ANALISIS ESTADISTICO MULTIVARIADO DE LOS RESULTADOS EN LAS PRUEBAS SABER PRO DEL PROGRAMA DE INGENIERÍA EN PRODUCCIÓN ACUÍCOLA EN LA UNIVERSIDAD DE NARIÑO 2016 – 2019.

LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

PROYECTO DE GRADO PARA OPTAR POR EL TÍTULO DE ESPECIALISTA EN ESTADISTICA APLICADA

FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES

PROYECTO DE GRADO

SAN JUAN DE PASTO

2020

ANALISIS ESTADISTICO MULTIVARIADO DE LOS RESULTADOS EN LAS PRUEBAS SABER PRO DEL PROGRAMA DE INGENIERÍA EN PRODUCCIÓN ACUÍCOLA EN LA UNIVERSIDAD DE NARIÑO 2016 – 2019.

PROYECTO DE GRADO PARA OPTAR POR EL TÍTULO DE ESPECIALISTA EN ESTADISTICA APLICADA

SANDRA MILENA CERÓN BENAVIDES

DIRECTORES: MANUEL FRANCISCO ROMERO OSPINA

LIDA RUBIELA FONSECA GÓMEZ

FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES

PROYECTO DE GRADO

SAN JUAN DE PASTO

2020

CONTENIDO

RESUMEN

El propósito de esta investigación fue definir las principales variables y la incidencia de ellas sobre el puntaje de las pruebas SABER PRO dentro del programa de Ingeniería en Producción Acuícola de la Universidad de Nariño en los últimos cuatro años 2016-2019.

Para el desarrollo de este proyecto se utilizaron 384 datos correspondientes a estudiantes de noveno semestre que presentaron las pruebas de estado SABER PRO en el programa de pregrado, se implementó un análisis descriptivo de los datos y técnicas multivariadas por componentes principales, Análisis de Regresión múltiple y cuantílica para establecer las principales variables y su incidencia en el puntaje global.

Los resultados arrojados por el primer análisis determino que todas las competencias genéricas inciden sobre las dos primeras componentes, pero existe la necesidad de fortalecer variables como Lectura crítica, Comunicación escrita y Competencia ciudadanas las cuales tienen contribución mayor a 50 sobre las tres últimas componentes, impidiendo que la variabilidad retenida en las primeras componentes sea mayor al 60%.

Por otro lado, haciendo caso omiso al incumpliendo del supuesto de homocedasticidad el Análisis de regresión múltiple generó un modelo con el 99% de capacidad predictiva, y el análisis de regresión cuantílica genero 5 modelos correspondientes a los cuantiles de 0.1, 0.25, 0.5(mediana), 0.75 y 0.9, sin diferencias significativas entre ellos. Además, todos los modelos sin excepción, demandan la participación de todas las competencias genéricas con el mismo coeficiente aproximado para cada una de ellas.

Como conclusión principal tenemos que todas las variables son indispensables para predecir el puntaje global, cada una con diferente grado de contribución, calidad de asociación y peso sobre

las combinaciones lineales de las dos primeras componentes. Además, el modelo generado por las diferentes técnicas de regresión fue el mismo, siendo cualquiera capaz de predecir el puntaje global de las pruebas SABER PRO para el programa de Ingeniería en Producción Acuícola.

Palabras claves: Análisis de Regresión múltiple, Análisis de Regresión cuantílica, Análisis de Componentes Principales, SABER PRO.

1. INTRODUCCIÓN

Los indicadores académicos revelan aquellas problemáticas que enfrenta la Educación Superior en nuestro país, entre ellos podemos hablar del nivel de deserción, el número de estudiantes no matriculados pero inscritos y admitidos, el nivel de desempeño con el que ingresan o egresan de su estudio de pregrado mediante el puntaje SABER 11° y SABER PRO, entre otros, que proporcionan un acercamiento al estado real de un programa, Institución o educación a nivel nacional.

Dependiendo del grado de influencia de estos indicadores se puede establecer la capacidad por parte de la institución, para solventar las principales problemáticas con el transcurrir del tiempo, lo anterior según la estructura del programa y/o facultad, medido por semestre, periodo o cohorte académico.

El presente trabajo tiene como objetivo establecer un análisis multivariado mediante el uso de diferentes técnicas con los datos generados en las pruebas SABER PRO dentro de los últimos cuatro años, lo anterior con estudiantes del programa de Ingeniería en Producción Acuícola del Departamento de Recursos Hidrobiológicos perteneciente a la Universidad de Nariño sede principal en el municipio de Pasto-Colombia.

2. PLANTEAMIENTO DEL PROBLEMA

La base de datos del Instituto Colombiano para el Fomento de la Educación Superior (ICFES) y el Portal Único del Estado Colombiano gov.com ofrecen información pública y relevante dentro del indicador académico SABER PRO, que facilita el seguimiento y el tratamiento de la misma por Institución y programa académico.

Además, en el programa de Ingeniera en Producción Acuícola perteneciente al Departamento de Recursos Hidrobiológicos de la Universidad de Nariño, existe la necesidad de evaluar el comportamiento de los indicadores académicos a través del tiempo, lo anterior mediante técnicas estadísticas multivariadas con el objetivo de facilitar la toma de decisiones en pro de una mejora continua en la educación superior.

Por todo lo anterior se establece el siguiente planteamiento; ¿Cuáles son las competencias genéricas que se relacionan e inciden en el puntaje en la prueba SABER PRO en el programa de Ingeniería en Producción Acuícola de la Universidad de Nariño 2016 – 2019?

3. OBJETIVOS

3.1.Objetivo general

Establecer mediante un análisis estadístico multivariado las principales variables y la incidencia de ellas sobre el puntaje de las pruebas SABER PRO dentro del programa de Ingeniería en Producción Acuícola de la Universidad de Nariño en los últimos cuatro años 2016-2019.

3.2.Objetivos específicos.

Analizar descriptivamente los datos del indicador SABER PRO del año 2016 al 2019 dentro del programa Ingeniería en Producción Acuícola de la Universidad de Nariño.

Aplicar un análisis multivariado por componentes principales para generar el menor número de variables que retengan la mayor variación del resultado Saber PRO y el nivel de correlación entre variables dentro de cada componente en los años del 2016 al 2019.

Realizar una regresión cuantílica y una regresión múltiple que permita comparar sus modelos a partir de las variables de los indicadores de la prueba SABER PRO del programa Ingeniería en Producción Acuícola de la Universidad de Nariño en los últimos cuatro años 2016-2019.

.

4. JUSTIFICACIÓN

La mayoría de estudios registrados con análisis multivariado para indicadores académicos, lo han realizado por separado para cada uno de ellos, algunos con respecto al rendimiento académico de una asignatura en específico o al de competencias para las pruebas de estado a nivel de programa académico, institución o diversas a nivel nacional. Entre ellos tenemos:

a. El estudio realizado y publicado por Cifuentes Garzón (2013), donde analiza algunas variables observables en el estudiante, con el objetivo de predecir el rendimiento en la prueba SABER PRO para el programa de Economía. Utilizó información contenida en las pruebas SABER 11 en cuanto al puntaje de cada componente, la edad y el género del estudiante. A través de regresión cuantílica se encontró un efecto positivo y significativo del género a favor de los hombres, el puntaje en matemáticas, en historia, en lenguaje y en idiomas sobre el resultado en la prueba SABER PRO, también determino que la edad predice mejores resultados en los estudiantes más jóvenes.

b. En referencia al estudio de Pérez-Pulido et al. (2016), se analizaron los resultados obtenidos en las pruebas de Estado Saber 11 por los estudiantes de recién ingreso a la Universidad de Santander para el periodo A-2016. Además de los puntajes se consideraron variables como: tipo de colegio, género, carrera a la que ingresa y región de procedencia. Se aplicó un análisis de correspondencias múltiples, un análisis factorial y una regresión cuantílica. Los resultados indican que los estudiantes que provienen de colegios privados y de Santander tienen mejores rendimientos en las competencias de inglés y matemáticas respectivamente.

c. En su trabajo de tesis, Marquín (2017), busco la predicción del rendimiento académico por medio de modelos de regresión lineal y análisis discriminante en la asignatura de algebra lineal, mediante el cual encontró las variables que inciden en el rendimiento con 131 estudiantes de la

Universidad. Como resultado se obtuvo que el modelo 3 de 11 realizados es el modelo con mayor porcentaje global de clasificación (90,8%), el modelo 10 aumento en un 1.6% con respecto a lo obtenido en el de regresión y el modelo 11 reduce su porcentaje en 2.3%, llegando a la conclusión de que ambos métodos son eficaces en la predicción del rendimiento académico.

d. En su trabajo de tesis Corredor Rivera et al. (2017) estudia la incidencia de diferentes dimensiones en las pruebas de matemáticas tipo ICFES mediante un análisis factorial y modelos de regresión ajustando un modelo semiparamétrico con distintos tipos de errores para determinar el modelo que mejor se ajuste, y determino que la dimensión cognitiva es afectada por la dimensión expresiva y afectiva en los estudiantes.

e. Manrique, Rodríguez et al. (2018) en su publicación estudia mediante un análisis multivariado las pruebas SABER PRO en el Departamento de Sucre, Colombia. Utilizo un análisis de clúster y un análisis de correspondencias simples para identificar las asociaciones de mayor peso entre los niveles y las competencias genéricas según la categorización en quintiles establecida por el ICFES, y concluye que las técnicas utilizadas para el análisis de los resultados en las competencias genéricas de las pruebas SaberPro fueron complementarías, indicando un nivel alto para Medicina, un nivel medio para Ingeniería, Ciencias Naturales y Exactas y niveles bajo y muy bajo para los demás grupos de referencia

5. MARCO TEÓRICO

5.1. Indicador

De acuerdo al DANE (2008), hace referencia a la relación de un conjunto de variables o características cuantitativas o cualitativas que muestran aspectos específicos y relevantes de una realidad social, cultural, financiera, administrativa o académica de una empresa, sector o institución. Se presenta una gran variedad de indicadores clasificados por la información que manejan junto con sus dos dimensiones, una cualitativa correspondiente a la descripción de la variable y la cuantitativa que se refiere a la expresión porcentual, numérica, promedio, etc. Además, su manejo o tratamiento posibilita establecer una evolución, desarrollo y predicción de una situación específica de aquella entidad o empresa.

5.1.1. Indicadores académicos. Proporcionan información relevante acerca de algún aspecto de la situación educativa, la mayoría de carácter cuantitativo, aunque se manejen datos como el sexo, y estrato, que complementan la realidad académica actual. Algunos de estos indicadores se clasifican como: inscritos, matriculados, egresados/graduados, deserción y puntaje de pruebas por asignaturas y/o competencias como es el caso de las pruebas de estado SABER 11 Y SABER PRO, estos indicadores que se asocian con diferentes características de la población estudiantil, de acuerdo con el SNIES, los define como:

Inscritos: Solicitud de aquellos que desean el ingreso a un programa académico específico dentro de una Institución.

Admitidos: Personas que han cumplido con todos los requisitos de ley que exige la Institución y es aceptado en calidad de estudiante.

Matriculados a primer curso: son aquellos estudiantes que formalizan sus matrículas en el primer curso o semestre en el programa académico dentro de la Institución que fue admitido.

Matriculados: Estudiantes de todas las cohortes o semestres en el o los programas académicos.

Egresados: De acuerdo SPADIES, es aquel estudiante que termina con sus materias y semestres pero que no ha obtenido el grado o título.

Estudiantes graduados: Según SPADIES, se refiere aquel estudiante que ha recibido el grado por parte de la Institución

Deserción: Estado de un estudiante que según SPADIES, no registra matricula por varios periodos consecutivos, tampoco se encuentra egresado o graduado.

Puntaje SABER 11°: De acuerdo con Icfes (2017), promedio generado por cada estudiante que finalizo el grado undécimo u obtuvo su título de bachiller permitiéndole ingresar a la educación superior.

Prueba SABER PRO: De acuerdo con Icfes, es un examen de estado para estudiantes que han finalizado el 75% de sus créditos y espera evaluar el grado de desarrollo por competencias generales y especificas del programa académico universitario profesional.

5.2. Análisis multivariado.

El análisis multivariado de acuerdo con Kachigan (1991), es un área de la estadística que provee mediante diferentes técnicas la interpretación adecuada de un conjunto de variables o características, la simplificación y la visualización de la relación existente entre ellas, medidas en un conjunto de objetos y/o personas.

El campo de aplicación está en todas las áreas de conocimiento, desde finanzas hasta la agricultura, y es fundamental para la toma de decisiones dentro del contexto en el que se encuentre el investigador.

De acuerdo con Salvador-Figuereas (2000) y Aldas Manzano and Uriel Jimenez (2017), las técnicas aplicadas dentro de este estudio se encuentran inmersas dentro de tres grandes grupos, dos de ellos son:

Métodos de dependencia, donde se busca determinar si el conjunto de variables independientes afecta al conjunto de variables dependientes, y se subdivide a su vez en dos grupos; dependiente métrica y no métrica, en estos se contemplan los análisis de regresión, supervivencia, correlación canónica, discriminantes, etc.

Métodos de interdependencia, donde por medio de datos métricos y no métricos, se busca identificar que variables están relacionadas, aunque no se pueda distinguir si estas son dependientes e independientes, mediante análisis de componentes principales, factorial, escalas multidimensionales y análisis de cluster.

5.2.1. Representaciones pictóricas de datos multivariados. Según Castaño (2017), son imágenes que representan tres o más variables de cada individuo, objeto o unidad experimental, no trasmiten información numérica, solo ayudan a reconocer observaciones similares y entre ellos menciona:

Gráfico de estrellas, el cual para Castaño (2017), si $p \geq 2$ variables, se construyen círculos de radio fijo con p rayos igualmente espaciados, donde las longitudes de los radios representan los valores de las variables.

