza total explicada por las dos primeras dimensiones fueron bajos, por lo que para alcanzar el 70 % de varianza se incluyó la tercera dimension en el análisis , por tratarse de matrices que concentran grupos de preguntas que podrian aportar información relevante para el objetivo de este trabajo.

Tabla 4.6: Grupos de Variables y Porcentaje total de varianza explicada por 3 dimensiones

Grupos de Variables	*Dimensión 1*	*Dimensión 2*	*Dimensión 3*	*Total*
Auto calificación del Estudiante	44,123 %	15,649 %	13,695 %	73,466 %
Uso de Computadora e Internet	30,614 %	25,473 %	16,023 %	72,109 %
Relación con los compañeros	35,405 %	19,556 %	16,385 %	71,346 %

Se muestran a continuación los detalle de estas matrices, presentadas en orden descendente de acuerdo a la cantidad de varianza mostrada por cada una.

4.5. Autocalificación del Estudiante vs. Puntaje en el test

- Varianza acumulada en tres dimensiones: 73,466 %.
- Cantidad de estudiantes analizados tras eliminación de datos nulos: 105.

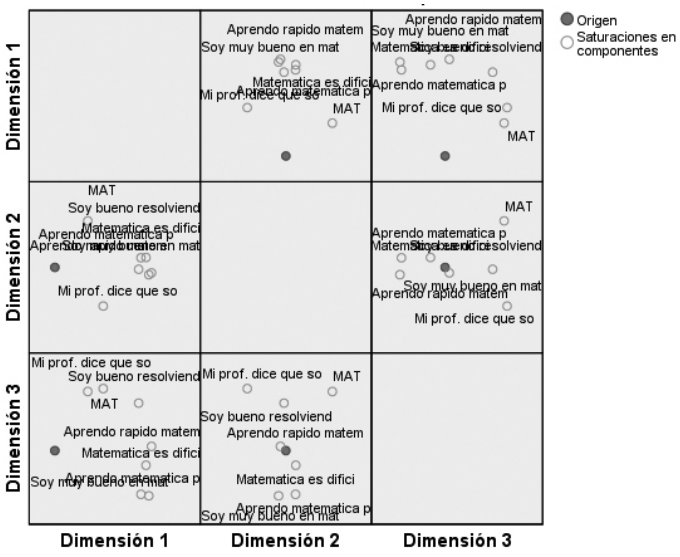

Figura 4.9: Gráfico de Saturación de Componentes Auto calificación del Estudiante vs. Puntaje en el test

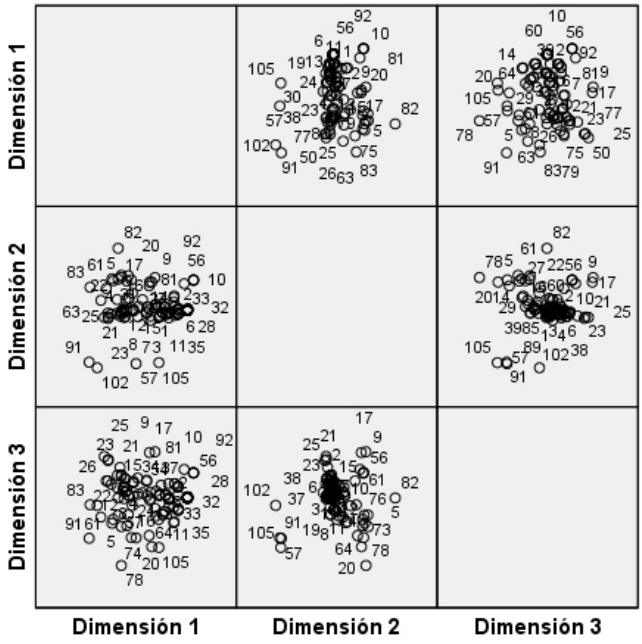

Figura 4.10: Puntos por objeto Auto calificación del Estudiante vs. puntaje Puntaje en el test

Tabla 4.7: Saturaciones en Componentes - Auto calificación del Estudiante

Variables	*Dimensión 1*	*Dimensión 2*	*Dimensión 3*
Matemática es difícil	0,761	0,164	-0,122
Soy muy bueno en matemáticas	0,784	-0,123	-0,377
Aprendo rápido matemática	0,807	-0,091	0,035
Aprendo matemática pero despacio	0,718	0,160	-0,364
Soy bueno resolviendo problemas difíciles	0,700	-0,032	0,396
Mi profesor dice que soy bueno en matemáticas	0,404	-0,645	0,517
MAT	-0,275	0,776	0,493

En columna de la primera dimensión se observa mayores pesos en las variables ***Matemática es difícil***, ***Soy muy bueno en matemáticas*** y ***Aprendo rápido matemática*** todas con signo positivo. Valores altos en esta dimensión corresponde a estudiantes que se consideran buenos en matemáticas y piensan que la matemáticas no son sencillas.

