

**Analítica de Datos para la Predicción del Desempeño en el Examen  
SABER11**

**Aron Rene Forero Africano**

**UNIVERSIDAD INDUSTRIAL DE SANTANDER  
FACULTAD DE INGENIERIAS FISICOMECHANICAS  
ESCUELA DE INGENIERIA DE SISTEMAS E INFORMÁTICA  
BUCARAMANGA  
2018**

**Analítica de Datos para la Predicción del Desempeño en el Examen  
SABER11**

**Aron Rene Forero Africano**

**TRABAJO DE GRADO PARA OPTAR POR EL TITULO DE  
INGENIERO DE SISTEMAS E INFORMATICA**

Director  
**GABRIEL RODRIGO PEDRAZA FERREIRA Ph.D.**  
Profesor EISI  
Codirector  
**RAUL RAMOS POLLAN Ph.D**  
Profesor UdeA

UNIVERSIDAD INDUSTRIAL DE SANTANDER  
FACULTAD DE INGENIERIAS FISICOMECHANICAS  
ESCUELA DE INGENIERIA DE SISTEMAS E INFORMATICA  
BUCARAMANGA  
2018

## ESPACIO PARA NOTA



UNIVERSIDAD INDUSTRIAL DE SANTANDER  
SISTEMA DE TRABAJOS DE GRADO  
ADMINISTRACIÓN DE TRABAJOS DE GRADO

Fecha Impresión:  
23 agosto 2018

Pág 1 de 1

Codigo:	18972	Fecha Presentacion:	15-dic-2017
<b>Título:</b> Analítica de datos para la predicción del desempeño en el examen SABER11.			
Nota Proyecto:	5.0	Fecha Registro Nota:	23-agosto-2018
Estado:	APROBADO		
Tipo Trabajo:	INVESTIGACION		
<b>Estudiantes</b>			
Código	Nombre	Programa Académico	
2130071	FORERO AFRICANO ARON RENE	11-INGENIERIA DE SISTEMAS	
<b>Directores</b>			
Documento	Nombre	Clase	Firma
C-91075781	GABRIEL RODRIGO PEDRAZA FERREIRA	DIRECTOR	
<b>Calificadores</b>			
Documento	Nombre	Firma	
C-13929892	FABIO MARTINEZ CARRILLO		
C-13812725	HECTOR NIÑO QUIÑONEZ		

## ESPACIO PARA CARTA AUTORIZACIÓN USO DE DATOS



### ENTREGA DE TRABAJOS DE GRADO, TRABAJOS DE INVESTIGACION O TESIS Y AUTORIZACIÓN DE SU USO A FAVOR DE LA UIS

Yo, Aron Rene Forero Africano, mayor de edad, vecino de Bucaramanga, identificado con la Cédula de Ciudadanía No. 1098783693 de Bucaramanga, actuando en nombre propio, en mi calidad de autor del trabajo de grado, del trabajo de investigación, o de la tesis denominada(o):

Analítica de datos para la predicción del desempeño en el examen SABER11,

hago entrega del ejemplar respectivo y de sus anexos de ser el caso, en formato digital o electrónico (CD o DVD) y autorizo a LA UNIVERSIDAD INDUSTRIAL DE SANTANDER, para que en los términos establecidos en la Ley 23 de 1982, Ley 44 de 1993, decisión Andina 351 de 1993, Decreto 460 de 1995 y demás normas generales sobre la materia, utilice y use en todas sus formas, los derechos patrimoniales de reproducción, comunicación pública, transformación y distribución (alquiler, préstamo público e importación) que me corresponden como creador de la obra objeto del presente documento. PARÁGRAFO: La presente autorización se hace extensiva no sólo a las facultades y derechos de uso sobre la obra en formato o soporte material, sino también para formato virtual, electrónico, digital, óptico, uso en red, Internet, extranet, intranet, etc., y en general para cualquier formato conocido o por conocer.

EL AUTOR – ESTUDIANTE, manifiesta que la obra objeto de la presente autorización es original y la realizó sin violar o usurpar derechos de autor de terceros, por lo tanto la obra es de su exclusiva autoría y detenta la titularidad sobre la misma. PARÁGRAFO: En caso de presentarse cualquier reclamación o acción por parte de un tercero en cuanto a los derechos de autor sobre la obra en cuestión, EL AUTOR / ESTUDIANTE, asumirá toda la responsabilidad, y saldrá en defensa de los derechos aquí autorizados; para todos los efectos la Universidad actúa como un tercero de buena fe.

Para constancia se firma el presente documento en dos (02) ejemplares del mismo valor y tenor, en Bucaramanga, a los 27 días del mes de Agosto de Dos Mil Dieciocho 2018.

**EL AUTOR / ESTUDIANTE:**

(Firma)

**Nombre** Aron Rene Forero Africano

## AGRADECIMIENTOS

El autor expresa su agradecimiento:

Al grupo de investigación de Computo Avanzado y a Gran Escala (CAGE) por acogerme en sus instalaciones y brindarme apoyo con el equipo de computo suficiente para el desarrollo del proyecto y los espacios apropiados para un ambiente de estudio productivo. Para el profesor Gabriel Rodrigo Pedraza por su apoyo como director frente a la escuela de ingeniería de sistemas e informática y en especial al profesor Raul Ramos Pollan por su tutoria, paciencia, dedicación y enseñanzas integrales que me brindo. A la Universidad Industrial de Santander, a la escuela de ingeniería de sistemas e informática y a todos aquellos profesores que se dedican verdaderamente a brindar una formación completa a los estudiantes, con sencillez y seguridad.

A mi novia Laura, y todos mis amigos y compañeros de la universidad, que me ayudaron antes y durante el proceso de desarrollo de este proyecto, y una mención especial para mis amigos del grupo de investigación CAGE, del semillero MACV, del grupo de investigación y desarrollo GID CONUSS y a mis amigos de *VOLLEYBALL* por su gran apoyo y contribución a mi crecimiento personal y profesional.

Finalmente, agradecer a mi familia, a mis hermanos, y a mis padres que con desinterés me ayudaron durante toda mi carrera universitaria y que de seguro yo les compensare en un futuro. Y por ultimo pero no menos importante a Dios por darme la oportunidad y poner a todas estas personas en mi camino.

## CONTENIDO

	pág.
<b>INTRODUCCION . . . . .</b>	<b>13</b>
<b>1. PLANTEAMIENTO Y JUSTIFICACION DEL PROBLEMA . . . . .</b>	<b>15</b>
<b>2. OBJETIVOS . . . . .</b>	<b>16</b>
2.1. OBJETIVO GENERAL . . . . .	16
2.2. OBJETIVOS ESPECIFICOS . . . . .	16
<b>3. MARCO CONCEPTUAL . . . . .</b>	<b>17</b>
3.1. EXPLORACION Y PRE-PROCESADO DE DATOS . . . . .	17
3.1.1. Exploración o análisis exploratorio de datos: . . . . .	17
3.1.2. Pre-procesado de datos: . . . . .	17
3.2. <i>MACHINE LEARNING</i> . . . . .	20
3.2.1. Regresión en <i>Machine Learning</i> . . . . .	21
3.2.2. Regresión Lineal: . . . . .	22
3.2.3. Árbol de decisión de regresión: . . . . .	22
3.2.4. Bosques aleatorios de regresión: . . . . .	24
3.2.5. Métrica de desempeño: . . . . .	24
3.2.6. Metodos de validación: . . . . .	24
<b>4. METODOLOGIA DESARROLLADA . . . . .</b>	<b>26</b>
4.1. CATALOGO DE DATOS Y ANALISIS EXPLORATORIO . . . . .	26
4.1.1. Catalogo de datos: . . . . .	26
4.1.2. Analisis exploratorio . . . . .	28
4.2. PRE-PROCESADO . . . . .	28
4.3. ANÁLISIS GENERAL Y PROFUNDO . . . . .	30
4.3.1. Tratamiento de la multicolinealidad: . . . . .	31
4.3.2. Visualización de Datos . . . . .	32
4.3.3. Análisis de los segmentos . . . . .	33

4.4. CONSTRUCCION Y SELECCION DE ALGORITMOS DE PREDICCIÓN	35
4.4.1. Tratamiento de datos . . . . .	38
4.4.2. Pruebas Planteadas . . . . .	38
4.4.3. Algoritmos seleccionados . . . . .	53
4.5. EXPERIMENTOS FINALES . . . . .	54
<b>5. EVALUACION DE RESULTADOS . . . . .</b>	<b>66</b>
<b>6. CONCLUSIONES . . . . .</b>	<b>68</b>
<b>7. PERSPECTIVAS . . . . .</b>	<b>69</b>
<b>BIBLIOGRAFIA . . . . .</b>	<b>69</b>

## LISTA DE FIGURAS

	pág.
1. Esquema de una predicción supervisada . . . . .	21
2. Regresion Lineal Simple . . . . .	23
3. Ejemplo de Árbol de Decisión de regresión . . . . .	23
4. Cantidad de Columnas por archivo . . . . .	28
5. Pesos en Megabytes [MB] de los archivos . . . . .	29
6. Datos sucios de una variable . . . . .	30
7. Matriz de correlación . . . . .	31
8. Visualizacion de <i>targets</i> . . . . .	32
9. Visualizacion Matematicas segmentada por ingresos familiares . . . . .	34
10. Visualizacion del Promedio de las Materias por ingreso familiar. Antes del año 2014 . . . . .	34
11. Visualización Matemáticas en segmentos geográficos antes del cambio del 2014	35
12. Visualización Matemáticas en segmentos geográficos después del cambio del 2014 . . . . .	36
13. Visualizacion del Promedio de las Materias por segmento geográfico. Antes del año 2014 . . . . .	36
14. Evidencia del tiempo de ejecución del SVR . . . . .	37
15. Feature Importances. Año 2000 y año 2014 . . . . .	39
16. Curva de aprendizaje Regresión Lineal año 2000 . . . . .	42
17. Curva de aprendizaje Regresión Lineal año 2014 . . . . .	44
18. Analisis de profundidad Y Curva de aprendizaje árbol de decisión año 2000 .	46
19. Analisis de profundidad Y Curva de aprendizaje árbol de decisión año 2014 .	48
20. Analisis de profundidad bosque aleatorio año 2000 . . . . .	50
21. Curva de aprendizaje bosque aleatorio año 2000 . . . . .	51
22. Analisis de profundidad Y Curva de aprendizaje bosque aleatorio año 2014 .	53
23. <i>Scores</i> Análisis individual Regresión lineal . . . . .	56
24. <i>Scores</i> Análisis individual Árbol de Decisión . . . . .	57
25. <i>Scores</i> Análisis individual Bosque Aleatorio . . . . .	58

26. <i>Scores</i> Análisis individual Regresión lineal despues del 2014 . . . . .	59
27. <i>Scores</i> Análisis individual Árbol de Decisión despues del 2014 . . . . .	60
28. <i>Scores</i> Análisis individual Bosque Aleatorio despues del 2014 . . . . .	60
29. <i>Scores</i> Análisis inter-semestral Regresión lineal primeros 14 años . . . . .	61
30. <i>Scores</i> Análisis inter-semestral árbol de decisión primeros 14 años . . . . .	62
31. <i>Scores</i> Análisis inter-semestral Regresión lineal primeros 14 años . . . . .	63
32. <i>Scores</i> Análisis inter-semestral Regresión lineal años posteriores al 2014 . . .	64
33. <i>Scores</i> Análisis inter-semestral árbol de decisión años posteriores al 2014 . .	64
34. <i>Scores</i> Análisis inter-semestral bosque aleatorio años posteriores al 2014 . .	65
35. Comparacion de desempeños predicción inter-semestral para matematicas . .	67

## LISTA DE TABLAS

pág.