Gráfico de Andrews: Identifica según Castaño (2017), agrupamientos de observaciones, el resultado para cada una es una onda formada por funciones seno y coseno.

Caras de Chernoff: Aquí según Castaño (2017), se usan varias características de la cara para representar los datos de las variables, según el programa estadístico que se utilice se puede representar de 18 a 20 variables (R, SYSTAT).

Caras asimétricas: En este grafico según Flury and Riedwyl (1981), los parámetros del lado derecho de la cara pueden variar independientemente de aquellos del lado izquierdo y puede expresar hasta 36 variables (faceplot).

Gráfico de barras e histograma: Estas representaciones de acuerdo con Espinel F. et al., (2009) son las más comunes y se las utiliza para trabajar con datos numéricos discretos, continuos, categóricos y grupo de intervalos, cada barra tiene la misma anchura y puede contener uno o varios niveles dentro de una categoría.

Caja y bigotes: Asimismo Espinel F. et al., (2009) menciona que las gráficas de cajas o grafico de tallos y hojas son adecuadas para la representación de la mediana, cuartiles, valor mínimo y máximo, permitiendo realizar comparaciones entre las diferentes cajas que representan la misma característica discriminada entre ellas.

5.2.2. Reducción dimensional. Recoge de acuerdo con Montanero Fernández (2018), aquellas técnicas que tiene como objeto simplificar el conjunto de datos, entre ellas el análisis de componentes principales.

Análisis de componentes principales (ACP). Este tipo de análisis fue la predicción de los trabajos de Karl Pearson sobre ajustes ortogonales, en esta técnica se puede reducir la dimensionalidad, de manera que se pueda explicar la información disponible en la menor cantidad

de variables las cuales estarán representadas mediante componentes con una mínima perdida de información, el análisis de los datos se puede realizar por filas(individuos) o por columnas (variables), y es relativamente fácil interpretar los resultados sobre los principales componentes o dimensiones(Peña-Méndez, 2014).

En primer lugar, se debe establecer si es necesario la centralización de las variables para que tengan todas media cero y de realizarse se da comienzo a la determinación de los componentes mediante el cálculo de los loadings con los que se maximiza la varianza, una forma de optimizar es mediante los eigenvector y eigenvalue, los componentes demostraran la relación lineal entre las variables originales y su grado de variabilidad, se puede tomar la ayuda de gráficos entre el valor propio (eje y) y su componente (eje x)(Joaquín, 2017).

El resultado inicial es una matriz la cual es difícil de interpretar por lo que es indispensable según Pérez and Medrano (2010), obtener una matriz adicional rotada (cuando hay más de un factor o componente en la solución), la rotación concentra la varianza de las variables en menos componentes, pueden ser ortogonales u oblicuas, y esto se logra por medio de procedimientos gráficos de rotación, métodos como Varimax y Promax o de paquetes estadísticos que la generan, de esta manera incrementan las correlaciones positivas extremas.

Los componentes tienen algunas propiedades como el conservar la variabilidad original, es decir que la suma de sus varianzas es igual a la inicial, también conservan la varianza generalizada, que es el determinante producto de los valores propios a la matriz de covarianzas (Everitt & Hothorn, 2011).

La inercia o varianza asociado a cada variable con su respectivo porcentaje acumulado, la cual es igual a el cociente entre su varianza, el valor propio asociado al vector propio y la suma de los valores propios de la matriz(Joaquín, 2017).

La contribución, calidad de asociación de cada variable y por último un análisis de varianza donde procuramos explicar las dos primeras dimensiones por las variables cuantitativas que los contienen en mayor proporción y correlación (Kassambara, 2017).

5.2.3. Análisis de regresión. Un análisis de regresión se resume en la modelación de la relación existente entre variables predictores en las variables de respuestas o dependientes (Das et al., 2019).

Regresión lineal simple y múltiple. Son las regresiones más comunes, la técnica lineal simple o bivariado trabaja con una sola variable independiente o explicativa y la técnica de regresión múltiple donde la cantidad de variables predictores es igual o mayor a dos, en este último análisis se tiene en cuenta la siguiente notación matemática de modelo o ecuación (Rodríguez & Catalá, 2001):

$$\mathrm{Y} = a + b1x1 + b2x2 + \cdots + bmxn + e$$

$$presente = a + b1pasado + b2futuro + \cdots + e$$

Donde; Y es la variable a predecir, e es el error que comentemos en la predicción de los parámetros y a, b1x1, b2x2, bmxn son parámetros a estimar.

Los pasos más relevantes en la regresión múltiple es la determinación de la bondad del ajuste, la elección del mejor modelo y la estimación o cumplimiento de los supuestos o condiciones estadísticas de los residuos (Rodríguez & Catalá, 2001).

Regresión cuantílica. Ford (2015) menciona que, aunque es común estimar en los análisis de regresión la media de la variable condicional para establecer la relación deseada, también se puede estimar la mediana, o algún cuantil, es ahí donde entra a trabajar la técnica de regresión cuantílica.

El poder realizar una regresión desde cualquier parte de la información permite conocer la influencia de los predictores desde el mínimo y máximo rango de la variable respuesta o dependiente, esta técnica es muy útil cuando el supuesto de varianza no se cumple y se quiere estimar múltiples cuantiles simultáneamente (Das et al., 2019), si la varianza dentro del conjunto de datos cambiar cada vez que lo hace el predictor no se estaría cumpliendo las condiciones necesarias para realizar una regresión por mínimos cuadrados y la regresión cuantílica junto con la de mínimos cuadrados ponderados son las únicas técnicas que trabajan con varianzas no constantes (Ford, 2015).

Tomando como referencia el trabajo realizado por Galvis A., (2012) y citado por (Ordoñez-Castaño & Sanabria-Domínguez, 2014) un modelo de análisis de varianza ANOVA simple entre cuartiles es importante para estimar la diferencia de la variable dependiente entre los individuos para cada percentil de la distribución calculada.

5.2.4. Supuestos estadísticos. Los tres supuestos que deben de verificarse en pro de no conducir a resultados erróneos, es decir antes de correr el métodos multivariados son: la normalidad, multicolinealidad y homocedasticidad, y se recomienda realizar un análisis exploratorio inicial de los datos, así se determinan datos atípicos (valores extremos), que distorsionan los resultados, ya que este tipo de análisis parte de una matriz de correlaciones entre variables, y se estiman en base a la media o promedio de los valores de dichas variables, el grafico de cajas y bigotes al igual que las distancias de Cook con valores menores a 1 determinan la no influencia de los datos atípicos en caso de su existencia (Cook, 1977).

Supuesto de normalidad. Se recomienda de acuerdo con Pérez and Medrano (2010), utilizar un análisis visual de los gráficos q-qplot al igual que las gráficas de normalidad, los cuales proporcionan una lineación de la distribución normal, ya que una prueba de Shapiro-Wilk y Kolmorogov-Smirnof, pueden ser demasiado sensibles a pequeñas desviaciones de la normalidad cuando se trabaja con muestras de gran tamaño, tal y como menciona López (2004). Otra alternativa es estimar los índices de asimetría y curtosis, si los valores están dentro del umbral +-1,5 indican variaciones leves de la normal y por ende es adecuado realizar los análisis (George and Mallery 2001, p. 3), además, también son muy utilizadas las técnicas multivaradas para normalidad denominadas Mardia, Henze-Zirkler, Royston y Doornik-Hansen (Doornik & Hansen, 1994; Porras Ceron, 2016).

Supuesto de linealidad y/o multicolinealidad. Este supuesto puede ser evaluado según J.F. *et al*. (1999), examinando los diagramas de dispersión, si los puntos se organizan a lo largo de una línea recta, puede mantenerse el supuesto de linealidad de las relaciones. Para una evaluación estadística puede realizarse una prueba de análisis de regresión múltiple. Además, acuerdo con Martínez Arias (1999), simplemente se puede observar la matriz de correlación y sus coeficientes entre cada par de variables incluyendo aquella dependiente, con esto se puede determinar la existencia o no de una correlación, dependencia lineal o la ausencia de la misma.

La realización de un análisis de inflación de la varianza (VIF) puede ser muy útil para comprobar la correlación entre variables, en este análisis valore inferiores a 10 son aptos para demostrar la baja correlación, valore entre 10 y 30 una correlación media y las variables que sobrepasen los 30 es recomendable descartarlas para evitar resultados erróneos o desfasados(Fernández Rodriguez, 2015).

Por último, la autocorrelación entre los residuos en una regresión también se puede comprobar determinado el *p-valor* mediante un test de Durbin-Watson, donde si el p-valor es >0.05 y si el coeficiente esta entre 0 a 4, siendo 2 prueba de no autocorrelación, cercano a cero correlación positiva y mayores a 2 autocorrelación negativa(Parra Rodríguez, 2016).

Supuesto de homocedasticidad (homogeneidad de varianzas). La detección de este supuesto se realiza por medio de diferentes test paramétricos como el de Bartlett y levene y/o test de Brown&Forsythe, los cuales contrastan si la varianza es igual o no en los diferentes grupos que componen la variable dependiente.(Brown & Forsythe, 1974; Correa et al., 2006; Parra Rodríguez, 2016).

6. METODOLOGÍA

6.1. Método

Estadístico e investigativo

6.2. Población

Estudiantes de último semestre que han cumplido con el 75% de sus créditos en el programa de Ingeniera en Producción Acuícola con resultados de sus pruebas SABER PRO en los últimos cuatro años (2016-2019).

6.3. Enfoque y tipo de variables

Enfoque Mixto, debido a que hay información cuantitativa y cualitativa analizadas por separado, las variables corresponden edad, genero, estrato y puntaje por competencias genéricas de las pruebas SABER PRO con su respectivo puntaje global o promedio.

6.4. Búsqueda, recolección, depuración y presentación de la información.

La base de datos del Instituto Colombiano para el Fomento de la Educación Superior (ICFES) y el Portal Único del Estado Colombiano, facilitan por medio de sus portales web la información y/o datos necesarios acerca de las pruebas SABER PRO entre el año del 2016 y 2019.

La información recolectada consolido una base de datos con un total de 9 variables con 64 registros para un total de 576 datos distribuidos entre las variables apartadas para el indicador de la prueba SABER PRO.

6.5. Variables

Las variables independientes o predictores objeto del presente estudio son:

- Periodo: Año en el que el estudiante presento la prueba de estado.

- Género: Sexo de cada estudiante registrado, se denota con una M para masculino ó una F para femenino.
- Estrato: Hace referencia a la estratificación socio-económica, el cual se registra de 1 a 3.
- Edad: Años de vida que presentaba el estudiante.
- Puntaje por competencias genéricas: es la cantidad de puntos obtenidos con respecto al global en su prueba SABER PRO por cada competencia de comunicación escrita, razonamiento cuantitativo, lectura crítica, competencia ciudadana e inglés.

La variable dependiente o de respuesta está dado por el puntaje global total, el cual corresponde solo al promedio de los puntajes de cada competencia genérica.

6.6. Instrumentos

Para la ejecución de todos los análisis de datos se construyó una matriz de organización de datos en Excel, se utilizó el programa estadístico Rstudio y algunos paquetes estadísticos como FactoMineR. Quantreg, Stats para análisis y factoextra para la visualización de datos, entre otros.

6.7. Modelo estadístico

Se utilizó el análisis descriptivo, el análisis de componentes principales, la regresión múltiple y cuantílica para caracterizar la población estudiantil y encontrar las principales variables con mayor incidencia sobre aquella dependiente o de respuesta (puntaje global) mediante modelos.

6.7.1. Análisis descriptivo. Todas las variables se sometieron a un análisis descriptivo mediante la representación de gráficos de barras, histogramas y boxplot mas el cálculo de las medidas de media y desviación estándar de cada competencia genérica a través del tiempo, lo anterior para facilitar la caracterización de la población estudiantil, sus resultados en las pruebas

SABER PRO y la presencia de datos atípicos en las diferentes observaciones, su determinación facilita la depuración e la base de datos original.

6.7.2. Análisis de componentes principales. En primer lugar, se estandarizaron las variables para que tengan media cero y desviación estándar 1, lo anterior facilitando el análisis en caso de que las varianzas entre las variables no sea la misma.

En segundo lugar, se calcularon los componentes principales y los pesos de cada uno, es decir la cantidad de información que recoge cada uno de ellos, para con ese valor establecer las combinaciones lineales de las variables originales y sus pesos correspondientes por cada componente.

En tercer lugar, se determinó la varianza retenida en cada componente principal y así apreciar sobre cuales recaía el mayor porcentaje (>50%).

En cuarto lugar, se determinó la contribución (contrib) para determinar las variables más relevantes, la calidad o grado de asociación entre variables(cos2) y la correlación entre las mismas (Coord.).

Y en último lugar, se determinó el nivel de significancia entre variables, calculando el p-valor para cada grupo de variables de mayor incidencia dentro de las dos primeras componentes.

6.7.3. Análisis de Regresión y supuestos estadísticos. Uno de los objetivos fue determinar los *p-valor* para cada supuesto estadístico, los cuales se determinaron para los residuos después de general el modelo de regresión y sobre todas las variables predictores y la variable respuesta antes del mismo análisis.

6.7.3.1. Condiciones o supuestos estadísticos.

Normalidad. Para este supuesto se quiso tener en cuenta diferentes test que confirmar la existencia o no de una distribución normal, se generaron las curvas de normalidad, los coeficientes de asimetría y curtosis estandarizados donde el rango optimo debe estar entre -2 y 2, los test de Mardia de asimetría y Mardia para curtosis, el test de Henze-Zirkler, Royston, Doornik Hansen, E-statistic y el test de Shapiro-Wilk, este último solo se trabajó para determinar la normalidad de los residuos.