Para la segunda dimensión las variables mas grandes son ***MAT*** que es el puntaje en el test de matemáticas con signo positivo y la variables ***Mi profesor dice que soy bueno*** con signo negativo, esto indica que existe una relación inversa entre lo que el profesor opina del alumno y el rendimiento en el test. Valores altos en esta dimensión indican altos puntajes en el examen de matemáticas y una opinión negativa del profesor respecto al desempeño del alumno en la materia de matemáticas.

Para la tercera dimensión la variable con más peso nuevamente es ***Mi profesor dice que soy bueno***. Valores altos en esta dimensión significa una opinión positiva del profesor.

La mayor concentración de estudiantes se encuentra en el espacio en el que se

encuentran estudiantes que se consideran buenos en matemáticas a pesar de que les es difícil la materia, que presentaron un bajo rendimiento en el test y cuyos profesores los consideran buenos en matemáticas. Por el otro lado la menor cantidad de estudiantes se ubican en el espacio que corresponde a los alumnos que piensan que la materia no es difícil pero no se considera bueno en ella, con bajo rendimiento en el test y cuyos profesores los consideran buenos en matemáticas.

4.6. Uso de Computadora e Internet vs. Puntaje en el test

- Varianza acumulada en tres dimensiones: 72,109 %.
- Cantidad de estudiantes analizados tras eliminación de datos nulos: 109.

Tabla 4.8: Saturaciones en Componentes - Uso de Computadora e Internet

Variables	*Dimensión 1*	*Dimensión 2*	*Dimensión 3*
Computadora en casa	0,407	0,707	-0,233
Internet en casa	0,093	0,595	0,670
Frecuencia de uso de computadora en casa	0,515	0,578	-0,399
Frecuencia de uso de computadora en la escuela	0,751	-0,314	-0,097
Frecuencia de uso de computadora en otro sitio	0,609	-0,453	-0,258
MAT	-0,680	-0,189	0,473

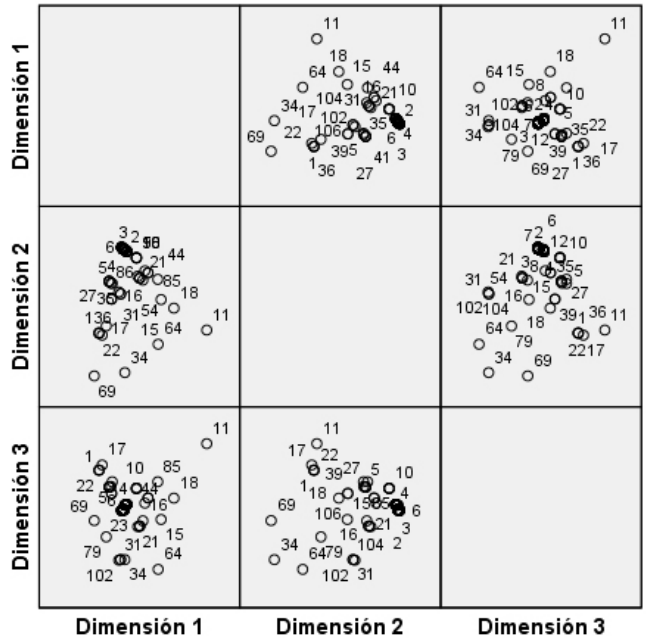

Figura 4.11: Gráfico de Saturación de Componentes Uso de Computadora e Internet vs. Puntaje en el test