1. Pre-seleccion de modelos de regresión lineal año 2000 semestre 1 . . . . .	41
2. Modelo de regresion lineal seleccionado para el año 2000 semestre 1 . . . . .	41
3. Pre-seleccion de modelos de regresión lineal año 2014 semestre 2 . . . . .	43
4. Modelo de regresion lineal seleccionado para el año 2014 semestre 2 . . . . .	44
5. Pre-seleccion de modelos de arbol de decision año 2000 semestre 1. . . . .	46
6. Modelo de arbol de decision seleccionado para el año 2000 semestre 1 . . . . .	46
7. Pre-seleccion de modelos de árbol de decisión año 2014 semestre 2. . . . .	48
8. Modelo de árbol de decisión seleccionado para el año 2014 semestre 2 . . . . .	48
9. Pre-seleccion de modelos de bosque aleatorio año 2000 semestre 1. . . . .	50
10. Modelo de bosque aleatorio seleccionado para el año 2000 semestre 1 . . . . .	51
11. Pre-seleccion de modelos de bosque aleatorio año 2014 semestre 2. . . . .	52
12. Modelo de bosque aleatorio seleccionado para el año 2014 semestre 2 . . . . .	53
13. Resumen de los modelos seleccionados año 2000 . . . . .	54
14. Resumen de los modelos seleccionados año 2014 . . . . .	54

## RESUMEN

**TITULO:** Analítica de Datos para la Predicción del Desempeño en el Examen SABER11.<sup>1</sup>

**AUTOR:** ARON RENE FORERO AFRICANO.<sup>2</sup>

**PALABRAS CLAVE:** ICFES, Examen SABER11, Desempeño en el Examen SABER11, Deserción estudiantil, Analítica de Datos, Machine learning, Inteligencia Artificial.

### DESCRIPCION:

La analítica de datos que es la ciencia que tiene como tarea examinar los datos en bruto, y sacar conclusiones útiles, y el *Machine Learning* es una importante área en la Inteligencia Artificial, dos de las más populares áreas del conocimiento en la actualidad y es porque buscan automatizar procesos que se realizan manualmente, pero que pueden ser llevados a cabo por una máquina. La educación colombiana si bien ha visto una evolución significativa en los últimos años, se sitúa muy atrás en comparación con otros países. Los procesos de seguimiento y mejoramiento de la educación no dan los mejores resultados, y los programas de apoyo a los estudiantes son muy ineficaces y lentos; por tanto en esta investigación se propone la construcción de una herramienta de *machine learning* que contribuya a la identificación de los factores socio-económicos que afectan el rendimiento académico de los estudiantes.

Se plantea utilizar los datos que proporciona las pruebas SABER11, desde el año 2000 hasta el año 2017. A estos datos se les hizo un análisis exploratorio y luego una limpieza profunda para que fuese posible su uso posterior en un modelo de *machine learning*. Con los datos ya limpios se hizo un análisis más profundo para tratar cualquier fenómeno particular que presentaran los datos. Lo siguiente fue plantear los modelos predictivos, realizar las respectivas pruebas, la selección de mejores algoritmos y la experimentación para obtener resultados. Se consiguió un error medio absoluto de 7.13 y finalmente se realizó un análisis a los resultados y se concluyó que factores como los ingresos familiares, la educación de los padres, algunos aspectos del colegio, entre otros, influyen en el rendimiento académico del estudiante.

---

<sup>1</sup>Trabajo de Grado

<sup>2</sup>Facultad de Ingenierías Físico-mecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: GABRIEL RODRIGO PEDRAZA FERREIRA. Codirector: RAUL RAMOS POLLAN

## ABSTRACT

**TITLE:** DATA ANALYTICS FOR PREDICTING PERFORMANCE IN THE SABER11 EXAM.<sup>3</sup>

**AUTHORS:** ARON RENE FORERO AFRICANO.<sup>4</sup>

**KEYWORDS:** ICFES, SABER11 Exam, Performance in SABER11 Exam, Data Analytics, Machine Learning, Artificial Intelligence.

### **DESCRIPTION:**

Data analysis is the science that has the task to examine the pure data and make useful conclusions, and machine learning a really important area of the artificial intelligence, are two of the most popular areas of the knowledge in the present and that is because are looking for automation in a lot of processes that are manually done, but could be done by a machine. The Colombian education although have seen a good improvement in the last years, it is located far back in comparison to other countries. The monitoring processes and the processes of improvement are not giving the better results, and the students support programs are so slow and ineffective; thus in this investigation the construction of a machine learning tool that contributes to identify the socio-economic factors that affects the academic performance of the student is proposed.

It is proposed to use the SABER11 data, since the 2000 to the 2017 year. An exploratory analysis was done and then a data cleaning to this data with the purpose of using it into a machine learning model. With the data already cleaned a deeper analysis was done to treat any particular phenomenon that could exist. The next step was purpose the predictive models, do the respective tests, the selection of the best algorithms and the experimentation to get the results. A mean absolute error of 7.13 was achieved and finally an analysis of the results was done, and it was conclude that factors such as family income, parents education, some aspects of the school among others, influence the academic performance of the students.

---

<sup>3</sup>Research Work

<sup>4</sup>School of Physical-Mechanical Engineering. Department of Systems Engineering and Informatics. Advisor, GABRIEL RODRIGO PEDRAZA FERREIRA. Advisor, RAUL RAMOS POLLAN

## INTRODUCCION

La búsqueda continua por el mejoramiento es una cualidad intrínseca en el ser humano, involucrada en la mayoría de las áreas del conocimiento humano. Este afán por el mejoramiento no es reciente, sino que es antiguo tanto como el ser humano; otrora más leve, pero en la actualidad es un afán fuerte que mueve a las personas hacia el futuro a pasos considerablemente grandes.

En Colombia, podemos hablar de mejoramiento en todas las áreas, pero a un paso más lento y con técnicas más rudimentarias que las que existen en otros lugares del mundo. En el área de la educación, según MINEDU, existe una educación de calidad “Cuando todos los niños y jóvenes, independientemente de sus condiciones socio-económicas y culturales, alcanzan los objetivos propuestos en el sistema educativo y realizan aprendizajes útiles para su vida y para la sociedad” <sup>5</sup>. El Ministerio de Educación por medio de herramientas como el ICFES, busca evaluar la educación con el ánimo de mejorarla. “Las instituciones educativas combinarán los recursos para brindar una educación de calidad, la evaluación permanente, el mejoramiento continuo del servicio educativo y los resultados del aprendizaje, en el marco de su Programa Educativo Institucional.” <sup>6</sup>, también indican que se deben “Formular planes anuales de acción y de mejoramiento de calidad, y dirigir su ejecución.” <sup>6</sup>.

El rendimiento académico ha sido estudiado desde diferentes áreas de trabajo: pedagógica, psicológica, sociológica, etc. Los estudios pedagógicos basados en las técnicas de enseñanza-aprendizaje, técnicas de estudio, etc. psicológicos basados en la motivación, personalidad, factores cognitivos, etc. y sociológicos basados en los factores extra-aula que influyen en el estudiante.

---

<sup>5</sup>LA LLAWE. Estándares, evaluación y mejoramiento. [En Linea]<<http://www.mineducacion.gov.co/1621/article-87448.html>>

<sup>6</sup>Ley 715 de 2001. Titulo II: Sector Educación, Capítulo III: De las instituciones educativas, los rectores y los recursos. Artículo 9 y Artículo 10.4 [En Linea] <[http://www.mineducacion.gov.co/1621/articles-86098\\_archivo\\_pdf.pdf](http://www.mineducacion.gov.co/1621/articles-86098_archivo_pdf.pdf)>

<sup>6</sup>Ley 715 de 2001. Titulo II: Sector Educación, Capítulo III: De las instituciones educativas, los rectores y los recursos. Artículo 9 y Artículo 10.4 [En Linea] <[http://www.mineducacion.gov.co/1621/articles-86098\\_archivo\\_pdf.pdf](http://www.mineducacion.gov.co/1621/articles-86098_archivo_pdf.pdf)>

Diversos autores se han interesado en cómo las condiciones sociológicas pueden influir en el rendimiento académico de los estudiantes, por ejemplo COLEMAN<sup>7</sup> indica que el desempeño académico de los estudiantes estadounidenses estaba asociado en gran medida con su origen socio-económico.

En el Caribe colombiano se realizó un estudio con el ánimo de identificar los determinantes del desempeño académico en la educación superior, partiendo de los resultados del examen SABER PRO realizado por el ICFES, concluyendo de que los antecedentes socioeconómicos de los estudiantes si tienen influencia en los resultados, pero su influencia es muy poco significativa, dejando así el camino libre para decir que el desempeño académico está muy ligado al tipo y calidad de la institución educativa y dejando al descubierto una brecha de género relevante<sup>8</sup>.

Toda la investigación realizada por el ministerio de educación contribuye de buena manera en el monitoreo, control y mejoramiento, pero en cuanto a la implementación de algún método de automatización para detectar los factores influyentes en el rendimiento académico, solo se encontró una investigación realizada en la Universidad Autónoma de Manizales, que toca el tema del rendimiento académico mediante de una investigación enfocada a la deserción, ya que una de las razones más comunes por que se presenta deserción es por un rendimiento académico bajo<sup>9</sup>.

Este trabajo busca aportar una herramienta que complemente estos procesos de monitoreo, control y mejoramiento. Y cabe resaltar que esta investigación puede llegar a ser de las primeras en tocar el tema desde una perspectiva de automatización, ya que como se mencionó anteriormente, en el momento de la creación de este trabajo no se encuentra una gran cantidad de investigaciones sobre el tema en la literatura.

---

<sup>7</sup>COLEMAN, J. et al. 1996 Equality of Educational Opportunity. Washington: US Government Printing Office. [En Linea] <<http://library.sc.edu/digital/collections/eeoci.pdf>>

<sup>8</sup>Determinantes del desempeño académico universitario. El caso de la Región Caribe colombiana (2014). [En Linea] <<http://www.icfes.gov.co/docman/investigadores-y-estudiantes-de-posgrado/resultados-de-investigaciones/equidad/989-determinantes-del-desempeno-academico-universitario-el-caso-region-caribe-colombiana>>

<sup>9</sup>SISTEMA DE APOYO PARA LA ACREDITACIÓN DE LA CALIDAD DE PROGRAMAS ACADÉMICOS DE LA UNIVERSIDAD DE CALDAS, APLICANDO TÉCNICAS EN MINERÍA DE DATOS. [En Linea] <[http://repositorio.autonoma.edu.co/jspui/bitstream/11182/350/1/Msc.GyDilloSoft\\_InformeFinal\\_JuanCarlosGonzalez.pdf](http://repositorio.autonoma.edu.co/jspui/bitstream/11182/350/1/Msc.GyDilloSoft_InformeFinal_JuanCarlosGonzalez.pdf)>

## 1. PLANTEAMIENTO Y JUSTIFICACION DEL PROBLEMA

El aprendizaje automatizado, las maquinas de aprendizaje o el *machine learning*, es una de las áreas mas importantes de la inteligencia artificial, la cual participa en la automatización de procesos que pueden ser llevados a cabo por máquinas. Desde la antigüedad, las personas han visitado a videntes con el fin de conocer su futuro, y en la actualidad esto puede ser posible, gracias al *Machine Learning*. Mediante una recolección previa de datos relevantes, que estén relacionados con este fenómeno que se desea predecir, y un pre-procesado, es posible tratarlos con un algoritmo de predicción diseñado especialmente para estos, que nos podrá dar una estimación muy aceptable de lo que pasará a corto, mediano o largo plazo dependiendo de cómo se haya diseñado el algoritmo.