Se establecieron las siguientes hipótesis teniendo en cuenta que el nivel de significancia es de Alfa=0.05

H0: La muestra proviene de una distribución normal (Si p-valor≥0.05 se acepta Ho)

H1: La muestra no proviene de una distribución normal (Si p-valor<0.05 se rechaza Ho)

Multicolinealidad. Por la matriz de correlación previo a la regresión se midió el coeficiente entre pares de variables y el coeficiente general, los cuales van de -1 a 1 y se estableció que entre más cercano a 1 este el coeficiente entre el predictor y la variable respuesta determina una dependencia o relación deseable.

Condiciones: Si $r = 1$, la relación es positiva perfecta, $0 < r < 1$ la relación es positiva, $r = 0$ no hay relación lineal, $-1 < r < 0$ la relación es negativa, $r = -1$ la relación es negativa perfecta

La Relación lineal entre predictores y variable respuesta posterior a generar el modelo por regresión se tuvo en cuenta mediante la representación gráfica de cada residuo con su variable o competencia genérica más el intercepto, cumpliendo la condición de que los residuos deben distribuirse lo más cercano a cero.

Independencia, No autocorrelación o No multicolinealidad entre predictores. Mediante la matriz de correlación se determinó la independencia entre predictores mientras el coeficiente este lo más cercano a cero y posterior a generar el modelo por regresión múltiple se realizó un análisis de inflación de la varianza (VIF), y un test de hipótesis de Durbin-Watson, con un alfa de 0.05 donde si $p>0.05$ se acepta la no autocorrelación entre predictores, es decir, se acepta la Hipótesis nula.

Homocedasticidad. Se trabajó con el test de Bartlet, levene y de Brown-Forsyth, este último teniendo en cuenta la mediana como unidad central, además, se manejó el test de Breusch-Pagan para determinar la homocedasticidad de los residuos, y con un alpha =0.05 se estableció las siguientes hipótesis:

H0: Se acepta la hipótesis nula si las varianzas son homogéneas si p-valor ≥ 0.05.

H1: Se rechaza hipótesis nula si las varianzas no son homogéneas si p-valor<0.05.

Datos atípicos influyentes. Este análisis se realizó posterior a la regresión múltiple por mínimos cuadrados y se determinó mediante la distancia de Cook donde se desea que los valores sean menores a 1 para concluir la no influencia de los datos atípicos.

6.7.3.2. Regresión múltiple. En primer lugar, se establecen las variables que se consideran predictores y aquella dependiente por medio del modelo que a nuestro criterio explica el desempeño de cada estudiante por medio del puntaje promedio o global de las pruebas saber PRO para el programa de pregrado.

En segundo lugar, se selecciona los mejores predictores determinando la calidad del modelo mediante el método AIC, el predictor con el mayor p-valor puede ser excluido y así reajustar el modelo excluyendo aquella variable.

Generado el nuevo modelo de observa si el nuevo valor de R^2 se modificó para determinar si la variable excluida afectaba en gran medida la variabilidad de y.

Como último paso se determinan los intervalos de confianza y de nuevo se validan los supuestos o condiciones estadísticas de los residuos, de no cumplirse algún supuesto se reajustará el modelo.

6.7.3.3. Regresión cuantílica. En primera instancia se establecen las variables que se consideran predictores, aquella variable dependiente y los cinco cuantiles sobre los cuales queremos general los modelos, 0.1, 0.25, 0.5(mediana), 0.75 y 0.90.

Se generaron los coeficientes de cada variable, el *p-valor* asociado a cada predictor y la por medio de las estimaciones del intercepto las representaciones de la pendiente de regresión para cada cuantil para determinar al influencia y significancia.

Por último, se generó varios ANOVAS por cada par de modelos generados para determinar la existencia u ausencia de diferencias significativas entre ellos.

7. ASPECTOS ÉTICOS O CONFIDENCIALIDAD DE LA INFORMACIÓN

La información correspondiente a las pruebas SABER PRO son generados por el ICFES, el cual facilita al público reportes con valores promedio y de desviación estándar por cada competencia genérica y/o especifica en la población estudiantil seleccionada, los resultados individuales de diferentes instituciones a nivel nacional la presenta el portal www.datos.gov.co , y al ser abiertos al público se tomó lo referente al programa de Ingeniería en producción acuícola de la Universidad de Nariño.

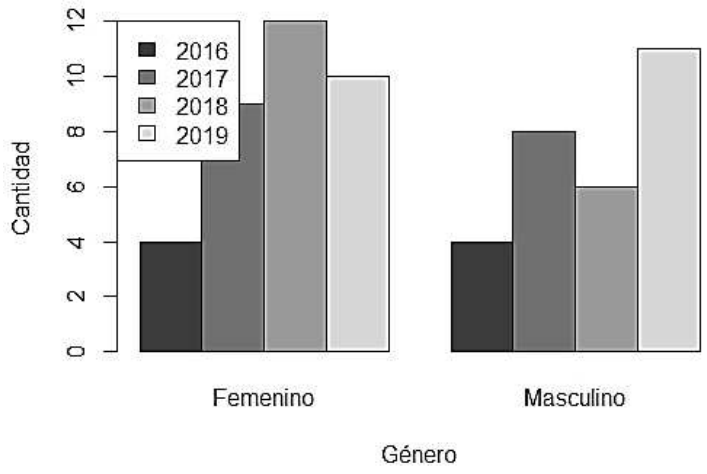
En el presente documento no se publica información individualizada con el fin de respetar la confidencialidad de cada estudiante.

8. RESULTADOS

8.1. Análisis descriptivo.

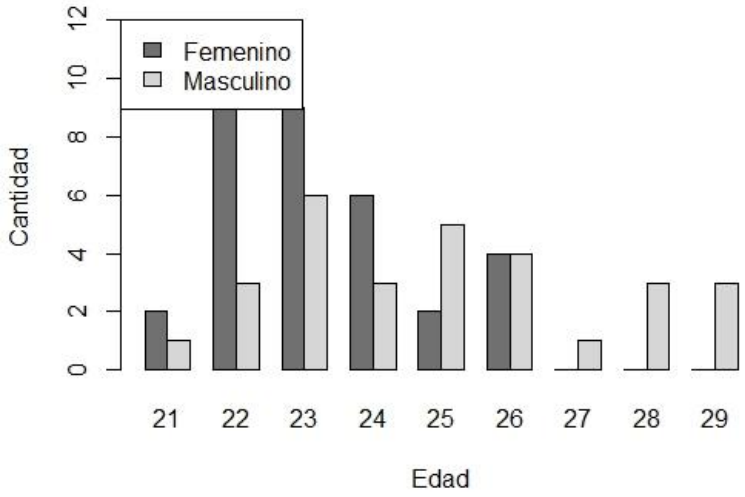
8.1.1. Género. Los resultados indican que, en el año 2016 se presentaron a la prueba de estado 8 estudiantes, 4 de ellos de género masculino y 4 femeninos, en el año 2017 con un total de 17 registros 8 de ellos hombres y 9 mujeres, el año 2018 presento 18 registros 6 de ellos hombres y 12 mujeres y, por último, el año 2019 con 11 hombres y 10 mujeres. En general existieron 35 registros femeninos y 29 masculinos (Figura 1).

Figura 1. Género de los estudiantes que participaron en la prueba SABER PRO durante el periodo del 2016 y 2019



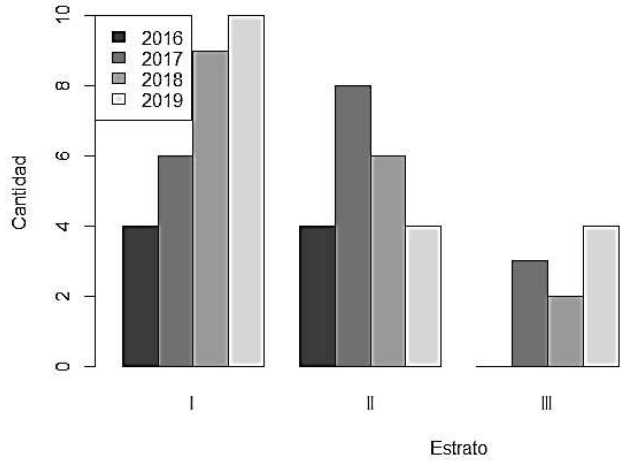
8.1.2. Edad. En la Figura 2 se presentan las edades de aquellos estudiantes que aplicaron la prueba de estado, corresponden a edades entre los 21 y 29 años, con una mayor participación de los estudiantes con 54 registros correspondiente al 84% de la población con edades entre los 22 y 26 años y una prevalencia del género femenino en este rango, mientras que los estudiantes con 27 años o más solo fueron 6 y de género masculino.

Figura 2. Edad por género de los estudiantes que lograron participar de el examen SABER PRO entre los años 2016 y 2019



8.1.3. Estrato. En la Figura 3 se puede apreciar que para el año 2016 solo se presentaron 8 estudiantes de estrato 1 y 2 en iguales proporciones, para el año 2017 se presentaron 6 estudiantes de estrato I, 8 de estrato II y solo 3 de estrato III, para el año 2018, se presentaron 9 de estrato I, 6 estudiantes de estrato II y solo 2 de estrato III, para el año 2019, se presentaron 10 estudiantes de estrato I, 4 de estrato II y 4 de estrato III. En los dos últimos años académicos, 2018 y 2019, no registraron estrato 4 estudiantes y se clasifican como valores perdidos.

Figura 3. Estrato socioeconómico presentado por los estudiantes en las pruebas SABER PRO



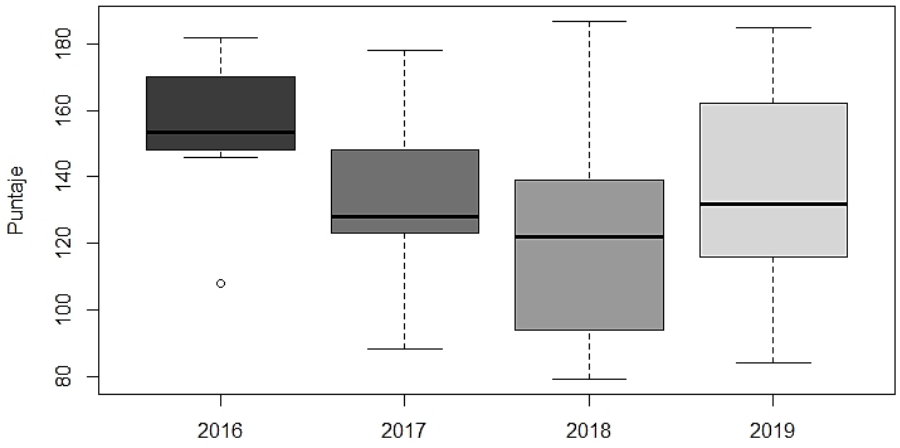
8.1.4. Competencia ciudadana. La Tabla 1 indica que para el año 2016 la media fue de 154 con una desviación estándar de más o menos 23puntos, para el año 2017 fue de 133±25puntos, el año 2018 presento un valor de 122±30puntos, el año 2019 una media de 136±26 puntos y un puntaje global promedio entre los cuatro años de 134±28 puntos

Tabla 1. Media y desviación estándar de competencia ciudadana por periodos 2016-2019

Año	2016	2017	2018	2019	Global
Media	154	133	122	136	134
Desviación estándar	23	25	30	26	28

Además, se puede observar en la Figura 4 que el año 2016 presenta puntajes desde los 146 a los 182 puntos con respecto al global, con una mediana de 155, el año 2017 demuestra valores entre los 88 a los 178 puntos y una mediana de 128, el año 2018 cuenta con puntajes entre los 79 y los 187 puntos con una mediana de 122 y el año 2019 presento datos entre los 84 a 185 puntos y una mediana de 132 puntos. Solo se observó un dato atípico, presente en el año 2016 con un valor de 108 puntos.

Figura 4. Puntajes anuales de competencia ciudadana con respecto al puntaje global.



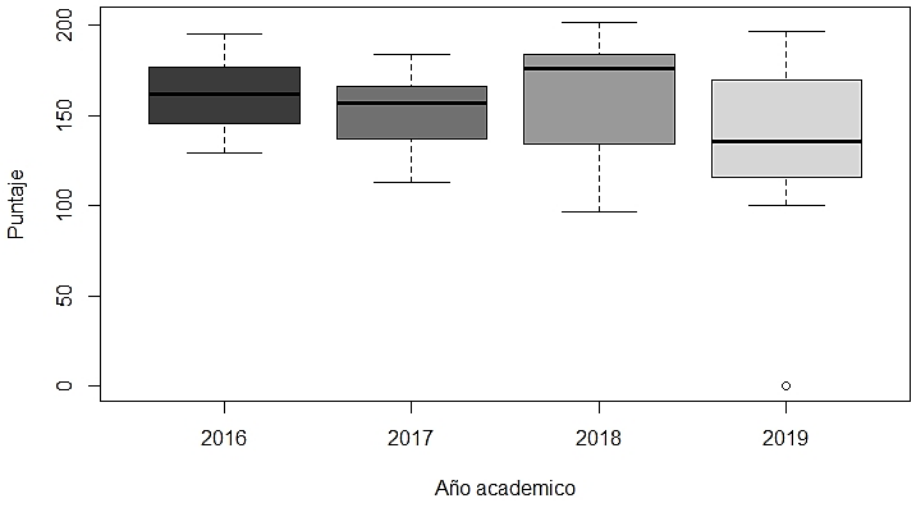
8.1.5. Comunicación escrita. La Tabla 2 indica que para el año 2016 la media fue de 162 con una desviación estándar de más o menos 24 puntos, para el año 2017 fue de 149±21puntos, el año 2018 presento un valor de 161±30puntos, el año 2019 una media de 130±52 puntos y un puntaje global promedio entre los cuatro años de 148±38 puntos.