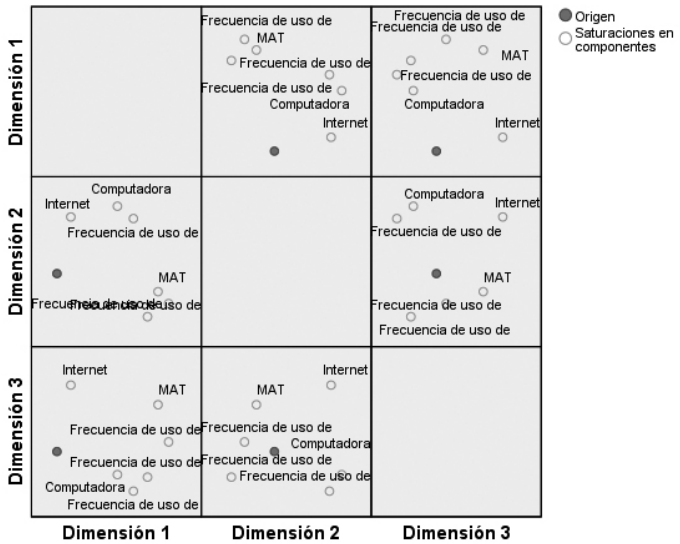

Figura 4.12: Puntos por objeto Uso de Computadora e Internet vs. puntaje Puntaje en el test

En la 1º dimensión las variables con mayor peso son ***Frecuencia de uso de computadora en la escuela***, ***Frecuencia de uso de computadora en otro sitio*** y ***MAT*** todas con signo positivo, con lo que se podría considerar que existe una relación directa entre el uso de computadoras fuera de la casa y el rendimiento en el test .Valores altos en esta dimensión señalan uso frecuente de la computadora fuera de la casa y buen desempeño en el test.

En la columna que corresponde a la segunda dimensión la variable con más peso es ***Tengo Computadora en casa***, ***Tengo Internet en casa*** y ***Frecuencia de uso de computadora en casa***. Valores altos en esta dimensión muestran uso frecuente de la computadora en casa.

Para la tercera dimensión la variable con más peso es ***Tengo Internet en casa nuevamente***. Valores altos en esta dimensión significa que el estudiante dispone de conexión a internet en casa.

En el gráfico de puntos de objetos por número de casos se ve una mayor concentración de estudiantes muestra una amplia mayoria de estudiantes que tienen computadora en casa con conexion a internet, que usan muy poco computadora fuera de la casa, con bajo desempeño en el test.

4.7. Relación con los compañeros vs. Puntaje en el test

- Varianza acumulada en tres dimensiones: 71,346 %.
- Cantidad de estudiantes analizados tras eliminación de datos nulos: 109.

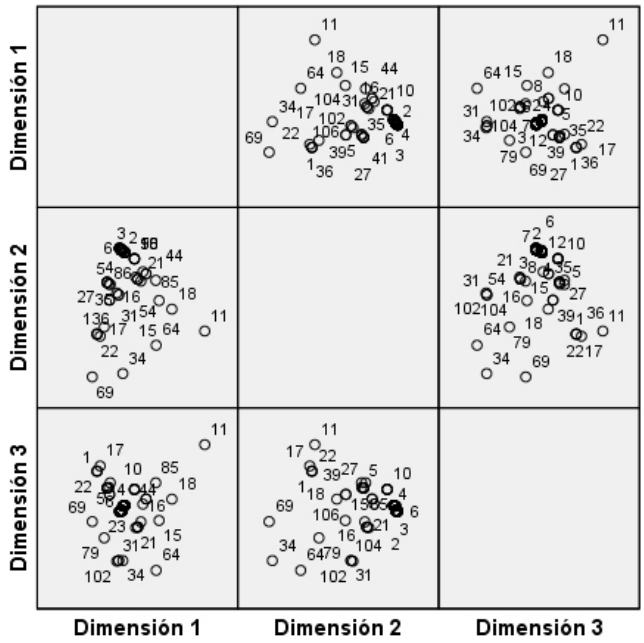

Figura 4.13: Gráfico de Saturación de Componentes Relación con los compañeros vs. puntaje en el test