La educación en Colombia ha tenido una evolución positiva a lo largo de los últimos años, pero el problema llega a la hora de analizar la gran cantidad de datos de evaluación de los estudiantes, ya que estas evaluaciones se realizan a los grados quinto, noveno, once y al finalizar los estudios universitarios, en todo el país. Este análisis puede llevar mucho tiempo al realizarse manualmente, y la toma de decisiones puede tardar mucho en ser llevada a cabo, lo cual podría conducir a que los problemas encontrados en este análisis ya no se presenten con la misma frecuencia y se pasen por alto detalles más relevantes. Sin embargo, estos escenarios se podrían complementar haciendo uso de técnicas de *Machine Learning*, partiendo desde la mejora de los métodos de recolección de datos, hasta poder reducir considerablemente los tiempos a la hora de analizar los datos para así poder tomar decisiones más rápidas, sumándole que el sistema de predicción podría otorgar una sugerencia bastante acertada o que al menos deberá ser considerada de gran importancia.

El principal objetivo de este proyecto es contribuir con la evolución de la educación, aportando una herramienta de Machine Learning que permita a los expertos tomar mejores decisiones. Partiendo de los datos obtenidos desde el repositorio del ICFES, se planea implementar la herramienta propuesta.

## **2. OBJETIVOS**

### **2.1. OBJETIVO GENERAL**

Identificar correlaciones entre datos socioeconómicos, con los resultados de la prueba SABER11 a través de la construcción de modelos predictivos.

### **2.2. OBJETIVOS ESPECIFICOS**

- Diseñar una metodología para catalogar los datos del repositorio del ICFES.
- Segmentar los datos según criterios como distribución geográfica, estratificación socioeconómica.
- Establecer métricas de desempeño para la capacidad predictiva y el tiempo de ejecución.
- Definir y ejecutar tareas de *Machine Learning* para construir modelos predictivos del desempeño en cada área del conocimiento evaluadas en el examen SABER11.
- Validar la factibilidad (utilidad que nos pueden dar los datos para realizar las tareas y la utilidad del recurso de computo) de generar modelos predictivos basados en las tareas definidas previamente.

### 3. MARCO CONCEPTUAL

#### 3.1. EXPLORACION Y PRE-PROCESADO DE DATOS

**3.1.1. exploración o análisis exploratorio de datos:** Es la primera tarea a realizar en cualquier trabajo que incluya la aplicación de *machine learning*, ya que es muy importante que el ingeniero, investigador o científico de datos, conozca los datos que va a manejar y a utilizar en sus modelos predictivos. Se dice que los datos que se utilizaran tienen ruido cuando poseen errores y/o datos vacíos, situación que no permitirá la construcción de un modelo predictivo y si la permite el modelo no se comportara de la manera como debería. Ademas de esto se puede recurrir a la analítica descriptiva para realizar una visualización de los datos, y así poder encontrar caminos mas rápidos para su entendimiento.

En este trabajo se realizo un análisis exploratorio breve, se construyeron gráficas empleando la analítica descriptiva, y se realizo un catalogo en donde se explico que y de que tipo de datos se poseían.

- **Estadística descriptiva:** se basa en el análisis y organización de un conjunto de datos, del cual se realizara una descripción, generalmente con ayuda de tablas, y gráficas para facilitar la identificación de patrones, tendencias o comportamientos particulares en una o varias variables.
- **Catalogo de datos:** se planteo como la construcción de una estructura de directorios, en donde se listan las componentes de la sección del SABER11 de la base de datos del ICFES, y se explica que datos contiene cada sub-carpeta y cuales se utilizaran para esta investigador, excluyendo algunos que podrían servir para futuros trabajos.

**3.1.2. pre-procesado de datos:** Uno de los pasos mas importantes y que mas tiempo toma en un proyecto de analítica de datos y/o que incluya machine learning es

el pre-procesado de los datos. El propósito de esta fase de pre-procesado es conseguir un conjunto de datos completamente limpio y preparado para su posterior uso en un modelo predictivo. Existen varias técnicas y generalmente queda a criterio del científico de datos escoger las mas convenientes y realizar la toma de decisiones sobre los datos tratados. En esta investigación fue necesario realizar una limpieza de datos, y un análisis de multicolinealidad.

- La **limpieza de datos** es completamente necesaria, pues hay que disminuir la cantidad de ruido que posean los datos ya que esto puede causar que no se obtengan buenos resultados con los modelos predictivos o incluso puede causar que no sea posible construir modelos predictivos a partir de estos datos. El ruido puede ser causado por datos erróneos (por ejemplo caracteres extraños en una variable numérica), por valores desconocidos, vacíos o faltantes, también llamados *Nan* o *Na* (*Not a Number, Not Available* respectivamente), los cuales se generan cuando no se tiene información sobre alguna variable en un registro.

Existen varias técnicas para tratar estos valores generadores de ruido, en este trabajo se utilizaron los método de *Back Fill* y *Forward Fill* que consisten en llenar los valores *Nan* con el ultimo valor valido encontrado, o con el siguiente respectivamente, y también se utilizo la técnica de llenado de valores *Nan* con la media aritmética de la variable.

Ademas de esto es necesario tener en cuenta que existen dos tipos de variables, las variables numéricas y las categóricas. Las variables numéricas son de tipo cuantitativo y no presentan ningún problema a la hora de construir un modelo predictivo, pero las variables categóricas son de tipo cualitativo y presentan un problema a la hora de realizar una predicción ya que no son datos manejables u operables numéricamente y no poseen una relación de orden es decir que ninguno vale mas que otro. El ejemplo mas trivial de una variable categórica es el genero de una persona, ya que puede ser hombre o mujer. Este inconveniente puede ser solucionado de manera sencilla reemplazando los valores categóricos por un valor numérico, a este proceso muchas veces se le conoce como codificación.

- **Análisis de multicolinealidad:** Luego de obtener un conjunto de datos limpios, es posible pasar a la utilización de este para la creación de un modelo predictivo,

pero, no se estaría teniendo en cuenta uno de los problemas mas influyentes a la hora de utilizar datos para predecir alguna variable, la multicolinealidad. Consiste en revisar si existen correlaciones fuertes entre las variables predictivas y tomar decisiones sobre que hacer con ellas. Generalmente se elimina una de las dos variables pues lo único que genera tener las dos variables en el conjunto de datos es una predicción con baja precisión o malos resultados. Este problema se aborda aplicando el **coeficiente de correlación de Pearson** [ecuación 3.1].

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \quad (3.1)$$

### Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson nos proporcionara un valor en el rango de  $[-1, 1]$ , y a partir de este valor se podrá clasificar la correlación como fuerte o débil, teniendo en cuenta que una correlación es fuerte y directa cuando el coeficiente es igual o mayor a 0.5 o que es fuerte e indirecta cuando el coeficiente es igual o menor a -0.5.

- ***Feature Engineering:*** Otro procedimiento importante que se realizó fue el *Feature Engineering* o ingeniería de características. Generalmente este paso se realiza antes de la construcción de modelos predictivos, agregando o quitando variables que el científico de datos considere que pueden aportar a la capacidad predictiva de los modelos así como la eliminación de variables realizada en el análisis de multicolinealidad, pero en esta investigación, se utilizó una característica de los bosques aleatorios, la cual nos devuelve la influencia de las variables en la predicción, de manera numérica más específicamente como un porcentaje de influencia en el resultado.
- ***Principal Component Analysis (PCA):*** Así como en el análisis de multicolinealidad y en el *Feature Engineering*, el método de PCA se basa en reducir la dimensionalidad de los datos. PCA es uno de los métodos más populares en la

actualidad y se utiliza para reducir la cantidad de variables predictivas, el tiempo de entrenamiento, y ademas de eso mejorar la capacidad predictiva de los modelos eliminando variables que generan ruido pero que no se pueden tratar manualmente, y dejando nuevas variables no correlacionadas entre si. Este es un proceso muy útil pero matemáticamente es algo complejo ya que hace uso de la descomposición en valores singulares <sup>1</sup>.

- **Polynomial Features** Es un proceso que consiste en calcular todas las combinaciones polinomiales de las variables, con un grado menor e igual al grado especificado por el científico de datos. De esta manera el conjunto de datos obtendrá muchas mas columnas, generadas en este proceso. Por ejemplo: si se pasa un conjunto de datos de dos dimensiones así: [a,b] con grado 2 el conjunto de datos resultante seria: [1, a, b, ab,  $a^2$ ,  $b^2$ ] <sup>2</sup>.
- **Segmentación de datos:** En este proyecto se planteo realizar una segmentación de datos para revisar si alguna de las divisiones obtenidas tenia alguna influencia significativa sobre los resultados de los estudiantes en las pruebas SABER11.

### 3.2. MACHINE LEARNING

En español maquinas de aprendizaje, aprendizaje de maquina, o aprendizaje automático, se puede pensar en *machine learning* como un conjunto de herramientas y métodos que tienen como fin inferir patrones y extraer ideas de un registro u observación del mundo real <sup>3</sup>. El *machine learning* es una rama de la inteligencia artificial, y consiste en entrenar un modelo predictivo con un conjunto de datos de entrenamiento, luego de esto, el modelo creado podrá ser utilizado para predecir el fenómeno con el que fue

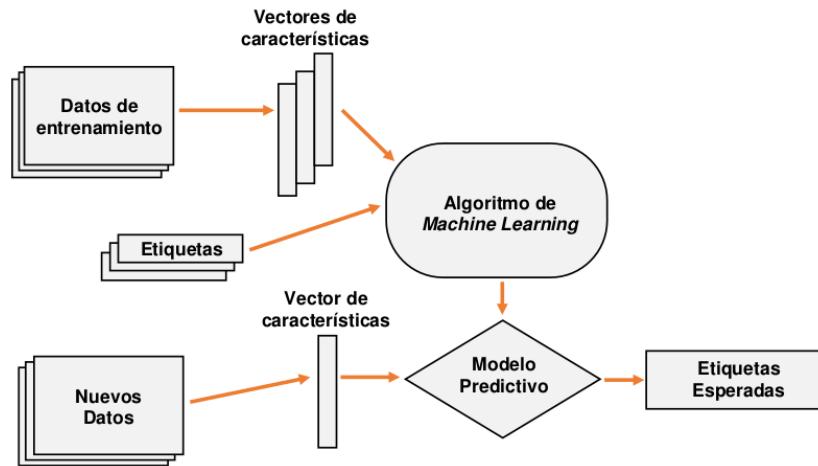
---

<sup>1</sup>SCIKIT-LEARN Developers. sklearn.decomposition.PCA [En Linea], [Revisado 26 Julio 2018]. Disponible en internet <<http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>>

<sup>2</sup>SCIKIT-LEARN Developers. sklearn.preprocessing.PolynomialFeatures [En Linea], [Revisado 26 Julio 2018]. Disponible en internet <<http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html>>

<sup>3</sup>CONWAY,Drew y MYLES,John. Machine Learning for Hackers. Estados Unidos de América: O'Reilly Media, Inc.:2012

**Figura 1:** Esquema de una predicción supervisada, se puede ver el proceso que tiene la creación de un modelo predictivo de izquierda a derecha y de arriba hacia abajo. Imagen recuperada de RUEDA,Edwin. Predicción de Series Financieras con redes neuronales recurrentes. Universidad Industrial de Santander.



entrenado, con los mismos o con datos diferentes a los que se utilizaron en el entrenamiento; finalmente, el modelo sera medido por su desempeño de acuerdo a una métrica escogida por el ingeniero de acuerdo a lo que se busque resolver.