Tabla 2. Media y desviación estándar de comunicación escrita por periodos 2016-2019

Año	2016	2017	2018	2019	Global
Media	162	149	161	130	148
Desviación estándar	24	21	30	52	38

Además, se puede observar en la Figura 5 que el año 2016 presenta puntajes desde los 129 a los 195 puntos con respecto al global, con una mediana de 162, el año 2017 demuestra valores entre los 113 a los 184 puntos y una mediana de 157, el año 2018 cuenta con puntajes entre los 97 y los 202 puntos con una mediana de 177 y el año 2019 presento datos entre los 100 a 197 puntos y una mediana de 137 puntos. Solo se observó dos datos atípicos, presentes en el año 2019 con un valor de 0 puntos.

Figura 5. Puntajes promedios anuales de comunicación escrita con respecto al global



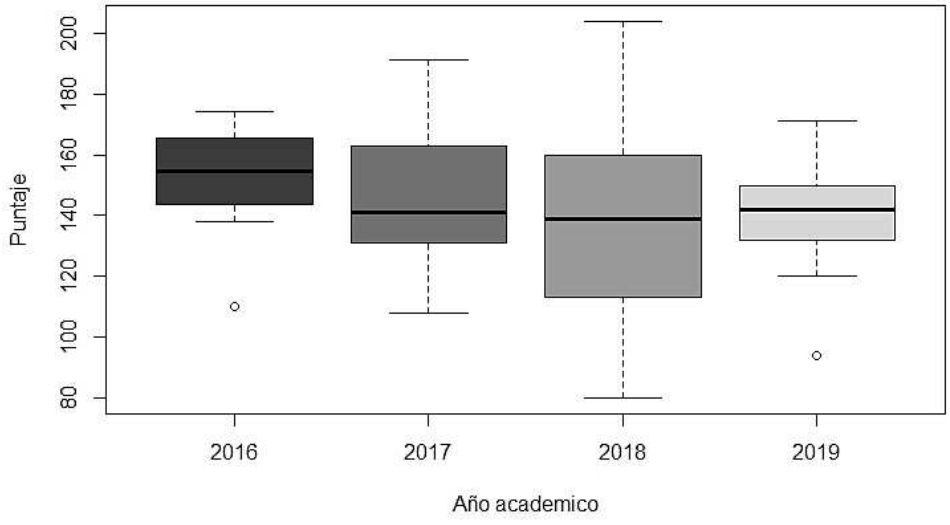
8.1.6. Lectura crítica. La Tabla 3 indica que para el año 2016 la media fue de 151 con una desviación estándar de más o menos 20 puntos, para el año 2017 fue de 146±24puntos, el año 2018 presento un valor de 141±30puntos, el año 2019 una media de 141±17 puntos y un puntaje global promedio entre los cuatro años de 144±23 puntos.

Tabla 3. Media y desviación estándar de lectura crítica por periodos 2016-2019

Año	2016	2017	2018	2019	Global
Media	151	146	141	141	144
Desviación estándar	20	24	30	17	23

Además, se puede observar en la Figura 6 que el año 2016 presenta puntajes desde los 138 a los 174 puntos con respecto al global, con una mediana de 156, el año 2017 demuestra valores entre los 108 a los 191 puntos y una mediana de 141, el año 2018 cuenta con puntajes entre los 80 y los 204 puntos con una mediana de 139 y el año 2019 presento datos entre los 120 a 171 puntos y una mediana de 143 puntos. Solo se observó dos datos atípicos, presentes en el año 2016 y 2019 con valores de 110 y 94 puntos respectivamente.

Figura 6. Puntajes promedios anuales de lectura crítica con respecto al global.



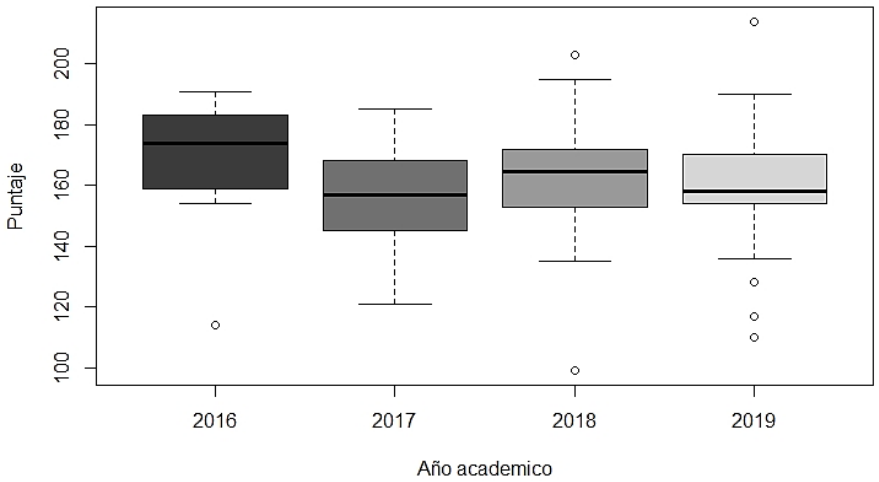
8.1.7. Razonamiento cuantitativo. La Tabla 4 indica que para el año 2016 la media fue de 167 con una desviación estándar de más o menos 25 puntos, para el año 2017 fue de 157±17puntos, el año 2018 presento un valor de 162±23puntos, el año 2019 una media de 160±24 puntos y un puntaje global promedio entre los cuatro años de 160±22 puntos.

Tabla 4. Media y desviación estándar de razonamiento cuantitativo por periodos 2016-2019

Año	**2016**	**2017**	**2018**	**2019**	**Global**
Media	167	157	162	160	160
Desviación estándar	25	17	23	24	22

Además, se puede observar en la Figura 7 que el año 2016 presenta puntajes desde los 154 a los 191 puntos con respecto al global, con una mediana de 177, el año 2017 demuestra valores entre los 121 a los 185 puntos y una mediana de 157, el año 2018 cuenta con puntajes entre los 135 y los 195 puntos con una mediana de 165 y el año 2019 presento datos entre los 136 a 190 puntos y una mediana de 167 puntos. Se observó siete datos atípicos, el valor de 114 para el periodo del 2016, valores de 99 y 203 para el 2018 y valores de 110, 117, 128 y 214 puntos para el 2017.

Figura 7. Puntajes promedios anuales de razonamiento cuantitativo con respecto al global



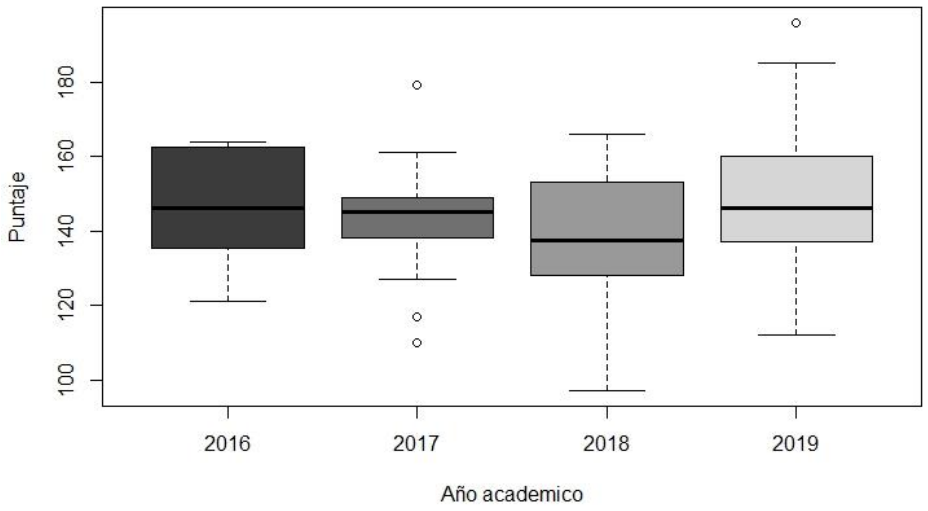
8.1.8. Inglés. La Tabla 5 indica que para el año 2016 la media fue de 146 con una desviación estándar de más o menos 16 puntos, para el año 2017 fue de 143±16puntos, el año 2018 presento un valor de 139±17puntos, el año 2019 una media de 148±21 puntos y un puntaje global promedio entre los cuatro años de 144±18 puntos.

Tabla 5. Media y desviación estándar de inglés por periodos 2016-2019

Año	2016	2017	2018	2019	Global
Media	146	143	139	148	144
Desviación estándar	16	16	17	21	18

Además, se puede observar en la *Figura 8* que el año 2016 presenta puntajes desde los 121 a los 164 puntos con respecto al global, con una mediana de 146, el año 2017 demuestra valores entre los 127 a los 161 puntos y una mediana de 145, el año 2018 cuenta con puntajes entre los 97 y los 166 puntos con una mediana de 138 y el año 2019 presento datos entre los 112 a 185 puntos y una mediana de 146 puntos. Se observó cuatro datos atípicos, el valor de 110, 117 y 179 para el periodo del 2017, y el valor de 196 puntos para el 2019.

Figura 8. Puntajes promedios anuales de inglés con respecto al global



8.2.Análisis Multivariados.

Previo a realizar los diferentes tipos de análisis se determinaron 13 observaciones atípicas, algunas coincidiendo dentro del mismo individuo o registro. La base de datos final conto con 53 registros y 5 variables de tipo cuantitativo (competencias genéricas), trabajando con un total de 265 datos para el análisis de componentes principales y con 318 observaciones para los análisis de regresión, donde se tuvo en cuenta la variable dependiente (puntaje global).

8.2.1. Análisis de Componentes Principales. Como se observa en la Tabla 6 las variables independientes cumplen el supuesto de normalidad al generar un p-valor en todos los test mayor a 0.05, por lo que se acepta la hipótesis nula y se genera el análisis de componentes principales.

Tabla 6 p-valor generado entre predictores en los diferentes test de normalidad.

Test	*p-valor*
Mardia-skewness	0.98
Mardia-kurtosis	0.23
Henze-zirkler	0.28
Royston	0.79
Doornik-Hansen	0.77
E-statistic	0.59

El primer resultado que obtenemos por medio del comando rotation en R son los diferentes valores de peso/importancia que tiene cada variable sobre cada componente (Tabla 7), cada uno como el resultado de la combinación lineal de su variable original.

Tabla 7. Pesos o cargas (eigenvector) para cada componente.

Competencia	PC1	PC2	PC3	PC4	PC5
Comunicación escrita (*ce*)	-0.204	0.61	-0.73	-0.233	-0.05
Razonamiento cuantitativo (*rc*)	0.52	-0.30	-0.46	0.29	-0.58
Lectura crítica (*lc*)	0.54	0.09	0.19	-0.81	-0.11
Competencia ciudadana (*cc*)	0.60	0.18	-0.15	0.28	0.71
Inglés (*i*)	0.18	0.70	0.45	0.36	-0.38

$$PC1 = -0.204ce + 0.525rc + 0.54lc + 0.60cc + 0.182i$$

$$PC2 = 0.61ce - 0.305rc + 0.09lc + 0.18cc + 0.70i$$

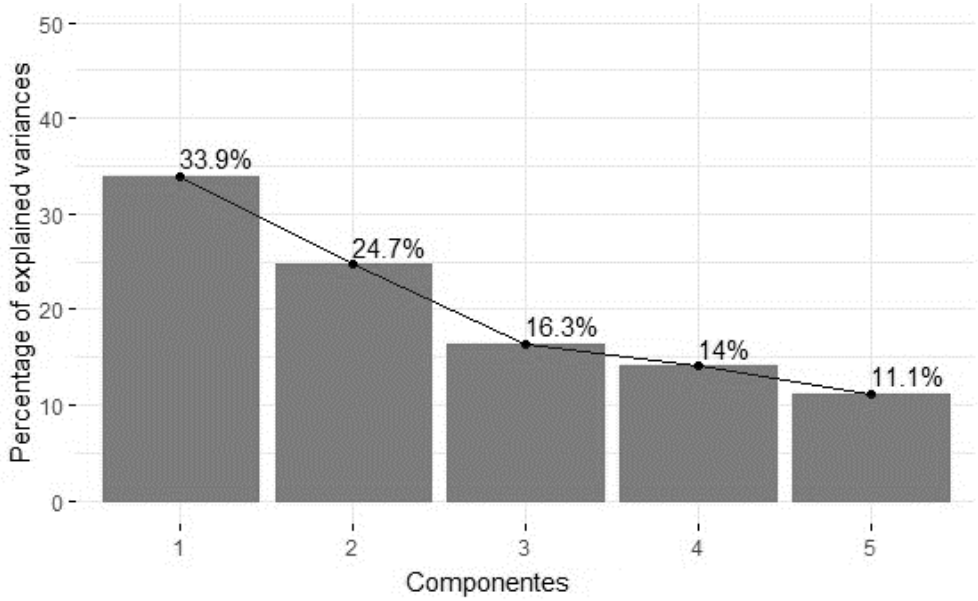
Los pesos asignados en la primera componente corresponden a la variable *Razonamiento cuantitativo y lectura crítica* con aproximadamente de 0.5, a *Competencia ciudadana* con 0.6 y a *Comunicación escrita e inglés* con -0.2 y 0.2 respectivamente. En la segunda componente, la variable *Comunicación escrita e inglés* tienen pesos de 0.6 y 0.7 mientras que *Razonamiento cuantitativo*, *Lectura crítica* y *Competencia ciudadana* valores de -0.3, 0.1 y 0.2 respectivamente.

Variación retenida por cada componente. En la Tabla 8 y en la Figura 9 se puede apreciar que la primera componente explica el 33.94% de la varianza observada en los datos y la segunda el 24.66%, explicando ambas el 58.60% de la varianza observada, mientras que las dos últimas componentes no superan por separado el 14% de la misma.

Tabla 8. Variación (eigenvalue) retenida por cada componente.