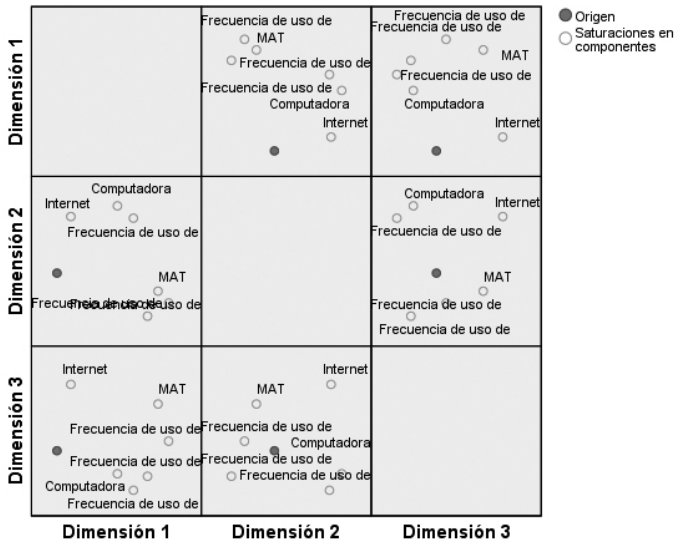

Figura 4.14: Puntos por objeto Relación con los compañeros vs. puntaje en el test

Tabla 4.9: Saturaciones en Componentes - Relación con los compañeros

Variables	*Dimensión 1*	*Dimensión 2*	*Dimensión 3*
No me dejan jugar con ellos	0,876	-0,172	-0,158
Se burlan de mí o ponen apodos	0,805	0,021	-0,106
Dicen mentiras de mí	0,681	-0,179	0,529
Soy obligado a hacer cosas que no quiero	0,123	-0,547	-0,687
Soy golpeado por compañeros	0,660	-0,116	0,025
Se pierden cosas mías o de mis compañeros	0,385	0,670	0,025
MAT	-0,011	-0,783	0,491

Para la primera dimensión se observa mayor peso en las variables ***No me dejan jugar con ellos***, ***Se burlan de mi o ponen apodos*** y ***Dicen mentiras*** de mi todas con signo positivo. Valores altos en esta dimensión indican malas relaciones con los compañeros en el colegio.

En la segunda dimensión las variables con más peso son ***MAT***, con signo negativo y ***Se pierden cosas mis o de mis compañeros*** con signo positivo, lo que indica una relación inversa entre el rendimiento en la prueba y la frecuente pérdida de objetos personales en el colegio. Valores altos en esta dimensión muestran bajo rendimiento en el test de matemáticas y alta frecuencia de perdida de objetos personales en el colegio.

Para la tercera dimensión la variable con más peso es ***Soy obligado a hacer cosas que no quiero***. Valores altos en esta dimensión indican que el alumno realiza actividades contra su voluntad con frecuencia.

En el gráfico de puntos de objetos por número de casos indica una ligera mayor concentración de estudiantes que señalan tener malas relaciones con los compañeros, buen rendimiento en el test de matemáticas, poca frecuencia en la perdida de objetos y que hacen actividades contra su voluntad. La segunda mayor concentración es la de estudiantes que indican malas relaciones con los compañeros en el colegio, que

realizan actividades contra su voluntad con frecuencia pero con buen rendimiento en el examen de matemáticas.

4.8. Identificación de Variables con Influencia en el Rendimiento en el Test

Del conjunto de todas las variables analizadas, fueron identificadas aquellas que muestran una relacion directa, lo que indica que crece junto al rendimiento del examen, es decir que al aumentar el valor de la variable también aumenta el rendimiento en el test

Tabla 4.10: Variables con Relación Directa con Calificación en el test

Variable	*Dimensión*	*Grupo de Variables*
Estudio Matemática cuando tengo tiempo	1º	Tareas para la Casa
Frecuencia de uso de computadora en la escuela	1º	Uso de Computadora e Internet
Frecuencia de uso de computadora en otro sitio	1º	Uso de Computadora e Internet

De igual modo fueron identificadas aquellas que muestran una relación inversa o negativa, lo que significa que al crecer una variable la otra disminuye, en otras palabras, al aumentar el valor de la variable, disminuye el rendimiento en el test.

4.9. Perfíl de estudiantes con los mejores puntaje

Como una información complementaria, en la tabla siguiente se muestra un perfíl que caracteriza a los 30 estudiantes que obtuvieron un puntaje global superior al 60 % .

Tabla 4.11: Variables con Relación Inversa con la Calificación en el test

Variable	*Dimensión*	*Grupo de Variables*
No deseo estudiar matemática	2º	Tareas para la Casa
La materia es aburrida	2º	Motivación Matemáticas
Me siento seguro en la escuela	2º	Preferencia por la Escuela
Mi profesor dice que soy bueno en matemáticas	2º	Auto calificación
Se pierden cosas mis o de mis compañeros	2º	Relación con los compañeros

Para cada pregunta expresada en la primera columna, en la segunda columna se registra la respuesta más frecuente (el modo estadístico) elegida por ese subconjunto de los estudiantes.