Existen varios tipos de algoritmos de predicción, pero en esta investigación vamos a tratar el *machine learning supervisado*, en el cual se utilizan unos datos de entrenamiento que vienen dados por pares. Estos pares de datos serán: los arreglos de dimensiones [1xn] y una componente de etiqueta que sera la variable a predecir, la cual puede ser tanto numérica como categórica, dependiendo del tipo de predicción que se intente realizar [Figura 1]. Este tipo de aplicación del *machine learning* es muy utilizado en diversos campos como: la bioinformática, el reconocimiento de visión por computador, clasificación de imágenes, etc.

**3.2.1. regresión en *machine learning*** : Es una de las ramas mas importantes del *machine learning* supervisado, en la cual se entrena el modelo predictivo con los datos escogidos y se predice una variable de tipo numérico que depende de una o

varias variables predictivas. A diferencia de los problemas de clasificación en *machine learning* en donde se puede encontrar que una observación pertenece a una u otra clase, y pueden existir  $n$  finitas clases, en la regresión la predicción de salida puede variar tanto como los números existentes en el intervalo estudiado, es decir, infinitas posibles predicciones.

**3.2.2. regresión lineal:** En inglés *Linear Regression*, es quizás el método más conocido para predecir el comportamiento de los datos, se utiliza para modelar la relación entre dos variables, una dependiente  $Y$ , y uno o más parámetros  $X$  [Figura 2]. La característica principal de este enfoque es que la relación entre las dos variables puede ser modelada con una simple ecuación lineal así:

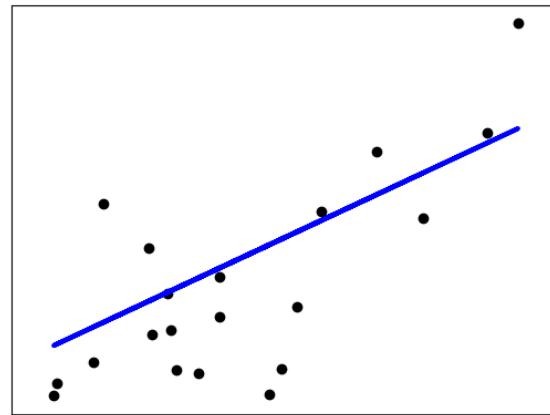
$$Y = mX + B \quad (3.2)$$

Cambien existe otro tipo de regresión lineal, que seria la regresión lineal múltiple. Este tipo de regresión es útil cuando se posee una variable a predecir que depende de varias variables y no solo de una como en la regresión lineal simple. Este tipo de regresión lineal correspondería a una ecuación lineal un poco mas compleja que la anterior, de esta forma:

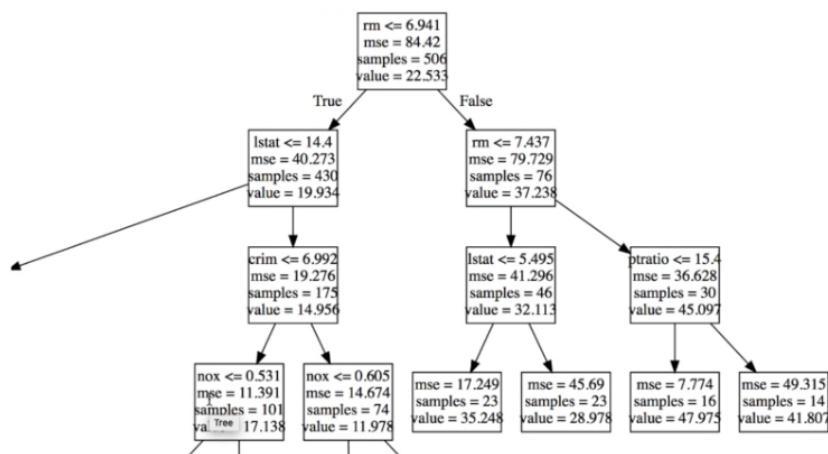
$$Y = B + m_1X_1 + m_2X_2 + \dots + m_nX_n \quad (3.3)$$

**3.2.3. árbol de decisión de regresión:** En inglés *Decision Trees* son una estructura básica en la informática, estos son muy útiles pues permiten visualizar las diversas opciones que se tienen y el camino que se debe seguir para llegar a ellas, y en la inteligencia artificial son utilizados como modelos de predicción, en los cuales, dado un conjunto de datos, fabrican diagramas de construcciones lógicas que sirven para categorizar y representar la serie de pasos o condiciones para llegar a la solución de un problema.

**Figura 2:** Regresion Lineal simple. En el eje X se situaría la variable predictiva que en este caso seria la variable dependiente, y en el eje Y se situaría la predicción. La recta azul intenta ajustarse a los puntos negros, y de esta manera estimar un valor de Y teniendo un valor del eje X. Imagen recuperada de SCIKIT-LEARN DEVELOPERS. Linear Regression Example [En linea]. <[http://scikit-learn.org/stable/auto\\_examples/linear\\_model/plot\\_ols.html](http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html)>



**Figura 3:** Ejemplo de Árbol de Decisión de regresión. Imagen recuperada de GOMILLA, Juan. Curso completo de Machine Learning: Data Science en Python. Udemy [En Linea] <<https://www.udemy.com/machinelearningpython/>>



**3.2.4. bosques aleatorios de regresión:** En inglés *Random Forest* son la combinación de árboles predictores que dependen de un vector aleatorio probado independientemente, y con la misma distribución para cada uno. En otras palabras, este método construye una larga colección de árboles no correlacionados y los promedia, reduciendo y/o controlando la varianza e incrementando la capacidad predictiva del modelo resultante. Los bosques aleatorios pertenecen a la familia de *Ensemble Methods* que simplemente consiste en combinar varios modelos del mismo o de diferente tipo para conseguir un modelo más robusto y predicciones más precisas.

**3.2.5. métrica de desempeño:** La métrica es una medida necesaria para determinar si el desempeño del modelo creado es bueno o malo y la métrica seleccionada se debe adaptar convenientemente al problema seleccionado y al entendimiento del cliente final al que se le mostrarán los resultados. La métrica que se seleccionó para este trabajo es el **promedio del valor absoluto**, que en inglés es *mean absolute error*, conocido como **MAE** y tiene la siguiente forma:

$$mae = \frac{\sum(Y - Y_p)}{n} \quad (3.4)$$

siendo  $Y_p$  la predicción obtenida,  $\mathbf{Y}$  el valor esperado o el valor real y  $n$  la cantidad de predicciones realizadas.

Esta métrica fue escogida pensando en la simplificación de la comprensión de los resultados de los modelos, por cualquier persona externa al proyecto, ya que representaría los puntos de error con respecto a cada puntaje en su misma escala o intervalo, por ejemplo: el modelo construido tiene un error absoluto promedio de 4 puntos sobre 100 (100 puntos ya que los puntajes del ICFES están en el intervalo de [0,100]).

**3.2.6. métodos de validación:** Existen diversas técnicas de validación, pero en esta investigación se utilizarán en las pruebas solo dos, la *Cross-validation* y la técnica

de *Shuffle Split*:

- **Cross-Validation:** Es una técnica de validación en la cual se dividen los datos en 2 partes, una parte para el entrenamiento y otra parte para la validación. Este proceso se realiza varias veces y se guardan los resultados para al final promediar los resultados y tener una estimación del resultado mucho mas precisa y confiable. Esta técnica garantiza que las validaciones serán siempre distintas en cada iteración, pues los datos que son usados en una validación, no serán usados para otra validación. Los tamaños de las partes tanto de entrenamiento y validación los puede fijar el científico de datos a su criterio.
- **ShuffleSplit:** Esta tecnica consiste en un muestreo aleatorio simple sin reemplazo, pero que en cada iteracion justo antes de que se tome la muestra para validacion, se mezcla la población (los datos en este caso), lo que permitiría que un elemento que apareció en la primera muestra de validacion aparezca en la segunda, o en la tercera, o en todas las muestras, esto permite obtener infinitos segmentos de validacion incluso cuando nuestro archivo es finito. Así se pueden obtener muestras que den mejores resultados de una manera practica y un poco mas acercada a la realidad.

## 4. METODOLOGIA DESARROLLADA

### 4.1. CATALOGO DE DATOS Y ANALISIS EXPLORATORIO

En este proyecto se plantea un catalogo de datos para definir y dejar claro que datos se utilizaran durante el proceso y que datos no. Ademas de esto se plantea realizar un análisis exploratorio de datos, con el fin de conocer los datos que se van a tratar. Con esto se pueden obtener características específicas de cada archivo como su tamaño, numero de observaciones, cantidad de variables predictivas, cantidad de variables a predecir, cantidad de variables numéricas, cantidad de variables categóricas, etc, datos que serán útiles para el desarrollo del proyecto.

**4.1.1. catalogo de datos:** Lo primero que se realizo fue el catalogo de datos del repositorio del ICFES, con el fin de organizar y así dejar claro con que datos se trabajaría en este proyecto.

El repositorio de datos del ICFES contiene los siguientes directorios:

- PERCE-1
- SERCE-1
- SABER 11
- SABER 359
- SABER PRO
- SABER TyT
- TERCE

De los cuales se utilizo solo el directorio que contiene los resultados de la prueba SABER11. Dentro de este directorio encontramos:

## ■ BASES DE DATOS

En este directorio se encuentran las bases de datos desde el año 2000 separadas por semestre hasta el año 2017-1 (hasta la fecha en que se inicio este proyecto). En cada fichero se encuentra información socioeconómica del estudiante, y de la institución y también los resultados de la prueba SABER11 realizada por el estudiante, y ademas de esto, la base de datos brinda anonimato completo a los participantes del examen. A pesar de que las bases de datos se encuentran agrupadas, se encuentra una diferencia entre archivos bastante significativa, esto debido a cambios que se hicieron en la prueba a medida que pasaba el tiempo y un cambio grande que sucedió en el año 2014 semestre 2.

## ■ CLASIFICACION DE PLANTELES

Aquí se encuentra una carpeta con bases de datos en las cuales existe información sobre la clasificación de los planteles educativos. En los cuales encontramos información de jornada, población(femenino, masculino o mixto), puntaje del colegio en general y en cada materia, calendario que maneja el plantel, e información geográfica.

## ■ INDICE SOCIOECONOMICO

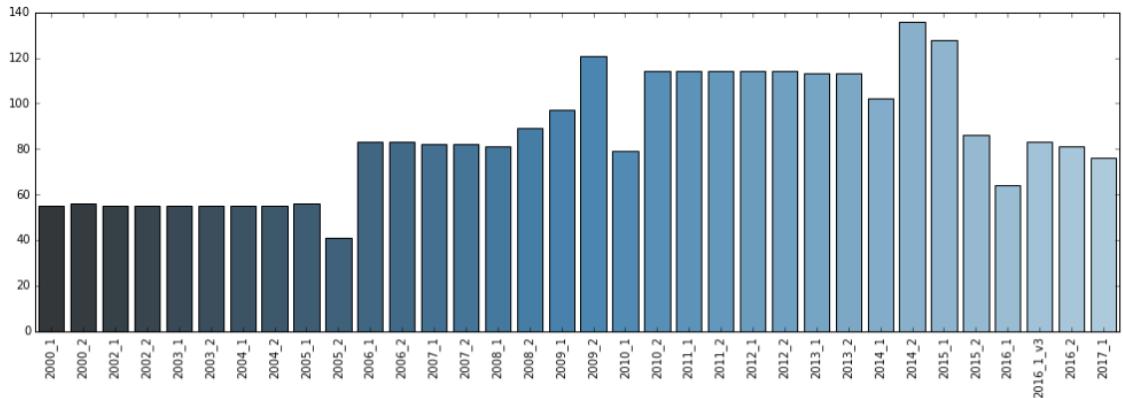
Aquí se encuentra una carpeta que posee los datos sobre el índice socio-económico para los estudiantes y colegios del 2008 y 2009 que presentaron la prueba.