	PC1	PC2	PC3	PC4	PC5
Eigenvalues	1.70	1.23	0.82	0.70	0.55
Porcentaje de la varianza	33.90	24.70	16.31	14.02	11.10
Porcentaje acumulado de la varianza	33.94	**58.60**	74.91	88.93	100.00

Figura 9. Porcentaje de la variación retenida por cada componente



Contribución (contrib). La variable comunicación escrita tiene una contribución en el componente 2 y 3 con 37.01 y 53.21 respectivamente, razonamiento cuantitativo en todos los componentes pero el mayor valor en componente 1 y 5 con 27.60 y 33.49 respectivamente, lectura crítica en los componentes 1 y 4 con 29.36 y 64.91 respectivamente, competencias ciudadanas dentro los componentes 1 y 5 con 35.58 y 50.81 respectivamente, ingles en la segunda componente con 49.65 y en la componente 3, 4 y 5 con valores menores a 20. Los espacios en blanco significan que no existe alguna contribución notable por parte de las variables dentro de aquella componente principal.

Figura 10. Contribución de las variables en los cinco primeros componentes ó dim.

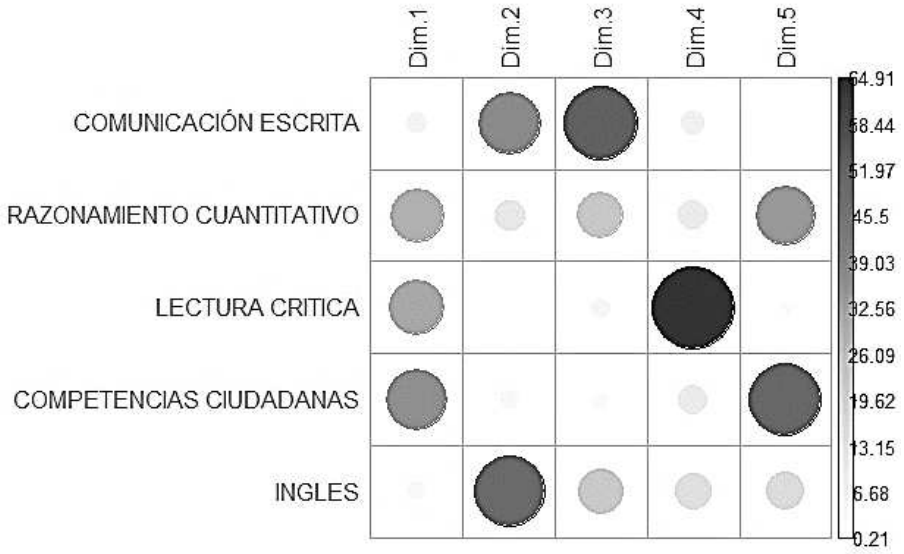


Tabla 9. Contribución en porcentaje de las variables sobre los componentes principales

Contribución	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
Comunicación escrita	4.16	37.01	53.21	5.41	0.21
Razonamiento cuantitativo	27.60	9.31	20.99	8.61	33.49
Lectura critica	29.36	0.84	3.58	64.91	1.32
Competencia ciudadana	35.58	3.19	2.36	8.06	50.81
Ingles	3.31	49.65	19.87	13.00	14.17

Calidad o grado de asociación entre variables (cos2). El grado de asociación se presenta en todos los componentes en diferente proporciones, en el componente 1, 2, 3, 4 y 5 la variable *Comunicación escrita* presenta un valor de 0.07, 0.46, 0.43, 0.04 y 0.00 respectivamente, la variable *Razonamiento cuantitativo* valores de 0.47, 0.11, 0.17, 0.06 y 0.19, Lectura crítica 0.5, 0.01, 0.03, 0.46 y 0.01, Competencia ciudadana con valores de 0.6, 0.04, 0.02, 0.06 y 0.28, y de igual forma Ingles que presenta valores de 0.06, 0.61, 0.16, 0.09 y 0.08 respectivamente (Tabla 10 y Figura 11).

Figura 11. Calidad o grado de asociación de todas las variables dentro de las cinco primeras componentes o dim.

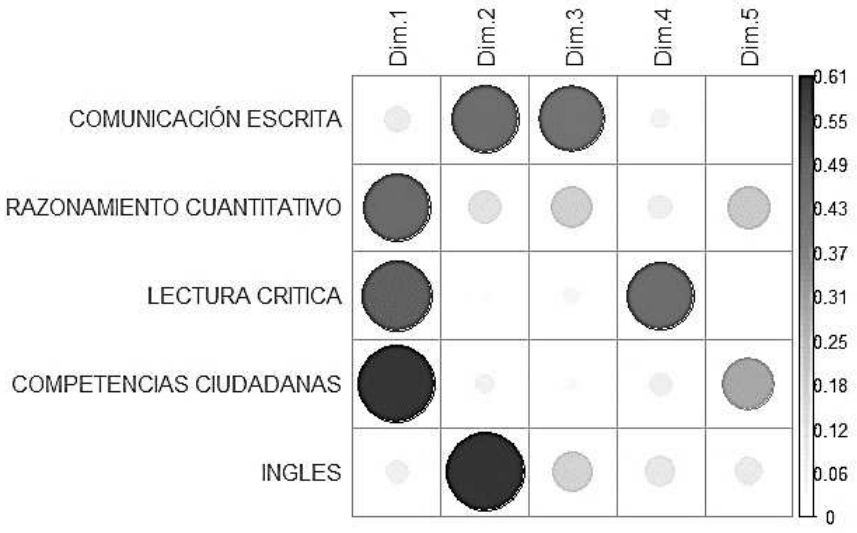


Tabla 10. Grado de asociación(cos2) de cada variable sobre cada componente principal

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
Comunicación escrita	0.07	0.46	0.43	0.04	0.00
Razonamiento cuantitativo	0.47	0.11	0.17	0.06	0.19
Lectura critica	0.50	0.01	0.03	0.46	0.01
Competencia ciudadana	0.60	0.04	0.02	0.06	0.28
Ingles	0.06	0.61	0.16	0.09	0.08

Correlación entre variables (Coord). La Tabla 11 indican el coeficiente de correlación existente entre cada variable y su componente, el componente 1, 2, 3, 4 y 5 con la variable *Comunicación escrita* presenta valores de -0.27, 0.68, 0.66, 0.19 y 0.03 respectivamente, la variable *Razonamiento cuantitativo* valores de 0.68, -0.34, 0.41, -0.25 y 0.43, Lectura crítica 0.71, 0.10, -0.17, 0.67 y 0.09, Competencia ciudadana con valores de 0.78, 0.20, 0.14, -0.24, -0.53, y de igual forma Ingles que presenta valores de 0.24, 0.78, -0.40, -0.30 y 0.28 respectivamente.

Tabla 11. Coeficiente de correlación entre variables por cada componente principal.

Correlación (Coord)	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
Comunicación escrita	-0.27	0.68	0.66	0.19	0.03
Razonamiento cuantitativo	0.68	-0.34	0.41	-0.25	0.43
Lectura critica	0.71	0.10	-0.17	0.67	0.09
Competencia ciudadana	0.78	0.20	0.14	-0.24	-0.53
Ingles	0.24	0.78	-0.40	-0.30	0.28

Nivel de significancia en la relación entre variables. Se puede apreciar en la Figura 12 y en la Tabla 12 que la variable Competencia ciudadana, lectura crítica y razonamiento cuantitativo presentan correlación con un p-valor < 0.05, lo anterior igual para inglés y comunicación escrita dentro de la segunda componente.

Figura 12. Clusters según el grado de relación entre variables para las dos primeras componentes o dim.

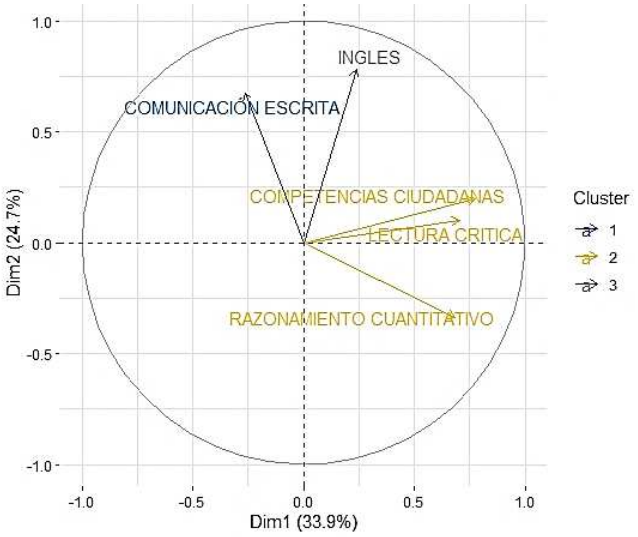


Tabla 12. Correlación y p-valor de las dos primeras componentes

Componente	Variable	Correlación	p-valor
1	Competencia ciudadana	0.78	7.90e-12
	Lectura critica	0.71	3.57e-09
	Razonamiento cuantitativo	0.68	1.60e-08
2	Ingles	0.78	4.52e-12
	Comunicación escrita	0.67	2.87e-08
	Razonamiento cuantitativo	-0.34	1.31e-02

8.2.2. Análisis de Regresión y supuestos estadísticos. En este análisis se generó modelos de regresión con pre y post cumplimiento de los supuestos o condiciones estadísticas, con la salvedad que aparte de las cinco competencias genéricas o predictores se suma la variable dependiente denominada Puntaje Global.

8.2.2.1. Supuestos estadísticos previos a la regresión. En la Tabla 13 se puede observar que la media entre variables va de 136 a 162 puntos, las medianas de 134 a 165 puntos y varianzas desde los 118.0 a 625.57 puntos con una desviación estándar mínima de 11 puntos y máxima de 25.

Tabla 13. Media, mediana, varianza y desviación estándar de cada variable.

	Com. escrita	Raz. cuantitativo	Lect. critica	Comp. ciudadana	Ingles	Global
Media	153	162	145	136	144	148
Mediana	157	165	142	134	145	146
Varianza	619.4	270.14	402.75	625.57	228.13	118.0
Des. estándar	25	16	20	25	15	11

Homocedasticidad. El test de Bartlet, de levene al igual que el de Brown-Forsyth generaron p-valores menores a 0.05, confirmando la ausencia de homogeneidad entre varianzas y rechazando la hipótesis nula (Tabla 14).

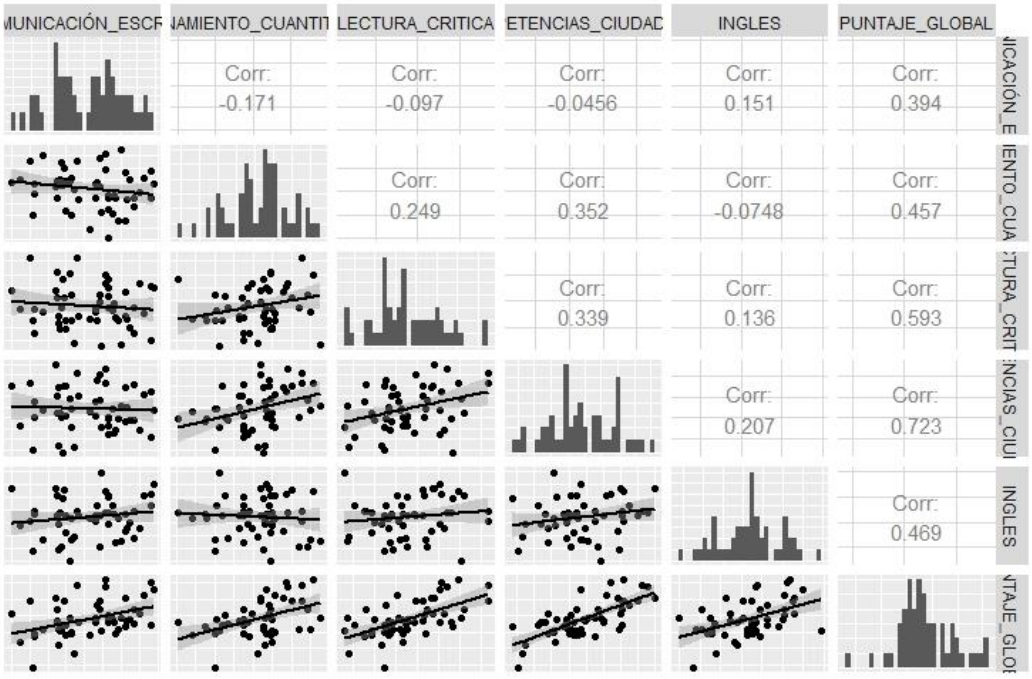
Tabla 14. Test de homocedasticidad

Tipo de test	p-valor
Bartlet	3.25e-09***
Levene	4.59e-09 ***
Brown-Forsyth	4.59e-09***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Multicolinealidad. La matriz de correlaciones y los gráficos de dispersión que se observan en la Figura 13 demuestran que cada par de variables predictores tiene valores cercanos a cero (-0.2 a 0.3), pero tendientes a 1 con el puntaje global(variable dependiente), competencias ciudadanas con 0.7, lectura crítica con 0.6 y razonamiento cuantitativo e inglés con 0.5 en ambas variables y comunicación escrita con 0.4, por lo tanto, se acepta la Hipótesis nula.

Figura 13. Diagramas de dispersión, valores de correlación para cada par de variables y la distribución de cada una de ellas.