Tabla 4.12: Perfíl de estudiantes con los mayores puntajes
Grupo Tareas para la casa

Pregunta	*Respuesta*
Frecuencia de tareas para la casa	1 vez por semana
Tiempo para resolver la tarea	1 a 15 minutos
Estudio Matemática cuando tengo tiempo	1 o 2 Veces por semana

Tabla 4.13: Perfíl de estudiantes con los mayores puntajes
Grupo Libros en Casa

Pregunta	*Respuesta*
Tenés libros para leer en casa	Si
Cuánto libros hay en tu casa	Más de 100

Tabla 4.14: Perfíl de estudiantes con los mayores puntajes
Grupo Motivación por las matemáticas

Pregunta	*Respuesta*
La materia es aburrida	Muy en desacuerdo
No deseo estudiar matemática	Muy en desacuerdo
Disfruto aprender matemáticas	Muy de acuerdo
Aprendo cosas interesantes	Muy de acuerdo
Es importante hacerlo bien en matemáticas	Muy de acuerdo
Me gusta la matemática	Muy de acuerdo

Tabla 4.15: Perfíl de estudiantes con los mayores puntajes
Grupo Preferencia por la Escuela

Pregunta	*Respuesta*
Me siento seguro en la escuela	Muy de acuerdo
Me gusta estar en la escuela	Muy de acuerdo
Prefiero la escuela que otro sitio	Parcialmente de acuerdo

Tabla 4.16: Perfíl de estudiantes con los mayores puntajes
Autocalificación del alumno

Pregunta	*Respuesta*
Matemática es difícil	Muy en desacuerdo
No soy muy bueno en matemáticas	Muy en desacuerdo
Aprendo rápido matemáticas	Muy de acuerdo
Aprendo matemáticas pero despacio	Poco en desacuerdo
Soy bueno resolviendo problemas difíciles	Poco en desacuerdo
Mi profesor dice que soy bueno en matemáticas	Muy de acuerdo

Tabla 4.17: Perfíl de estudiantes con los mayores puntajes
Uso de Computadoras e Internet

Pregunta	*Respuesta*
Frecuencia de Uso de computadora en casa	Siempre
Frecuencia de Uso de computadora en la escuela	1 o 2 veces por semana
Frecuencia de Uso de computadora en otro sitio	Nunca o casi nunca
Computadora en casa	Si
Conexión a internet en casa	Si

Tabla 4.18: Perfíl de estudiantes con los mayores puntajes
Relación con mis Compañeros

Pregunta	*Respuesta*
Mis compañeros no me dejan jugar con ellos	Nunca
Mis compañeros se burlan de mí o me ponen apodos	Nunca
Mis compañeros dicen mentiras sobre mí	Nunca
Soy obligado a hacer cosas que no quiero	Nunca
Soy golpeado o lastimado por otros compañeros	Nunca
Se pierden cosas mías o de mis compañeros	1 o 2 veces por mes

4.10. Perfil de la mayoría de los estudiantes

En lo que se refiere al perfil presentado por la mayoria de los alumnos estudiados, el rendimiento no fue bueno, la mayoria presentouna rendimiento menor al 50 % de la prueba, que indicando que se les asigna mucha tarea, que las resuelvenrápidamente, dicen tener en casa una gran cantidad de libros, computadora con conexion a internet y que utilizan muy poco computadoras fuera de casa.

También la mayoria expresó su gusto por las matemáticas pero con disconformidad con la forma en que las estudian, les gusta estar en la escuela y se sienten seguros en ella. Se consideran buenos en matemáticas a pesar de que les es difícil la materia y cuyos profesores los consideran buenos en matemáticas. Ademas indicaron tener no buenas relaciones con los compañeros, que en ocasiones hacen actividades contra su voluntad y poca frecuencia en la pérdida de objetos personales.

4.11. La Variable Edad

Adicionalmente, con la intensión de extraer la máxima información a las distintas matrices formadas por los grupos de preguntas de la encuesta y las calificaciones del examen, se agregó una variable mas, que es la edad de los estudiantes, quedando las

matrices formadas por variables de la encuesta, calificación del examen y la edad. Se procedió a hacer los análisis con la misma metodología con la que se procesaron los primeros grupos de matrices.