## ■ PRESABER

se encuentran los resultados de la prueba PRESABER, otra prueba realizada con el ánimo de dar un diagnóstico antes de la prueba SABER11 medidos en deciles. En estos datos no se encuentran datos socio-económicos, pero sí datos sobre propiedades de la familia.

Todos los datos tienen que ver con el examen SABER11, pero al tener esta característica de anonimato, no existe una manera clara de relacionar los ficheros, por esta razón se decide trabajar solo con el directorio BASES DE DATOS, ya que se considero que en estos archivos se tiene información suficiente para realizar buenas predicciones.

**Figura 4:** Cantidad de Columnas por archivo. Se puede ver la cantidad de columnas que posee cada archivo antes de la limpieza.



**4.1.2. análisis exploratorio** Antes que nada es necesario hacer un análisis exploratorio de cada conjunto de datos del que se disponga, para adquirir una vista general de lo que se va a tratar. Se encontró una gran variación en la cantidad de columnas que poseía cada archivo como se ve en la [Figura 4].

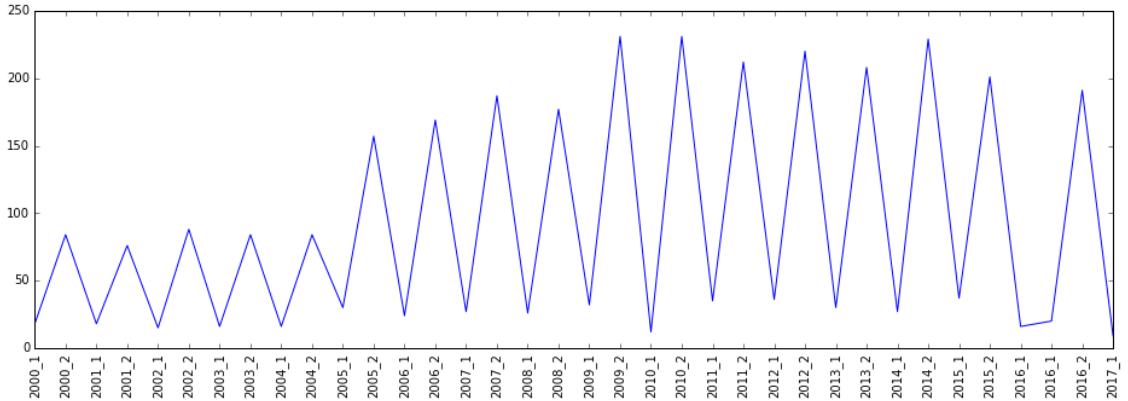
Ademas de esta variación, se encontró una diferencia en la cantidad de observaciones que poseían los archivos de primer semestre con los del segundo semestre en todos los años como se ve en la [Figura 5].

Esto se debe a que la cantidad de personas que presentan el examen SABER11 en el segundo semestre del año, es mucho mayor a la cantidad de personas que lo presentan en el primer semestre, ya que los estudiantes de grado once de los colegios lo presentan en el segundo semestre del año.

## 4.2. PRE-PROCESADO

Habiendo terminado el análisis exploratorio de datos, y seleccionado los datos con los que se trabajara a lo largo de la investigación, se plantea realizar una limpieza de datos, para eliminar o rellenar cualquier cantidad de valores vacíos (*NaN* o *Null*) que puedan

**Figura 5:** Pesos en Megabytes [MB] de los archivos



existir y cualquier elemento extraño que se encuentre en las columnas de los conjuntos de datos y no tenga nada que ver con lo que describe la variable. Finalmente, poder obtener conjuntos de datos planos y totalmente en formato numérico para que sea posible realizar una regresión, la cual sera nuestro tipo de predicción a implementar.

**Limpieza de datos:** Ya que los datos que se utilizaron en esta investigación fueron bajados directamente del repositorio del ICFES <sup>1</sup> y presentan muchos errores en los primeros años, y valores vacíos *NaN* o *Null*, fue necesario realizar una limpieza a todos y cada uno de los archivos obtenidos.

Consistió en encontrar los datos extraños de cada columna y a su vez de cada archivo, por ejemplo: guiones (-), asteriscos (\*), arrobas (@), espacios, entre otros( como en la [Figura 6]), en donde solo deberían existir datos numéricos, o en el caso de que se tratara de una variable categórica, datos que no tienen nada que ver con los posibles valores que podría tomar cierta variable; al encontrarlos, estos datos fueron convertidos en valores desconocidos o valores *NaN*.

---

<sup>1</sup>Repositorio de Datos del ICFES<<ftp://ftp.icfes.gov.co>>

**Figura 6:** Muestra de Datos antes de la limpieza, se pueden apreciar arrobas y espacios vacíos en una variable de tipo *string* o cadena, llena de valores numéricos.

```
['0', '00 1', '0000', '0001', '001', '0052', '0070', '01', '0128', '017  
...', '@818', '@819', '@821', '@823', '@85', '@92', '@999', '@@', '@@4',  
'@@@@'].
```

**Tratamiento de los valores *NaN*** Luego de la limpieza inicial, se decidió que los valores *NaN* resultantes se deberían distribuir de manera proporcional entre los valores realmente validos que tenia la columna; Esto para no romper las proporciones que tenían los valores antes del tratamiento, y aun así conseguir un conjunto de datos plano y sin huecos. El proceso anterior se realizo para una gran cantidad de variables, pero se implemento un tratamiento distinto para variables numéricas como los mismos puntajes, en donde se reemplazo los valores *NaN* por la media aritmética de la columna. Y en algunos casos, en donde se encontraban valores *NaN* en variables que tenían por ejemplo información de posición geográfica, se utilizaron los métodos de *Back Fill* y *Forward Fill*.

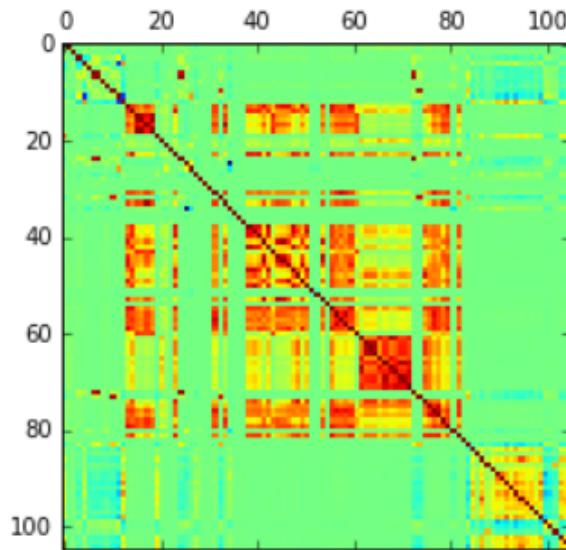
**Tratamiento de las variables categóricas** Como es común, se encontraron variables categóricas, y para resolver esto se construyo una función que codificar los valores de cada variable categórica encontrada en valores numéricos. De esta manera se obtuvieron archivos listos para ser utilizados por un modelo predictivo.

### 4.3. ANÁLISIS GENERAL Y PROFUNDO

Después del pre-procesado de datos, se plantea realizar un análisis mas profundo de estos datos seleccionados y ya limpios, con el animo de encontrar alguna tendencia en las variables a predecir, y aun mas importante revisar si existe algún problema de multicolinealidad, tomar alguna decisión para resolverlo.

Los siguientes análisis se realizaron para el primer archivo (año 2000 semestre 1) y

**Figura 7:** Matriz de Correlación del año 2014 en donde se pueden apreciar colores suaves como el azul celeste, el verde aguamarina o el amarillo claro, que muestran una correlación baja entre las variables, y también colores mas sólidos como Naranjas, Rojos y Azules oscuros que muestran una correlación alta.



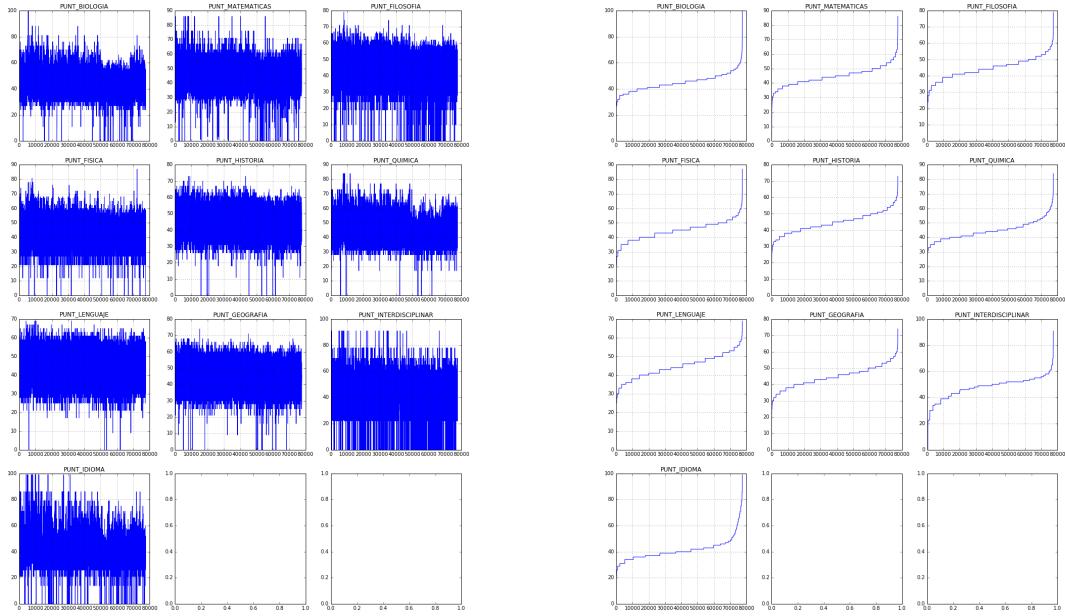
para el archivo del año 2014 semestre 2, ya que en este punto temporal tenemos este cambio bastante significativo en la metodología de evaluación del ICFES en el examen SABER11.

**4.3.1. tratamiento de la multicolinealidad:** El primer paso que se realizo fue crear una matriz de correlación la cual implementa el coeficiente de correlación de Pearson ecuación 3.1 definido anteriormente, para poder ver las correlaciones entre variables predictivas.

Para hacer mas fácil la tarea de encontrar las correlaciones fuertes, se gráfico la matriz de correlación obtenida como un mapa de calor [Figura 7], así seria posible relacionar los colores mas fuertes u oscuros con las relaciones fuertes entre columnas.

Al encontrar las correlaciones fuertes, se tomo la decisión de eliminar una de las dos

**Figura 8:** Targets u objetivos a predecir. A la izquierda se pueden ver los *targets* tal y como estan indexados en la tabla, y a la derecha se pueden ver los *targets* ordenados de menor a mayor.



variables fuertemente correlacionadas, para así mejorar la capacidad predictiva del modelo.