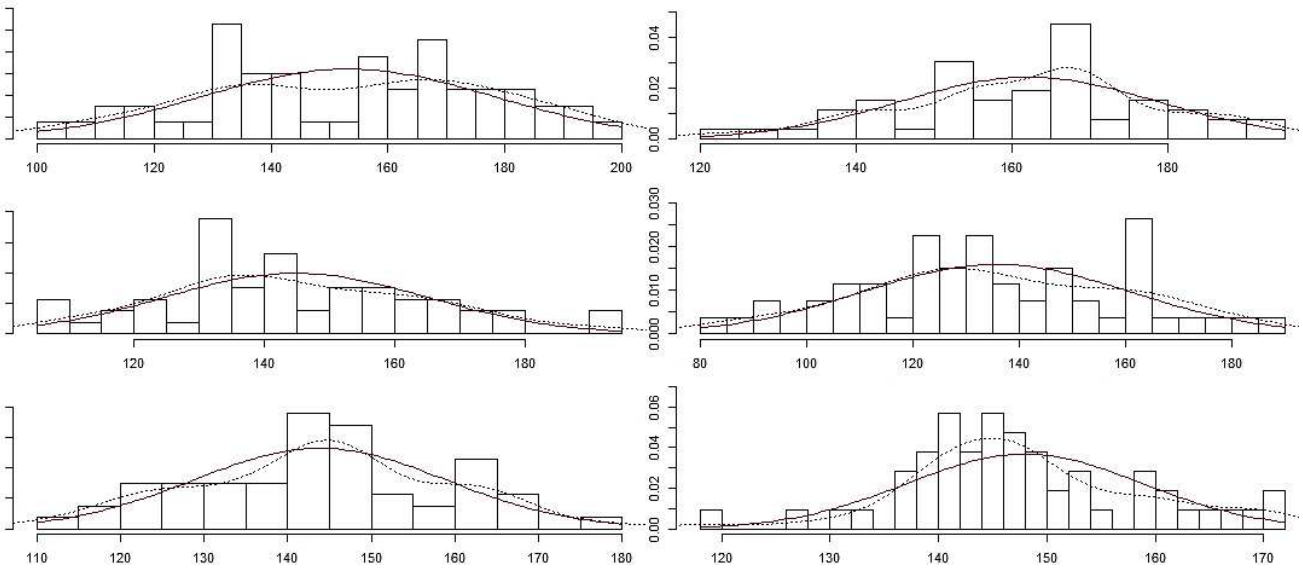


Normalidad. El coeficiente de asimetría y curtosis para la variable dependiente son de 0.16 y 0.23 respectivamente, y los valores estandarizados son 1.15 y 1.70 de signo positivo, los p-valor obtenidos en los diferentes test son superiores a 0.05 como se observan en la Tabla 15, por lo que se acepta la hipótesis nula. Además, en la Figura 14 se puede observar la curva de normalidad en todas las variables.

Tabla 15. Test multivariados para normalidad.

Test	**p-valor**
Asimetria	0.16 (1.15)
Curtosis	0.23 (1.70)
Mardia Skewness	0.99
Mardia kurtosis	0.24
Henze-Zirkler	0.45
Royston	0.58
Doornik-Hansen	0.12
E-statistic	0.84

Figura 14. Curva de normalidad sobra cada competencia genérica y su puntaje global. Comunicación escrita, Razonamiento cuantitativo, Competencia ciudadana, Ingles y el puntaje global respectivamente.



8.2.2.2. Regresión múltiple. Obviando el incumplimiento del supuesto de homocedasticidad se determinó el modelo por mínimos cuadrados, el cual cumplió un R^2 de 0.99 y un p-valor significativo de 2.2e-16 (Tabla 16).

Tabla 16. Coeficientes, error estándar, t value y Pr para cada competencia genérica

	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**
(Intercept)	-0.09	0.631	-0.16	0.875
Com. Escrita	0.201	0.002	125.43	<2e-16***
Raz. cuantitativo	0.202	0.003	77.22	<2e-16***
Lect. critica	0.200	0.002	95.60	<2e-16***
Comp. ciudadanas	0.197	0.002	111.40	<2e-16***
Ingles	0.200	0.003	74.04	<2e-16***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 0.2799

Selección de los mejores predictores. Al tener solo cinco predictores se establece con *Akaike (*AIC) que todos son esenciales para el modelo lineal y este método arroja una sola opción:

Puntaje_global ~ Com. Escrita + Raz.cuantitativo+ Lect. Critica+ Comp.ciudadanas+ Ingles

Tabla 17. Calidad del modelo lineal con el valor AIC=129.34

	DF	**Sum of Sq**	**RSS**	**AIC**
(none)			3.68	-129.34
Ingles	1	429.42	433.11	121.34
Raz. cuantitativo	1	467.20	470.88	125.77
Lect. critica	1	715.93	719.61	148.25
Comp. ciudadanas	1	971.53	975.21	164.35
Com. Escrita	1	1232.40	1236.09	176.92

Intervalo de confianza. Se pudo determinar que el rango de valores en donde se localiza la media poblacional para cada coeficiente esta entre 0.19 a 0.21 (Tabla 18).

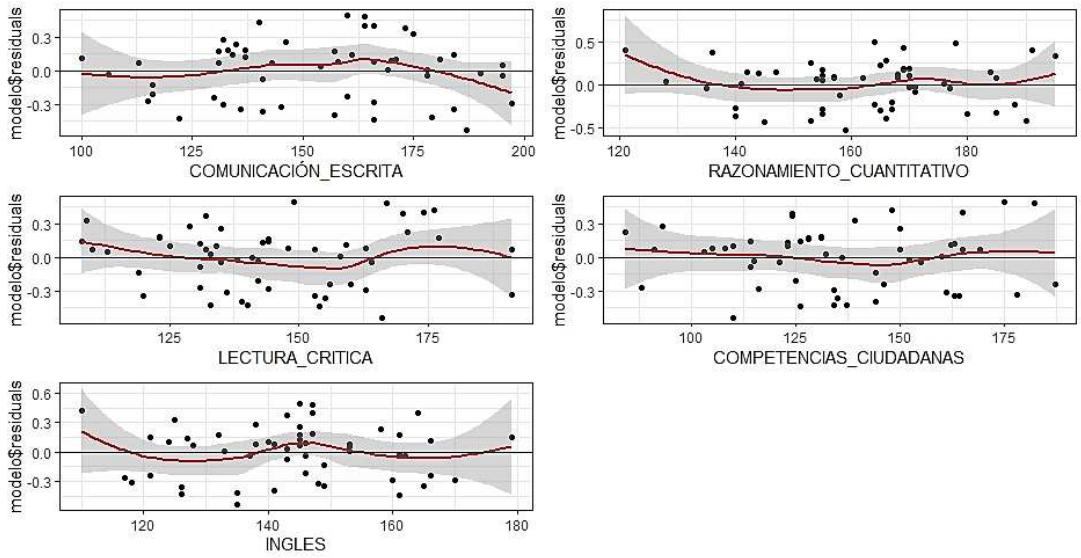
Tabla 18. Intervalos de confianza para cada coeficiente.

	2.5%	**97.5%**
(Intercept)	-1.369	1.169
Com. Escrita	0.198	0.204
Raz. Cuantitativo	0.197	0.207
Lec. Critica	0.196	0.205
Comp. ciudadana	0.193	0.200
Ingles	0.195	0.206

Validación de los supuestos estadísticos del modelo. Se determinaron siete supuestos entre ellos, linealidad entre variables, normalidad de residuos, homocedasticidad de residuos, no multicolinealidad, no autocorrelación, inflación de la varianza y datos atípicos influyentes, todas las condiciones fueron aprobados validando el modelo generado.

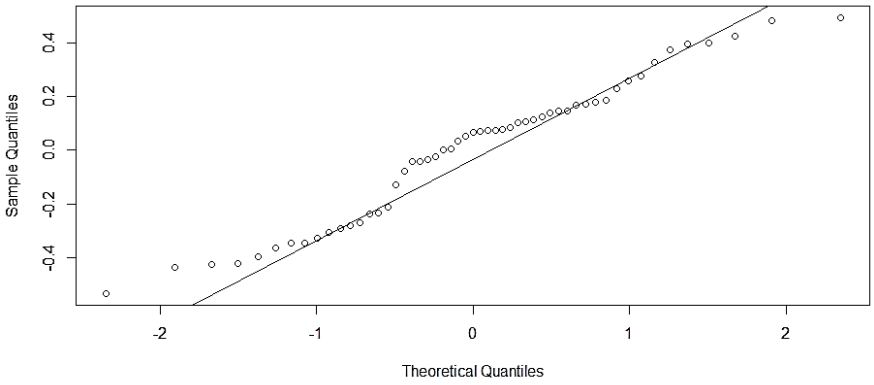
Linealidad entre predictores y variable respuesta. En la Figura 15 se puede observar la linealidad entre todos los predictores, los residuos se distribuyen aleatoriamente en torno a 0 con una variabilidad constante a lo largo del eje X y la posible presencia de algunos datos atípicos.

Figura 15. Relación lineal entre los predictores y variable respuesta.



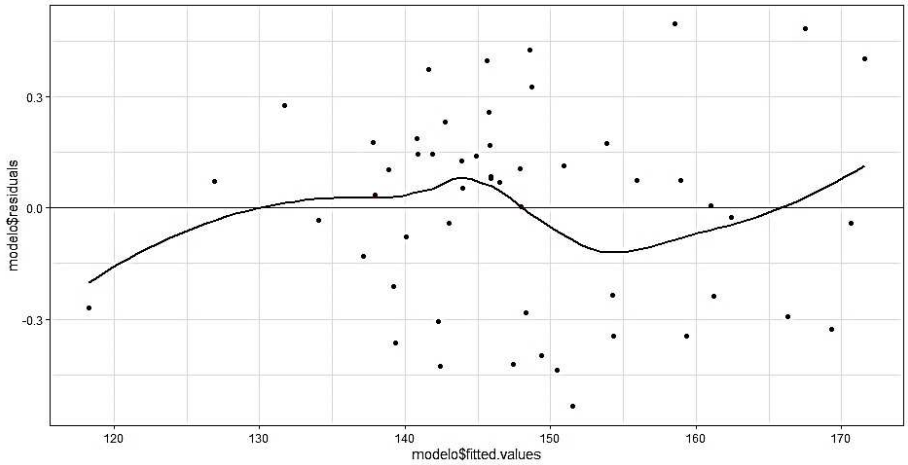
Normalidad de los residuos. En la Figura 16 podemos observar la distribución normal de los residuos al encontrarse lo más alineados entorno a la recta, además el test de Shapiro-Wilk nos arroja un p-valor de 0.093<0.05, por lo que se acepta la hipótesis nula.

Figura 16. Distribución normal de los residuos



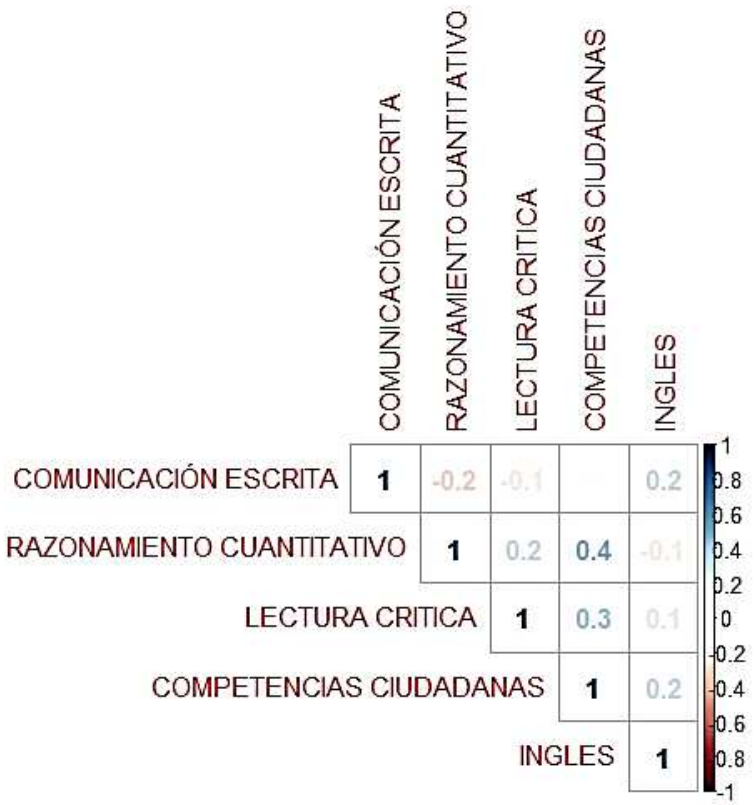
Homocedasticidad u homogeneidad de varianza en los residuos. Mediante la representación de la Figura 17 y el test de Breusch-Pagan se pudo establecer un p-valor es igual a 0.12>0.05, por ende también se acepta la Ho.

Figura 17. Residuales



No multicolinealidad. Mediante la matriz de correlaciones se estableció la no linealidad o dependencia entre predictores, los coeficientes de correlación entre cada par de variables es menor a 1 con valores entre -0.2 a 0.4(Figura 18).

Figura 18. Coeficiente de correlación entre predictores.



Análisis de inflación de varianza (FIV). No hay predictores que muestren una correlación lineal muy alta, de acuerdo a la Tabla 19, todos los valores de FIV están entre 1.1 a 1.3 por lo que al no ser mayor a 10 no se evidencia un grado de colinealidad preocupante.

Tabla 19. Inflación de la varianza para cada coeficiente

Com. escrita	Raz. cuantitativo	Lect. critica	Comp. ciudadana	Ingles
1.06	1.22	1.17	1.30	1.11

Autocorrelación. No hay evidencia de autocorrelación entre los residuos al presentar un p-valor de 0.7 > 0.05 y un coeficiente de Durbin-Watson aproximado a 2, por lo tanto, se acepta la Hipótesis nula.