Los resultados obtenidos mostraron lo que se podría suponer de antemano, en general los estudiantes con mayor edad presentaron rendimiento ligeramente superior a los estudiantes de menor edad, lo que se debería a que están más familiarizados con los conceptos presentados en los ejercicios por estar hace más tiempo resolviendo esos tipos de problemas matemáticos. Por otra parte no se observaron mucha variación en cuanto a la carga de varianzas, e inclusive en algunas matrices se observó una disminución en la cantidad de informacion concentrada en las primeras 2 componentes principales.

Capítulo 5

Conclusiones y Trabajos Futuros

Este estudio de caso permitió descubrir tendencias relacionadas con el rendimiento estudiantil en las pruebas de matemáticas, utilizando como herramienta el análisis de componentes principales categóricos (CatPCA). Con esta técnica se pudo representar en pocas dimensiones una realidad multidimensional al identificar las componentes principales, que podriamos denominar como variables de variables; variables que son combinación de las variables originales. Estas combinaciones son interesantes en sí mismas, porque ayudan a formar un conglomerado de variables combinadas de una forma que, en realidad, reflejan la vida interna y la relación que tienen ellas entre sí en cuanto a la covariación conjunta.

El estudio se basó en una encuesta de aspectos de estudio personal y familiar de la vida de los estudiantes y en los resultados de la prueba OMAPA que analiza los juegos olímpicos de matemáticas, teniendo en cuenta tanto el grado de examen global como la influencia que la escuela y el entorno humano pueden tener. También ilustramos el uso de CatPCA para la exploración de datos en presencia de algún conocimiento previo: dividimos los datos en varias ejecuciones, agrupando las preguntas relacionadas de la encuesta junto con los resultados matemáticos, y la herramienta ayudó a identificar qué agrupaciones tuvieron mayores contribuciones al total diferencia. A continuación, se presentan las conclusiones sobre los principales resultados.

Tarea asignada: observamos que la asignación frecuente de tarea y la velocidad para resolver tareas no están asociadas a las altas calificaciones; pero [redacted] tiempo total dedicado a estudiar matemática si está ligado al buen rendimiento [redacted] En este caso de estudio, existe una gran dispersión de resultados matemáticos bajos y altos en estudiantes con poca tarea de la escuela, así como en estudiantes con muchas asignaciones.

La motivación y simpatía de los estudiantes por las matemáticas escolares: en nuestra interpretación, esta dimensión expresa entusiasmo o emoción por las matemáticas, lo que demuestra que los estudiantes pueden sentirse atraídos por la materia, pero que no necesariamente muestran buen rendimiento en pruebas de matemáticas, lo que lleva a pensar que la simpatía por la materia está débilmente relacionada con el buen desempeño en una prueba de matemáticas, tal vez en contra de las creencias generales. Las expectativas relacionadas con las asignaturas de matemáticas se manifestaron como un componente separado de la motivación, es decir, muchos estudiantes podrían estar motivados hacia las matemáticas, pero al mismo tiempo pueden tener poco deseo de estudiarlo en la forma en que la escuela presenta las matemáticas. Estos aspectos deberán ser mejor aclarados en estudios posteriores.

Preferencia y seguridad: la mayoría de los estudiantes prefieren o les gusta su escuela, junto con la protección que la institución les brinda. Sin embargo, aquellos que dicen sentirse seguros, tienden a lograr un rendimiento inferior en matemáticas como un talento joven.

En otros temas, como el seguimiento de los padres o las instalaciones físicas para el alumno, que a menudo se consideran fundamentales, no se han encontrado ninguna dimensión con un alto porcentaje de varianza explicada. Sin embargo, esos factores

pueden ser importantes para el éxito en áreas de estudio fuera de las matemáticas, que no fueron consideradas aquí.

Uso de Computadora e Internet: Se muestra que el uso de computadoras fuera de casa influye en el rendimiento del estudiante en la materia, lo que se podria interpretar que el uso de computadoras en ambientes controlados y en forma colaborativa tiene efecto sobre su desempeño, dado que en la actualidad los padres están mucho tiempo fuera de casa se podría suponer que el uso de la computadora en casa es menos supervisado que en la escuela.

Relación con los compañeros: Este grupo de variables muestra que no hay una relación directa entre las malas relaciones con los compañeros de escuela y el rendimiento en el test de matemáticas, pero si sugiere que existe una leve influencia de la perdida de objetos personales en el desempeño en la prueba. Este grupo de preguntas amerita un estudio más minucioso y extenso por tratarse de un tema muy delicado y sensible como es el acoso escolar (bullying).