**4.3.2. visualización de datos** Ademas de las técnicas explicadas con anterioridad se realizo una visualización de los datos de interés, que son los *targets*, nuestros objetivos o variables a predecir; que son los puntajes de las materias. Y se comportan como se ve en la [Figura 8].

Se puede visualizar que al estar desordenados, los puntajes tienen una forma caótica, sin una tendencia clara; mientras que cuando están ordenados, se puede evidenciar una tendencia parecida a la gráfica de:

$$Y = X^3 \quad (4.1)$$

**Segmentación** Los segmentos que se tuvieron en cuenta para esta investigación fueron de tipo:

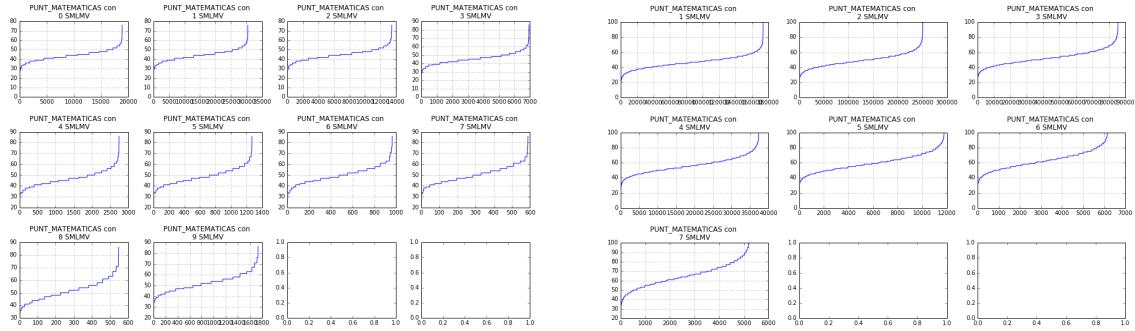
1. Socio-económico Los datos de cada semestre se dividieron por el ingreso familiar mensual, en Colombia la medida estándar es el salario mínimo, entonces se tienen grupos de acuerdo a cuantos salarios mínimos gana la familia mensualmente.
2. Geográfico En este caso los datos de cada semestre se vieron divididos por departamento, en Colombia encontramos 32 departamentos, aunque en los archivos no se encuentran todos los departamentos registrados ya que hay departamentos que no poseen ningún individuo presentando el examen SABER11.

Este proceso de segmentación sera realizado en los archivos en donde se a realizado el análisis profundo anterior, se buscara la columna que mas se ajuste al criterio de segmentación, ya que alguna de las variables clave para la segmentación pudo haber sido eliminada en la limpieza de datos.

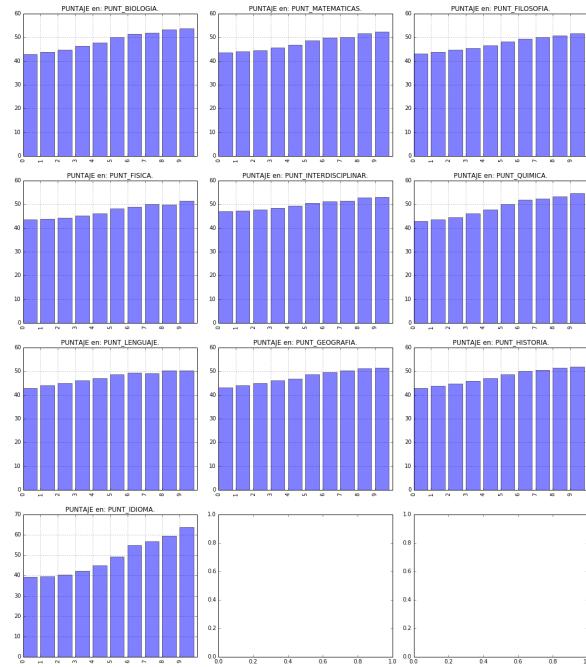
**4.3.3. análisis de los segmentos** El análisis se realizo para la misma materia con los dos tipos de segmentación, para así poder compararla con mas seguridad.

- **Análisis de segmentación Socio-económica:** Si se analiza una materia, distribuida teniendo en cuenta el ingreso familiar mensual en salarios mínimos antes del cambio en el 2014 [Figura 9] y después del cambio [Figura 9]. Se puede observar que la tendencia es la misma para todas las cantidades de ingresos y para antes o después del cambio del 2014. Pero el rendimiento promedio si se ve ligeramente afectado por los ingresos familiares, se ve un pequeño incremento a medida que crecen los ingresos familiares [Figura 10].
- **Análisis de segmentación Geográfica:** Si se analiza una materia, distribuida de manera geográfica antes del cambio en el 2014 [Figura 11] y después del cambio [Figura 12]. Se puede observar que la tendencia no cambia entre divisiones

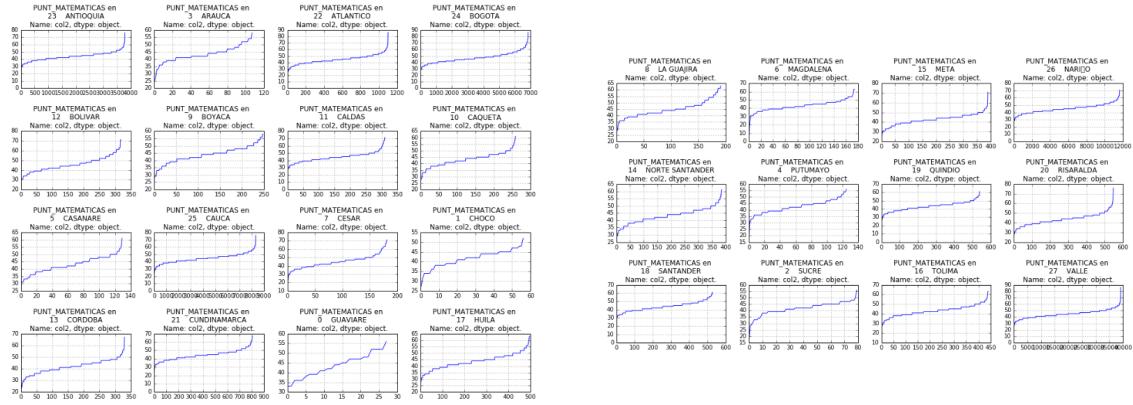
**Figura 9:** Visualizacion Matematicas segmentada por ingresos familiares ante del cambio del año 2014 a la izquierda y despues del cambio del año 2014 a la derecha. Misma tendencia, y un leve incremento a medida que se aumentan los ingresos.



**Figura 10:** Visualización del Promedio de las Materias por ingreso familiar mensual. Antes del cambio en el año 2014. Se puede notar un incremento a medida que crecen los ingresos familiares.



**Figura 11:** Visualización del puntaje de Matemáticas en segmentos geográficos antes del cambio del 2014. Misma tendencia para los segmentos graficados, y las diferencias en promedio son muy pequeñas



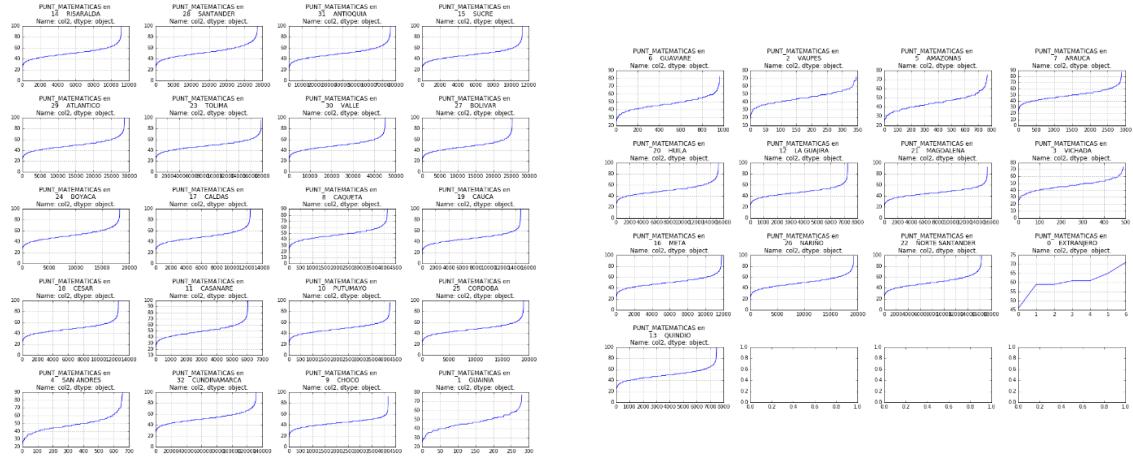
geográficas, ni tampoco con el cambio del año 2014, y que el promedio de las notas de los estudiantes es en gran medida el mismo, pues se mantiene entre 40 y 50 puntos, sin importar si es antes o después del cambio del año 2014, pero si se puede ver que en algunos departamentos el promedio se mantiene un poco mas alto [Figura 13], pero entre 40 y 50 puntos siempre. Lo mismo sucede con las demás materias, no cambia la tendencia, ni tampoco hay un cambio significativo en el promedio.

#### 4.4. CONSTRUCCION Y SELECCION DE ALGORITMOS DE PREDICCIÓN

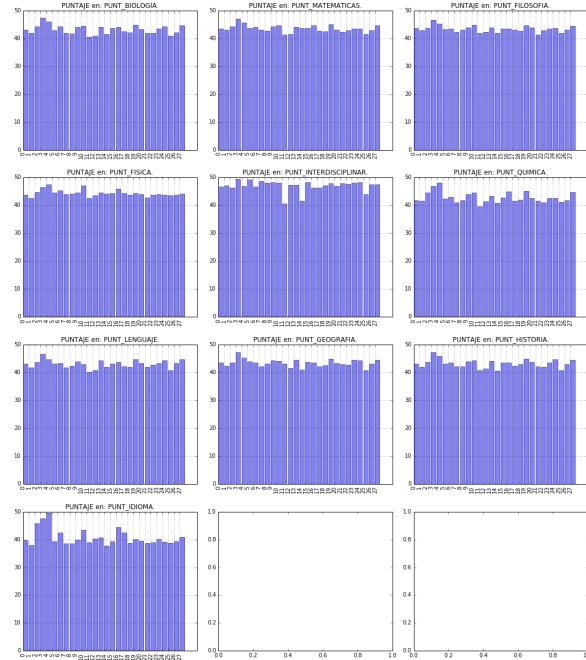
Con el análisis en profundidad de los datos hecho, se realizaran las primeras pruebas, en cada prueba se proporcionaran los mismos datos pero tratados de distintas maneras. Al terminar las pruebas se escogerá la manera de tratamiento de datos que mejores resultados proporcione de acuerdo a la métrica seleccionada, en combinación con cada algoritmo propuesto.

En esta investigación solo cabe un tipo de predicción: la regresión, ya que buscamos predecir un numero que seria el puntaje de una materia en particular, para un estudiante. Las pruebas se realizaron con los algoritmos mas famosos y/o conocidos: regresión

**Figura 12:** Visualización del puntaje de Matemáticas en segmentos geográficos después del cambio del 2014. Misma tendencia para los segmentos graficados, y las diferencias en promedio son muy pequeñas



**Figura 13:** Visualización del Promedio de las Materias por segmento geográfico o departamento. Antes del cambio en el año 2014. El promedio se mantiene entre 40 y 50 puntos en su mayoría, y se ve una diferencia muy pequeña entre departamentos.