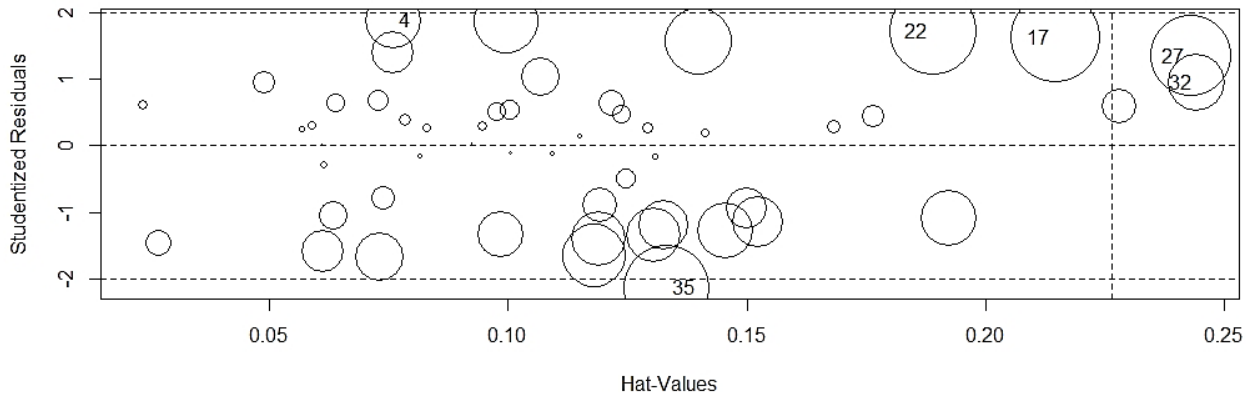
Tabla 20. Coeficientes de autocorrelación

lag	Autocorrelación	D-W statics	p-valor
1	0.038	1.891	0.682

Hipótesis alterna rho!=0

Identificación de posibles valores atípicos influyentes. La Figura 19 no presenta ningún punto atípico influyente.

Figura 19. Valores atípicos, no influyentes en los cinco predictores



Observaciones que son significativamente influyentes en al menos uno de los predictores. Se puede apreciar que la observación 16, 4, 17, 22, 27, 32 y 35 son atípicas no influyentes y según el valor de la distancia Cook en todos es menor a 1 por lo que ninguno es preocupante(Tabla 21).

Tabla 21. Distancias de Cook para datos atípicos.

Dato atípico	4	16	17	22	27	32	35
Distancia Cook	0.05	0.02	0.12	0.11	0.10	0.05	0.11

El mejor modelo resultante con los coeficientes aproximados a dos cifras decimales es:

$$\text{Puntaje global} = 0.20ce + 0.20\text{rc} + 0.20lc + 0.20cc + 0.20i$$

Donde:

ce: Comunicación escrita

rz: Razonamiento cuantitativo

lc: Lectura critica

cc: Competencia ciudadana

i: inglés

8.2.2.3. Regresión cuantílica. Como se aprecia en la Tabla 22 los valores de la mediana entre los predictores y variable tienen una variación entre ellas de ±31 datos.

Tabla 22. Mediana de cada variable dependiente e independiente (puntaje global)

Com. escrita	Raz. cuantitativo	Lect. critica	Comp. ciudadana	Ingles	Global
157	165	142	134	145	146

Se generaron 5 modelos cuantílicos con diferentes cuantiles, entre ellos 0.1, 0.25, 0.5 (mediana), 0.75 y 0.9, todos los modelos presentaron coeficientes aproximados igual a 0.20 sobre cada predictor, por lo que al realizar los ANOVAS se presentaron p-valor entre 0.22 a 0.69, valores mayores a 0.05 (Tabla 23, Tabla 24, Tabla 25, Tabla 26, Tabla 27 y Tabla 28).

Modelo1/Cuantil 0.1:

$$P.global = Com.escrita + Raz.cuantitativo + Lect.critica + Comp.ciudadana + Ingles$$

Tabla 23. Resumen del modelo de regresión para el cuantil 0.1

	Coeficientes	Lower.bd	Upper.bd	Std. Error	T.value	Pr(>|t|)
(intercept)	0.285	-1.870	0.598	0.062	4.615	0.000
Com. Escrita	0.199	0.198	0.203	0.001	303.701	0.000
Raz. Cuantitativo	0.199	0.197	0.215	0.001	197.108	0.000
Lect. Critica	0.198	0.190	0.203	0.000	561.808	0.000
Comp. Ciudadana	0.197	0.191	0.206	0.001	172.711	0.000
Ingles	0.203	0.200	0.209	0.002	260.798	0.000

Modelo 2/Cuantil 0.25:

$$P.global = Com.escrita + Raz.cuantitativo + Lect.critica + Comp.ciudadana + Ingles$$

Tabla 24. Resumen del modelo de regresión para el cuantil 0.25

	Coeficientes	Lower.bd	Upper.bd	Std. Error	T.value	Pr(>|t|)
(intercept)	-0.806	-0.827	1.381	0.949	-0.849	0.400
Com. Escrita	0.202	0.198	0.204	0.002	103.479	0.000
Raz. Cuantitativo	0.205	0.193	0.213	0.004	49.475	0.000
Lect. Critica	0.197	0.188	0.208	0.003	66.633	0.000
Comp. Ciudadana	0.195	0.191	0.202	0.002	82.337	0.000
Ingles	0.204	0.193	0.208	0.003	61.726	0.000

Modelo3/Cuantil 0.5:

$$P.global = Com.escrita + Raz.cuantitativo + Lect.critica + Comp.ciudadana + Ingles$$

Tabla 25. Resumen del modelo de regresión para el cuantil 0.5

	Coeficientes	**Lower.bd**	**Upper.bd**	**Std. Error**	**T.value**	**Pr(>|t|)**
(intercept)	0.419	-2.066	1.081	0.885	0.473	0.638
Com. Escrita	0.199	0.199	0.203	0.002	86.588	0.000
Raz. Cuantitativo	0.203	0.195	0.208	0.004	52.428	0.000
Lect. Critica	0.199	0.196	0.204	0.003	64.767	0.000
Comp. Ciudadana	0.195	0.194	0.201	0.002	80.935	0.000
Ingles	0.199	0.189	0.205	0.004	53.241	0.000

Modelo 4/Cuantil 0.75:

$$P.global = Com.escrita + Raz.cuantitativo + Lect.critica + Comp.ciudadana + Ingles$$

Tabla 26. Resumen del modelo de regresión para el cuantil 0.75

	Coeficientes	**Lower.bd**	**Upper.bd**	**Std. Error**	**T.value**	**Pr(>|t|)**
(intercept)	-1.245	-1.450	1.896	0.874	-1.378	0.175
Com. Escrita	0.203	0.198	0.207	0.001	157.232	0.000
Raz. Cuantitativo	0.206	0.196	0.206	0.003	72.074	0.000
Lect. Critica	0.201	0.199	0.205	0.001	139.404	0.000
Comp. Ciudadana	0.198	0.195	0.202	0.002	128.118	0.000
Ingles	0.202	0.195	0.204	0.003	74.838	0.000

Modelo 5/Cuantil 0.9:

$$P.global = Com.escrita + Raz.cuantitativo + Lect.critica + Comp.ciudadana + Ingles$$

Tabla 27. Resumen del modelo de regresión para el cuantil 0.9.

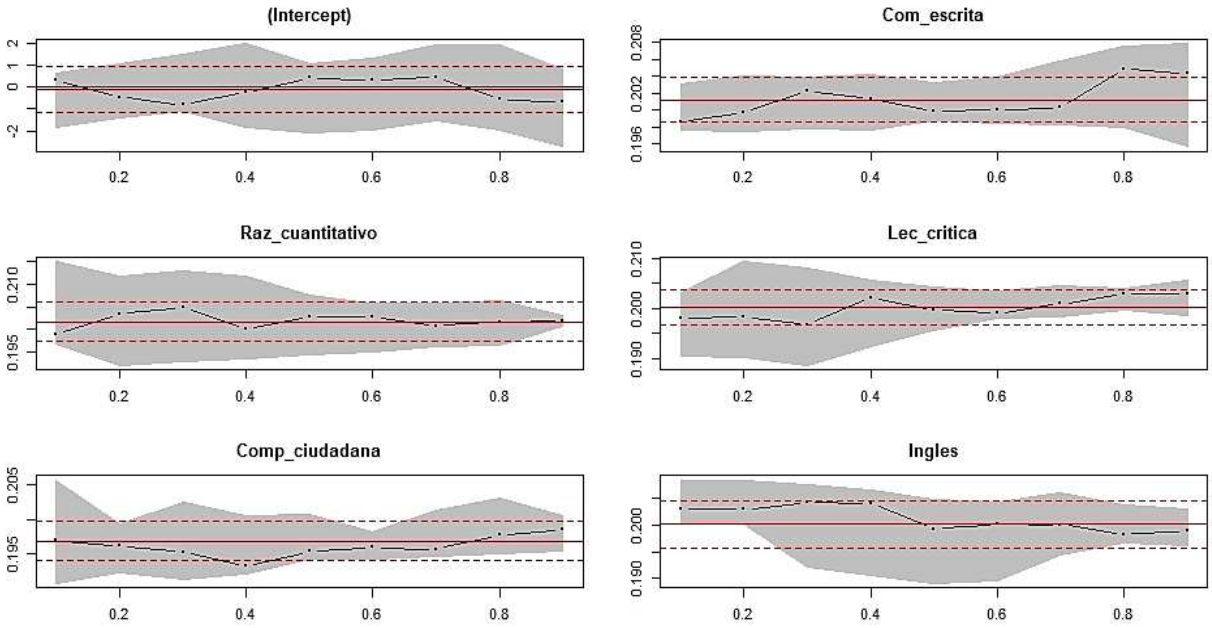
	Coeficientes	**Lower.bd**	**Upper.bd**	**Std. Error**	**T.value**	**Pr(>|t|)**
(intercept)	0.686	-2.736	0.799	1.047	-0.656	0.515
Com. Escrita	0.204	0.196	0.208	0.004	52.556	0.000
Raz. Cuantitativo	0.202	0.201	0.203	0.005	43.220	0.000
Lect. Critica	0.203	0.199	0.206	0.002	83.518	0.000
Comp. Ciudadana	0.199	0.196	0.201	0.002	87.760	0.000
Ingles	0.199	0.196	0.203	0.002	97.896	0.000

Tabla 28. ANOVA entre modelos cuantiles.

	Mod01/25	Mod25/50	Mod50/75	Mod75/90	Mod01/50	Mod90/50
F-valor	1.36	0.108	1.26	0.67	1.61	1.43
P-valor	0.25	0.37	0.28	0.64	0.69	0.22

En la Figura 20 se puede apreciar los intervalos de confianza representados por las lineas punteadas de rojo, la recta de regresion de cada cuantil por la linea negra dentro del intervalo y la linea continua roja que representa la regresion normal de cada variable independiente y su intercepto, el predictor Competencia ciudadana y Comunicación escrita en el cuantil 0.8 y 0.4 respectivamente sobrepasan el intervalo minimamente.

Figura 20. Coeficientes de la regresión de cada cuantil sobre la pendiente en cada variable, su recta de regresión (roja continua) e intervalos de confianza.



9. DISCUSIÓN DE RESULTADOS

En el análisis descriptivo podemos establecer que en general durante los cuatro años existió una mayor cantidad de estudiantes de género femenino con un porcentaje del 55% con respecto al masculino, este género predomina también en los reportes nacionales según el ICFES (2018a) y estudios en otras áreas como en el de Orjuela, (2013). La mayoría presentaron edades entre los 22 a 26 años ya que se encontraban cursando noveno semestre y por lo general mantienen este rango de edades, a excepción de algunos pocos que reingresan al programa años más tarde o se nivelan con el pensum académico. Además, el 80% de la población es de estrato bajo, I y II, característico de la mayoría de estudiantes pertenecientes a Universidades públicas, resultados que también se observan en los documentos de análisis de resultados generados por el ICFES cuando caracterizan a los estudiantes de universidades públicas y privadas por competencia genérica (ICFES, 2018b).

La cantidad de estudiantes que presentaron el examen de estado varia cada año dependiendo de aquellos que lograron cumplir con el 75% de los créditos académicos del programa de pregrado, durante el 2016 se presentó la menor cantidad de ellos con tan solo 8 registros y en el 2019 la mayor cantidad con 21 registros, esta baja cantidad de estudiantes no contribuye junto a otros programas a que Nariño sea considerado por el ICFES como uno de los municipios con el mayor número de evaluados a nivel nacional(ICFES, 2018a), algunos valores como el de desviación estándar variaron en este estudio según la cantidad de datos existentes por año en cada variable o competencia genérica.

Además, sin tener en cuenta datos atípicos podemos cuestionar los valores de las medias, medianas, desviaciones estándar y rango de puntajes en cada año.

Por ejemplo, con el valor mínimo y máximo entre puntajes, se pudo determinar qué el año 2016 entre todas las variables mantuvo el menor rango, entre 121 a 195 puntos, esto se pudo deber a la baja cantidad de estudiantes que presentaron la prueba durante el año en mención, por lo tanto, la variación fue baja entre los puntos acertados de los 8 estudiantes.

La variable Competencia ciudadana, Comunicación escrita y Razonamiento cuantitativo presentaron la mayor media en el año 2016, en cambio, la variable Ingles fue la única que presento su media más alta en el año 2019, lo anterior debido a que este año conto con estudiantes que generaron un alto puntaje en su prueba de inglés debido a su destreza y experiencia en la segunda lengua, Rstudio no los reporto como atípicos.

Las mayores y menores desviaciones estándar se presentaron en todos los años, el 2016 se caracterizó por contener en las variables de Competencia ciudadana y Lectura critica la más alta desviación con ±30 puntos, mientras que los años 2017, 2018 y 2019 en las competencias de Razonamiento cuantitativo, Ingles y Lectura crítica respetivamente los menores valores con ±17puntos.

En el análisis de componentes principales se pudo establecer previamente el cumplimiento del supuesto de normalidad y generar los pesos asignados a cada componente, donde se puede observar que las dos primeras recogen la mayor información correspondiente a las competencias genéricas y pueden explicar casi el 60% de la varianza total.