En el perfíl presentado de los mejores 30 estudiantes y sus respuestas más frecuentes a estos grupos de preguntas de la encuesta, muestra opiniones y características que generalmente se esperan de los estudiantes con buen rendimiento. Teniendo como punto resaltante el que se refiere a la poca cantidad de tareas (1 por semana) para la casa que los estudiantes dicen tener y al poco tiempo semanal (1 o 2 veces por semana) que le dedican al estudio de matemáticas

Estos análisis son todos exploratorios, pero sugieren ir a más estudios sobre la motivación, la cantidad de tarea asignada y la preferencia por la escuela, mediante el empleo de métodos más orientados a detectar las relaciones de causa y efecto. Probablemente, la disección de estos aspectos de la vida escolar permitirá encontrar

características de los estudiantes con alto rendimiento en matemáticas y también características de aquellos que no alcanzan ese nivel. La identificación de las variables involucradas en el rendimiento matemático de los alumnos mediante la minería de datos puede ayudar a los educadores a dirigir los cambios curriculares y evaluar los efectos de los cambios implementados, y posiblemente, a mejorar la generalización de las soluciones.

Finalmente, y no por eso menos importante, independientemente de las técnicas de análisis de datos utilizada y lo optimizada que puedan ser, es primordial la correcta elaboración de los items del instrumento utilizado para la recolección de datos, ya sea encuestas, test o entrevistas, es necesario evitar la ambigüedad en los items, formulando preguntas que sean claras para los encuestados, esto es fundamental para conseguir resultados válidos y exactos, que reflejen la realidad del caso estudiado y ayuden a tomar desiciones correctas sobre tales temas [**1**].

Bibliografía

[1] ElGamal, A.F. *An Educational Data Mining Model for Predicting Student Performance in Programming Course*. Department of CS Mansoura University, Egypt, 2013.

[2] La Red Martínez, David; Karanik, Marcelo; Giovannini, Mirtha; Báez, María Eugenia; Torre,Juliana,. *Descubrimiento de perfiles de rendimiento estudiantil*. Universidad Tecnologica Nacional.Argentina.2016

[3] S. Bloom, Benjamin. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. McKay.1956

[4] Rensis Likert. *A Technique for the Measurement of Attitudes*. Archives of Psychology.1932

[5] Alvarez Figueroa, Lorena. *Análisis Estadístico Multivariado Impacto Emigracion Familiares Estudiantes Adolescentes*. Escuela Superior Politécnica del Litoral, Ecuador.2005.

[6] IBM Knowledge Center *Categorical Principal Components Analysis (CATPCA)*. https://www.ibm.com/support/knowledgecenter/en/.

[7] OECD, Organisation for Economic Cooperationand Development. *Programme for International Student Assessment (PISA)*. http://www.oecd.org/pisa/pisaenespaol.htm, Francia.2015

[8] UNESCO, Oficina Regional de Educación para América Latina y el Caribe. *Tercer Estudio Regional Comparativo y Explicativo: Factores Asociados*. Francia.2015

[9] PSU, Pennsilvania State University Eberly College of Science. *Stat 505 Applied Multivariate Statistical Analysis.* https://onlinecourses.science.psu.edu/stat505/node/49.2017

[10] USDE, United States Department of Education. *Science, Technology, Engineering and Math: Education for Global Leadership*. https://www.ed.gov/stem.2015

[11] OMAPA, Organización Multidisciplinaria de Apoyo a Profesores y Alumnos. *Olimpiadas Nacionales de Matemática*. http://www.omapa.org/.

[12] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyths, Padhraic. *From Data Mining to Knowledge Discovery in Databases*. Asociación Estadounidense de Inteligencia Artificial.1996.

[13] Han, Sun Young; Capraro, Robert M.; Capraro, Mary Margaret,c. *How science, technology, engineering and mathematics STEM project-based learning affects high, middle and low achievers differently: The impact of student factors of achievement. In: International Journal of Science and Mathematics Education 2014*. https://www.researchgate.net/publication/271658486.l.2014.

[14] Daniel Bogoya Maldonado. *Hoja de Vida. http://scienti.colciencias.gov.co:8081/cvlac/visualizador/generarCurriculoCv.do ?cod_rh=0000257591*