Las variables presentaron bajos coeficientes de correlación dentro del supuesto de multicolinealidad, mas sin embargo en el análisis de componentes principales, la variable Comunicación escrita e inglés demuestran que tiene un alto grado de asociación(Cos2), contribución (Contrib.) y correlación (Coord.) sobre la segunda componente, mas no son variables

predictores relevantes en la primera, mientras que Razonamiento cuantitativo, Lectura crítica y Competencia ciudadana si lo son. Lo anterior se pudo confirmar con el circulo de correlaciones y los coeficientes de correlación que generaron valores superiores a 0.7. Al igual que en el trabajo de Martínez & Mendoza, (2016) se apreció la relación entre variables dentro de cada componente, y es normal ya que existen según el ICFES, (2017) subcompetencias en cada módulo, por ejemplo, Razonamiento cuantitativo implica la interpretación y argumentación de sus resultados y es apenas normal que esté relacionado con lectura crítica y Competencia ciudadana mas no Inglés o Comunicación escrita, los cuales son evaluados por niveles de desempeño.

Todos los supuestos estadísticos se cumplieron a excepción de la homogeneidad de varianzas, mas sin embargo la media y mediana tuvo poca variación entre competencias, similar a los resultados presentados por Rodriguez Manrique et al., (2018) en su estudio.

En el análisis de regresión múltiple, el ajuste de bondad generado fue un R^2 de 0.99, lo que implica que más del 90% de la variabilidad del puntaje global esta explicada en el modelo, similar a lo ocurrido con Marquín (2017), con un p-valor significativo de 2.2e-16 se puede decir que al menos uno de los coeficientes parciales es diferente de cero y que las variables explicativas tienen asociación con la variable dependiente o criterio (Rodriguez Ayán, 2007), que aquella parte del promedio global o variable que dejamos por explicar corresponde a 0.28 como error residual estándar.

Las regresiones cuantílicas muestran que a partir del primer cuantil al 0.9 los coeficientes crecen en Comunicación escrita, Lectura crítica y Competencia ciudadana, mientras que existe un decrecimiento leve en Inglés y el Intercepto, solo Razonamiento cuantitativo presentó un comportamiento aparentemente constante, todos estos cambios se presentaron dentro del intervalo de confianza establecido para cada competencia y cuantil, que en general fueron valores

aproximados entre 0.19 a 0.21, solo Comunicación escrita en el cuantil 0.8 y Competencia ciudadana en el cuantil 0.4 presentaron coeficientes mínimamente establecidos por encima y debajo del intervalo, estos cambios y similares intervalos se presentaron en el estudio de Pérez-Pulido et al., (2016) donde el intercepto si presento crecimiento a medida que aumento el cuantil en cada regresión.

10. CONCLUSIONES Y RECOMENDACIONES

Las 5 variables independientes son indispensables para predecir el puntaje global, cada una con diferente grado de contribución, calidad de asociación y peso sobre las combinaciones lineales de las dos primeras componentes.

Todos los modelos son eficaces en la predicción del puntaje global, ya que el modelo por regresión múltiple contó con los mismos coeficientes y variables predictores que los modelos por regresión cuantílica sin diferencia significativa entre ellos, por lo tanto, se confirma la confiabilidad del mismo con un 99% de capacidad predictiva.

La media total estuvo por debajo de la mitad (148 puntos), presentándose los más bajos promedios en Lectura crítica, Competencia ciudadana e Inglés, los más altos puntajes en aquellos años donde existió el mayor número de estudiantes, (Año 2018 y 2019) con puntajes entre 185 y 204 puntos.

La variación entre medias y medianas no influyo en el cumplimiento de la normalidad y multicolinealidad entre las variables, sin embargo, se recomienda realizar una previa transformación de datos para suplir el supuesto de homocedasticidad entre predictores y así poder realizar un análisis factorial de las variables.

Se recomienda mejorar el rendimiento del estudiante en las pruebas SABER PRO fortaleciendo aquellas competencias genéricas que no presentaron Contribución mayor a 50 sobre las dos primeras componentes, como lo fue en Lectura crítica, Comunicación escrita y Competencia ciudadana, de esta forma la variabilidad retenida en las dos componentes iniciales seria mayor al 60%.

Se espera que las pruebas SABER PRO para el programa de Ingeniería en Producción Acuícola puedan evaluar no solo las habilidades y conocimiento básicos mediante competencias genéricas sino también las habilidades y conocimientos específicos del programa académico que cualquier recién egresado debe tener.

Es necesario establecer estudios con la existencia de nuevas variables independientes en los modelos de regresión, como el estrato socioeconómico, edad, genero, puntaje de las pruebas de estado saber 11° y saber pro en el programa de pregrado, de esta forma se podrá establecer las áreas donde el estudiante presenta falencias desde un inicio de su carrera y poder fortalecerlas cumpliendo el criterio de alta calidad en la educación superior.

Se recomienda realizar un tratamiento estadístico a todos los indicadores académicos, entre ellos al nivel de deserción, al número de estudiantes egresados graduados y no graduados, al desempeño por periodos y a los puntajes de las pruebas de estado facilitando la toma de decisiones en pro de la calidad académica de los futuros profesionales.

11. REFERENCIAS

(ICFES), I. C. para la E. de la E. (2017). *Guia de Orientación. Saber Pro Módulos de Competencias Genéricas.* (p. 90).

Aldas Manzano, J., & Uriel Jimenez, E. (2017). Introducción. In S. A. Paraninfo (Ed.), *Análisis multivariantes aplicado con R* (2.a, pp. 19–28).

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. In *Journal of the American Statistical Association* (Vol. 69, Issue 346). https://doi.org/10.1080/01621459.1974.10482955

Castaño, E. (2017). Introducción al análisis de datos multivariados en ciencias sociales. *XII Seminario de Estadística Aplicada III Escuela de Verano VII Coloquio Regional de Estadística*, 1–243.

Cifuentes Garzón, J. A. (2013). *Predicción del resultado en la prueba SABER PRO para Economía a partir de la información disponible en el proceso de admisión*. Pontifica Universidad Javeriana.

Cook, R. (1977). Detection of Influential Observation in Linear Regression. *Rechnometrics*, *19*(1), 1–494. https://doi.org/10.1007/978-3-319-55252-1

Correa, J. C., Iral, R., & Rojas, L. (2006, June). Estudio de potencia de pruebas de homogeneidad de varianza. *Revista Colombiana de Estadistica*, *29*(1), 57–76.

Corredor Rivera, D. A., Gómez, L. F., & Rios Pineda, W. (2017). *Estudio de la incidencia actitudinal de los estudiantes en pruebas de matemáticas tipo ICFES : Una aproximación semiparamétrica.* Universidad Santo Tomas.

DANE. (2008). *Manual de indicadores* (pp. 1–30). Departamento Administrativo Nacional de Estadística.

Das, K., Krzywinski, M., & Altman, N. (2019). Quantile regression. *Nature Methods*, *16*(6), 451–452. https://doi.org/10.1038/s41592-019-0406-y

Doornik, J., & Hansen, H. (1994). An Omnibus Test for Univariate and Multivariate Normality. *Nuffield College*, 1–16.

Espinel F., C., González A., T., Bruno C., A., & Pinto S., J. (2009). Las gráficas estadísticas. In L. Serrano (Ed.), *Tendencias actuales de la investigación en educación estocástica* (pp. 133–155). Universidad de Granada.

Everitt, B., & Hothorn, T. (2011). *(Use R) An Introduction to Applied Multivariate Analysis with R*. (R. Gentleman, G. Parmigiani, & K. Hornik (eds.)). Springer US.

Fernández Rodriguez, L. (2015). *Análisis De La Dependencia De Las Variables Macroeconómicas En El Rating Soberano Y Prima De Riesgo Española*. Universidad Pontificia ICAI&ICADE.

Flury, B., & Riedwyl, H. (1981). Graphical representation of multivariate data by means of asymmetrical faces. *Journal of the American Statistical Association*, *76*(376), 757–765. https://doi.org/10.1080/01621459.1981.10477718

Ford, C. (2015). *Getting Started with Quantile Regression*. Research Data Services +Sciencies. http://data.library.virginia.edu/getting-started-with-quantile-regression/

Galvis A., L. A. (2012). Informalidad laboral en las areas urbanas de Colombia. *Coyuntura Economica: Investigacion Economica y Social*, *42*(1), 15–51.

http://www.fedesarrollo.org.co/publicaciones/publicaciones-periodicas/coyuntura-economica/ediciones-anteriores/%5Cnhttp://search.ebscohost.com/login.aspx?direct=true&db=ecn&AN=1387669&site=ehost-live&scope=site

George, D., & Mallery, P. (2001). *SPSS for Windows step by step: a simple guide and reference, 10.0 update*. Allyn and Bacon.

Icfes. (2017). *Información general del examen saber 11.°*. © 2016 Instituto Colombiano Para La Evaluación de La Educación ICFES. http://www2.icfes.gov.co/estudiantes-y-padres/saber-11-estudiantes/informacion-general-del-examen

ICFES. (2018a). *Informe nacional Saber Pro 2016 - 2018 1*.

ICFES. (2018b). *Las diferencias y el contexto en los módulos específicos. Documento de análisis de resultados Examen SABER PRO 2017. Módulo de Anális Económico*.

J.F., H., R.E., A., R.L., T., & W. Black. (1999). *Análisis Multivariante*. Prentice-Hall.

Joaquín, A. R. (2017). Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. *RStudio Pubs*. https://www.cienciadedatos.net/documentos/35_principal_component_analysis#t-sne%0Ahttps://rstudio-pubs-static.s3.amazonaws.com/287787_79ce0f01d0a941d8a38caffdb209922b.html#introducción

Kachigan, S. K. (1991). Multivariate statistical analysis: A conceptual introduction. In *New York Radius Press* (2nd ed., Vol. 2nd).

Kassambara, A. (2017). Practical Guide to Principal Component Methods in R. In *Multivariate*

Analysis II (p. 170). STHDA.

López, C. (2004). *Técnicas de Análisis Multivariante de Datos. Aplicaciones con spss.* Pearson Education.

Marquín, M. J. (2017). *Predicción del rendimiento académico mediante técnicas del análisis multivariado en la asignatura de Álgebra Lineal* [Universidad Tecnologíca de Pereira UTP]. http://repositorio.utp.edu.co/dspace/bitstream/handle/11059/8356/37126M357.pdf?sequence=1

Martínez Arias, M. R. (1999). *El análisis multivariante en la investigación científica.* La Muralla.

Martínez, C. Y., & Mendoza, L. F. (2016). Análisis de los Resultados de la Evaluación en Competencias Genéricas de las Pruebas SABER PRO 2014 en Programas de Licenciatura en el Área de las Ciencias Naturales y Educación Ambiental de Cuatro Universidades del pais. Universidad Distrital Fancisco José de Caldas.

Montanero Fernández, J. (2018). Manual Abreviado de Estadistica Multivariante (pp. 1–104). Universidad de Extremadura.

Ordoñez-Castaño, I. A., & Sanabria-Domínguez, J. A. (2014). Retornos de la educación para los trabajadores formales e informales en Cali: Una aproximación con regresiones cuantílicas y splines lineales. *Estramado*, *10*, 12–22.

Orjuela, J. (2013). Análisis del Desempeño Estudiantil en las Pruebas de Estado para Educación Media en Colombia mediante Modelos Jerárquicos Lineales. *Ingeniería*, *18*(2), 54–67.

https://doi.org/10.14483/udistrital.jour.reving.2013.2.a04

Parra Rodríguez, F. (2016). *Curso de Estadística con R. Documentos técnicos.* (2).

Peña-Méndez, D. P. (2014). *Análisis de componentes principales en la estimación de indices de empoderamiento en mujeres de Colombia*. Universidad de Granada.

Pérez-Pulido, M. O., Aguilar-Galvis, F., Orlandoni-Merli, G., & Ramoni-Perazzi, J. (2016). Análisis estadístico de los resultados de las pruebas de estado para el ingreso a la educación superior en la Universidad de Santander, Colombia. *Revista Científica*, *27*, 328–339. https://doi.org/10.14483/udistrital.jour.rc.2016.27.a3

Pérez, E. R., & Medrano, L. (2010). Análisis Factorial Exploratorio: Bases Conceptuales y Metodológicas. *Revista Argentina de Ciencias Del Comportamiento*, *2*, 58–66. www.psyche.unc.edu.ar/racc

Porras Ceron, J. C. (2016, September). Comparación de Pruebas de Normalidad Multivariada. *A. Científicos*, *77 (2)*(Zimmerman 2011), 141–146. http://simposioestadistica.unal.edu.co/fileadmin/content/eventos/simposioestadistica/documentos/memorias/Memorias_2016/Posters/16._Pruebas_Normalidad_Cortes_Rave___Hernandez.pdf

Rodriguez Ayán, M. N. (2007). *Análisis multivariado del desempeño académico de estudiantes universitarios de Química*. Universidad Autónoma de Madrid.

Rodríguez, M., & Catalá, R. (2001). Análisis de Regresión Múltiple. In U. de Alicante (Ed.), *Estadística Informática: casos y ejemplos con el SPSS* (pp. 3–17). http://rua.ua.es/dspace/bitstream/10045/8143/1/Regresion MUTIPLE.pdf

Rodriguez Manrique, J. A., Ruiz Escorcia, R. R., & Cohen Manrique, C. S. (2018). *Análisis Multivariado Aplicado a la Evaluación de Competencias Saber-Pro en el Departamento de Sucre, Colombia* (No. 16). https://doi.org/10.18687/LACCEI2018.1.1.16

Salvador-Figuereas, M. (2000). *Introducción al análisis multivariante*. http://www.5campus.com/leccion/anamul

SNIES. (n.d.). *Estadísticas*. Sistema Nacional de Información de La Educación Superior. Retrieved March 20, 2019, from https://www.mineducacion.gov.co/sistemasinfo/Informacion-a-la-mano/212400:Estadisticas

SPADIES. (n.d.). *Glosario*. Sistema Para La Prevención de La Deserción de La Educación Superior. Retrieved March 19, 2019, from https://www.mineducacion.gov.co/sistemasdeinformacion/1735/w3-article-254707.html