**Figura 14:** Tiempo de ejecución del SVR. Se puede ver que el algoritmo de maquinas de soporte vectorial esta corriendo en un *job* desde hace 5 días, 3 horas, 58 minutos y 34 segundos, lo cual es un tiempo inaceptable para el cronograma establecido.

JOBID	PARTITION	NAME	USER	ST	TIME	NODES	NODELIST(REAISON)
94976	all	vasp:tes	aggh	PD	0:00	1	(Resources)
94977	all	vasp:tes	aggh	PD	0:00	1	(Priority)
94978	all	vasp:tes	aggh	PD	0:00	1	(Priority)
94979	all	vasp:tes	aggh	PD	0:00	1	(Priority)
94980	all	Temp283.	jorgeleo	PD	0:00	2	(Resources)
94193	all	md0G1298	jpwillab	R	14-18:28:13	1	guane02
94196	all	anti.sh	jpwillab	R	14-18:19:55	1	guane02
94592	all	vasp:tes	stiven.s	R	1-09:26:28	2	guane[09-10]
94603	all	qe:WSe_C	cbeltran	R	1-09:26:23	2	guane[15-16]
94608	all	vasp:tes	stiven.s	R	9-04:56:32	2	guane[03-04]
94723	all	vasp:tes	Valderra	R	5-05:15:25	2	guane[08-09]
94741	all	python.b	Forero	R	5-03:58:34	1	guane04
94826	all	vasp_mpi	ccelis	R	3-16:13:51	1	guane11
94847	all	sml_jupy	sergioml	R	2-13:30:04	1	guane01
94908	all	time	vsbasto	R	1-20:55:29	1	guane01
94910	all	time	vsbasto	R	1-20:53:29	1	guane01
94939	all	vasp:tes	Valderra	R	11:35:36	2	guane[12,14]
94962	all	Temp300.	jorgeleo	R	20:09:44	2	guane[06-07]
94967	all	vasp:tes	dayharri	R	16:13:09	2	guane[01,13]

lineal, arboles de decisión, y bosques aleatorios. Ademas cabe resaltar que estas pruebas se realizaron en el archivo del año 2000 semestre 1 y en el archivo del año 2014 semestre 2, en total dos veces debido al cambio que se menciono con anterioridad. Las pruebas realizadas se validaron con datos que no fueron utilizados en el entrenamiento, haciendo uso de los metodos de *cross-validation* y *ShuffleSplit*.

Ademas de los algoritmos mencionados, se intento utilizar maquinas de soporte vectorial (SVM) para realizar las regresiones, pero se encontró que no era factible utilizarlas ya que tomaba demasiado tiempo en el proceso de entrenamiento y validación [Figura 14], debido a la dimensión de los datos.

Los algoritmos que se utilizaran para las pruebas y experimentación serán:

- Regresion Lineal: La regresión lineal que se probara en esta investigación sera múltiple, ya que no solo se posee una variable predictiva y un *target*, sino que se poseen muchas variables predictivas y se quiere predecir una variable, que seria nuestro *target*.
- Árbol de decisión de regresión: Generalmente se habla de arboles de decisión para la clasificación, pero en este proyecto se hizo uso de los arboles de decisión para

regresión.

- Bosques aleatorios de regresión: Al igual que los arboles de decisión, los bosques aleatorios son utilizados generalmente en clasificaciones, pero en esta investigación se utilizara la implementación de bosques para regresiones.

**4.4.1. tratamiento de datos** Lo primero que se definió antes de realizar los experimentos fueron las maneras de implementar el algoritmo en su entrenamiento y validación, se planteo hacer uso del proceso de *cross-validation* y de *ShuffleSplit*, con un tamaño de entrenamiento del 80 % de los datos y con un tamaño para la validación de un 20 % de los datos. Se decidió obtener solo 5 validaciones por prueba con cada uno de los métodos. Ademas de estos métodos de validación, se decidió hacer uso del proceso de *Polynomial Features* y del proceso PCA.

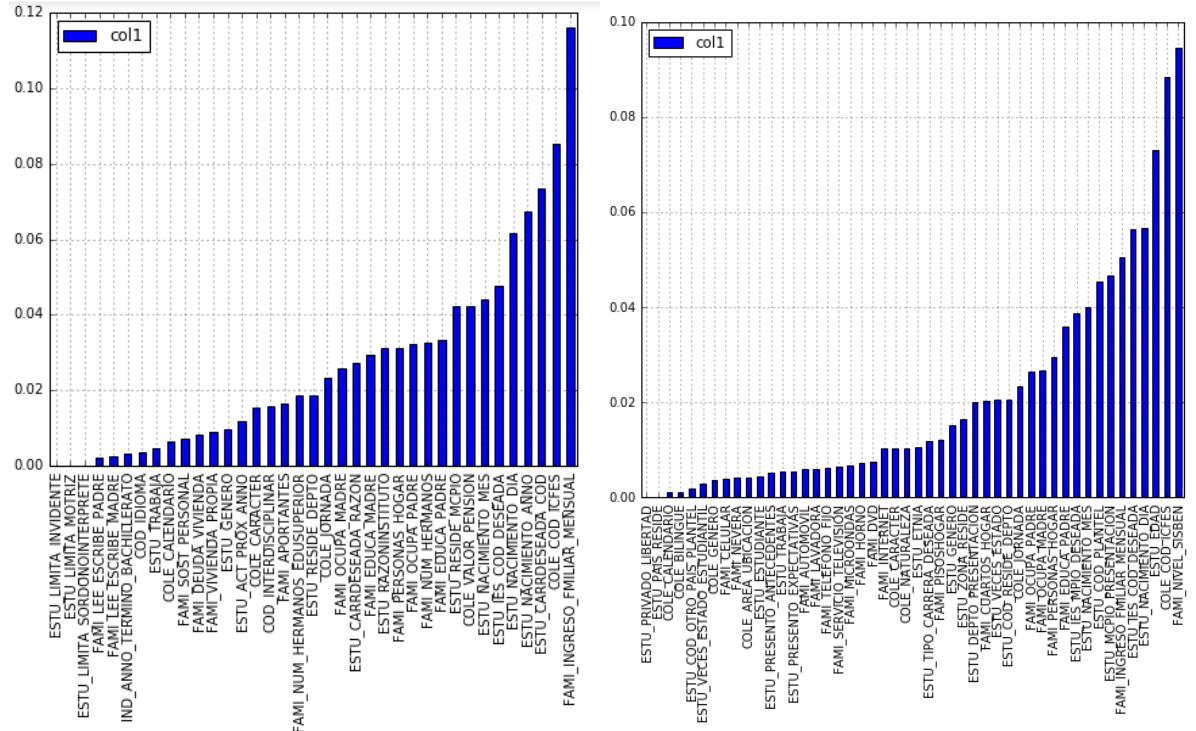
**4.4.2. pruebas planteadas** Estas pruebas se realizaron con los 3 algoritmos escogidos. Cuando se hace referencia a los datos originales, se refiere a todo el conjunto de datos obtenido después de la limpieza, y después de quitar los *targets*, pero sin ningún tipo de tratamiento extra. Y cuando se hace referencia a los datos nuevos, se refiere a las variables escogidas de acuerdo a los resultados obtenidos en el atributo de salida *feature importances* del *random forest*.

1. La primera prueba se hará con los Datos originales.
2. La segunda prueba se realizara con los Datos originales elevados al cuadrado.
3. La tercera prueba se realizara con los datos Originales elevados al cubo.
4. La cuarta Prueba se realizara con datos transformados por medio del proceso de *Polynomial Features* con grado 2
5. La quinta prueba se realizara con datos transformados por medio del proceso de *Polynomial Features* pero esta vez con grado 3

6. Finalmente la sexta prueba se realizara con datos transformados por medio del proceso PCA y la cantidad de componentes que tendrá el PCA sera ajustado por medio de fuerza bruta.

Y ademas de esto se utilizo un atributo de salida del *Random Forest Regressor* llamado *Feature Importances* [Figura 15], el cual permitió reducir la cantidad de variables predictivas, dejando así solo las que mas influyeran en la predicción. Se escogieron solo aquellas variables que tuviesen al menos un 1 % de influencia en la predicción, las demás variables se desecharon. Este nuevo conjunto de datos se utilizo, solo en los algoritmos que se consideraron mejores después de realizarlos y analizarlos con el conjunto de datos original.

**Figura 15:** Feature Importances. A la izquierda el Año 2000 y a la derecha el año 2014. Se puede observar el porcentaje que influye cada una de las variables en la predicción.



## Regresión Lineal

Análisis para el año 2000 semestre 1:

**1. Datos Puros, o sin ningún tratamiento:**

	Kfolds	ShuffleSplit
Biología	5.10	4.43

**2. Datos elevados al cuadrado**

	Kfolds	ShuffleSplit
Biología	5.83	4.43

**3. Datos elevados al cubo**

	Kfolds	ShuffleSplit
Biología	5.87	4.46

**4. Datos transformados con PolynomialFeatures grado 2**

	Kfolds	ShuffleSplit
Biología	13.63	8.29

**5. Datos transformados con PolynomialFeatures grado 3**

	Kfolds	ShuffleSplit
Biología	8.80	8.84

**6. Datos transformados con PCA**

	ShuffleSplit
Biología	4.42

Los algoritmos que presentaron el mejor desempeño son (Todos fueron entrenados con datos tratados con ShuffleSplit):

**Tabla 1:** En esta tabla se puede ver las mejores implementaciones obtenidas con la regresión lineal en el año 2000 semestre 1

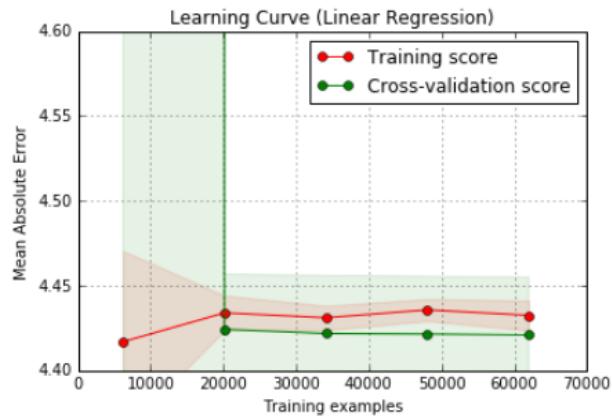
	Score/time	Score/time	Score/time	Score/time
Algoritmo	Normal	EXP 2	EXP 3	PCA 35 componentes
Datos Originales	4.43/3.68	4.43/2.73	4.46/2.93	4.43/2.31
Datos nuevos	4.46/2.05	4.43/2.58	4.48/2.48	4.44/2.21

Al escoger estos, y probarlos con los nuevos datos, se tiene un mejor panorama del rendimiento de cada algoritmo. Así entonces se escogió solo uno, que sera el algoritmo implementado a la hora de hacer la experimentación, y fue el siguiente:

**Tabla 2: Regresion lineal + Datos Originales + PCA de 35 Componentes + ShuffleSplit.** Seleccionado por su mejor desempeño frente a los demás. Ultimo Score Obtenido:

SCORE [MAE]	TIME [s]
4.43	2.31

**Figura 16:** Curva de aprendizaje Regresión Lineal año 2000. Muestra el comportamiento del modelo de regresión lineal escogido, a medida que se incrementa la cantidad de datos de entrenamiento



Análisis para el año 2014 semestre 2:

### 1. Datos Puros, o sin ningún tratamiento:

	Kfolds	ShuffleSplit
Biología	8.85	6.71

### 2. Datos elevados al cuadrado

	Kfolds	ShuffleSplit
Biología	8.88	6.75

### 3. Datos elevados al cubo

	Kfolds	ShuffleSplit
Biología	8.95	7.01

#### 4. Datos transformados con PolynomialFeatures grado 2

	Kfolds	ShuffleSplit
Biología	8.82	6.78

#### 5. Datos transformados con PolynomialFeatures grado 3

	Kfolds	ShuffleSplit
Biología	MEMORY ERROR	MEMORY ERROR

#### 6. Datos transformados con PCA

	ShuffleSplit
Biología	6.71

Los algoritmos que presentaron el mejor desempeño son (De nuevo todos fueron entrenados con datos tratados con *ShuffleSplit*):

**Tabla 3:** En esta tabla se puede ver las mejores implementaciones obtenidas con la regresión lineal en el año 2014 en el semestre 2

	Score/time	Score/time	Score/time	Score/time
Algoritmo	Normal	EXP 2	EXP 3	PCA 45 componentes
Datos Originales	6.71/12.60	6.75/12.21	6.78/654.14	6.71/21.14
Datos nuevos	6.73/9.97	6.77/11.03	6.58/155.09	ERROR

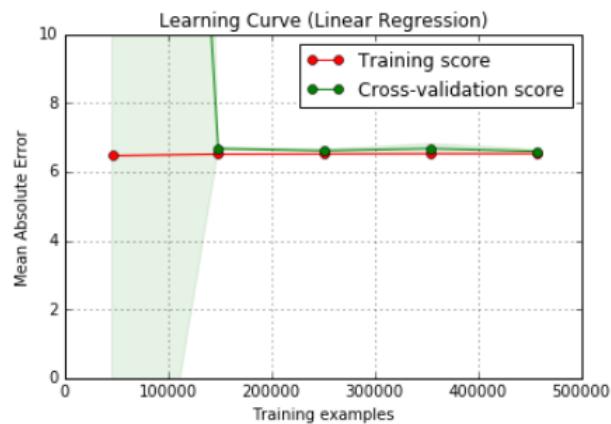
Se presento un error en el PCA con los datos nuevos porque el PCA requeria 45 componentes y el nuevo conjunto de datos tenia solo 27 variables.

Al escoger estos, y probarlos con los nuevos datos, se tiene un mejor panorama del rendimiento de cada algoritmo. Así entonces se escogió solo uno, que sera el algoritmo implementado a la hora de hacer la experimentación, y fue el siguiente:

**Tabla 4: Regresion Lineal + Nuevos datos elevados al cuadrado + *ShuffleSplit*.** Seleccionado por su mejor desempeño frente a los demás. Ultimo Score Obtenido:

SCORE [MAE]	TIME [s]
6.77	11.03

**Figura 17:** Curva de aprendizaje Regresión Lineal año 2014. Muestra el comportamiento del modelo de regresión lineal escogido, a medida que se incrementa la cantidad de datos de entrenamiento.



## Arboles de Decisión

Análisis para el año 2000 semestre 1:

### 1. Datos Puros, o sin ningún tratamiento:

	Kfolds	ShuffleSplit
Biología	8.56	9.04

### 2. Datos elevados al cuadrado

	Kfolds	ShuffleSplit
Biología	8.57	9.05

### 3. Datos elevados al cubo

	Kfolds	ShuffleSplit
Biología	8.56	9.06

### 4. Datos transformados con PolynomialFeatures grado 2

	Kfolds	ShuffleSplit
Biología	8.6	6.25

### 5. Datos transformados con PolynomialFeatures grado 3

	Kfolds	ShuffleSplit
Biología	7.57	6.26

### 6. Datos transformados con PCA

	ShuffleSplit
Biología	4.45

Los algoritmos que presentaron el mejor desempeño son (notese que todos fueron entrenados con datos tratados con *ShuffleSplit*):

**Tabla 5:** En esta tabla se puede ver las mejores implementaciones obtenidas con un arbol de decision en el año 2000 en el semestre 1

	Score/time	Score/time	Score/time
Algoritmo	Poly 2	Poly 3	PCA 33 Componentes
Datos Originales	6.25/355.02	6.26/14274.55	4.49/10.29
Datos nuevos	6.28/114.5	6.25/1324.16	ERROR

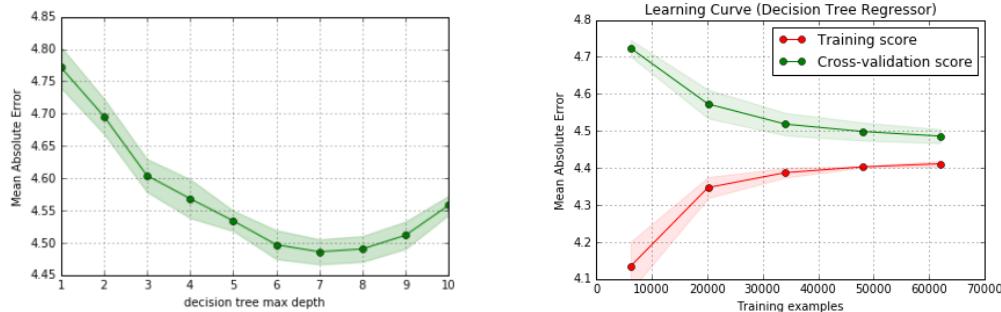
Se presento un error en el PCA con los datos nuevos porque el PCA requería 33 componentes y el nuevo conjunto de datos tenia solo 25 variables.

Al escoger estos, y probarlos con los nuevos datos, se tiene un mejor panorama del rendimiento de cada algoritmo. Así entonces se escogió solo uno, que sera el algoritmo implementado a la hora de hacer la experimentación, y fue el siguiente:

**Tabla 6: Arbol de decision con profundidad maxima de 7 + Datos Originales + PCA con 33 componentes + *ShuffleSplit*.** Seleccionado por su mejor desempeño frente a los demás. Ultimo Score Obtenido:

SCORE [MAE]	TIME [s]
4.49	11.74

**Figura 18:** A la izquierda el análisis de profundidad y a la derecha la curva de aprendizaje obtenida con el modelo de árbol de decisión escogido para el año 2000



Análisis para el año 2014 semestre 2:

**1. Datos Puros, o sin ningún tratamiento:**

	Kfolds	ShuffleSplit
Biología	11.45	9.43

**2. Datos elevados al cuadrado**

	Kfolds	ShuffleSplit
Biología	11.45	9.43

**3. Datos elevados al cubo**

	Kfolds	ShuffleSplit
Biología	11.45	9.44

**4. Datos transformados con PolynomialFeatures grado 2**

	Kfolds	shufflesplit
Biología	11.45	9.45

**5. Datos transformados con PolynomialFeatures grado 3**

	Kfolds	Shufflesplit
Biología	ERROR	ERROR

**6. Datos transformados con PCA**

	Shufflesplit
Biología	7.78

Los algoritmos que presentaron el mejor desempeño son (note se que todos fueron entrenados con datos tratados con *ShuffleSplit*):

**Tabla 7:** En esta tabla se puede ver las mejores implementaciones obtenidas con un árbol de decisión en el año 2014 en el semestre 2

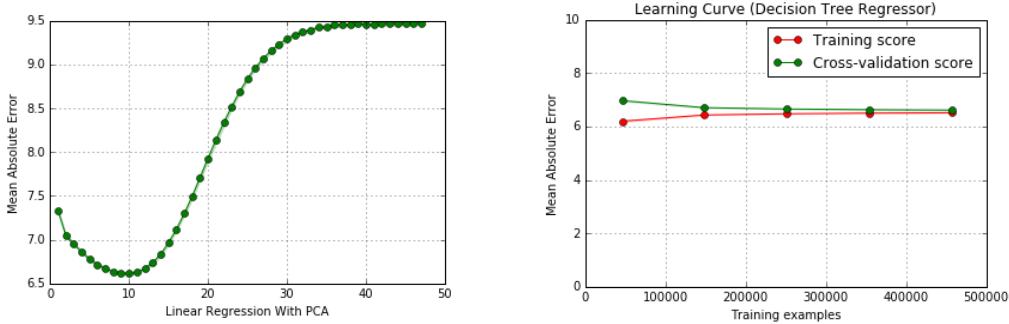
	Score/time	Score/time	Score/time	Score/time
Algoritmo	Normal	EXP 2	EXP 3	PCA 1 componente
Datos Originales	9.43/227.72	9.43/239.01	9.44/252.01	7.78/46.01
Datos nuevos	9.46/57.17	9.45/108.65	9.47/53.51	7.79/15.19

Al escoger estos, y probarlos con los nuevos datos, se tiene un mejor panorama del rendimiento de cada algoritmo. Así entonces se escogió solo uno, que sera el algoritmo implementado a la hora de hacer la experimentación, pero en este caso se noto una mejora particular al fijar la profundidad del árbol y supero en precisión a los demás.

**Tabla 8: Árbol de decisión con profundidad máxima de 10 + Datos Originales + *ShuffleSplit*.** Seleccionado por su mejor desempeño frente a los demás. Ultimo Score Obtenido:

SCORE [MAE]	TIME [s]
6.61	32.92

**Figura 19:** A la izquierda el análisis de profundidad y a la derecha la curva de aprendizaje obtenida con el modelo de árbol de decisión escogido para el año 2014



## Bosques Aleatorios

Análisis para el año 2000 semestre 1:

### 1. Datos Puros, o sin ningún tratamiento:

	Kfolds	ShuffleSplit
Biología	6.002	4.53

### 2. Datos elevados al cuadrado

	Kfolds	ShuffleSplit
Biología	6.005	4.52

### 3. Datos elevados al cubo

	Kfolds	ShuffleSplit
Biología	6.01	4.53

### 4. Datos transformados con PolynomialFeatures grado 2

	Kfolds	ShuffleSplit
Biología	5.99	4.51

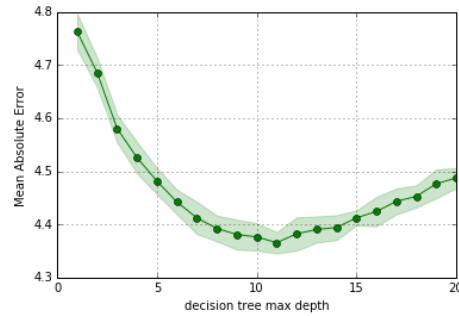
### 5. Datos transformados con PolynomialFeatures grado 3

	No Validacion Cruzada
Biología	5.47

## 6. Datos transformados con PCA

	ShuffleSplit
Biología	4.43

**Figura 20:** Análisis de profundidad obtenida con el modelo de árbol de decisión.



Los algoritmos que presentaron el mejor desempeño son (note se que todos fueron entrenados con datos tratados con *shuffleSplit*). Ademas se realizo el análisis de profundidad al *random forest* y se hicieron otras pruebas con los mejores algoritmos (y sus respectivos conjuntos de datos) y se añadieron los resultados a la tabla.

**Tabla 9:** En esta tabla se puede ver las mejores implementaciones obtenidas con un bosque aleatorio en el año 2000 en el semestre 1

	Score/time	Score/time	Score/time
Algoritmo	EXP 2	Poly 2	PCA 33 Componentes
Datos Originales	4.53/44.38	4.51/1417.17	4.43/23.58
Datos nuevos	4.52/35.6	4.54/703.7	ERROR
Profundidad 11 niveles	4.33/8.97	4.34/134.22	4.37/37.48

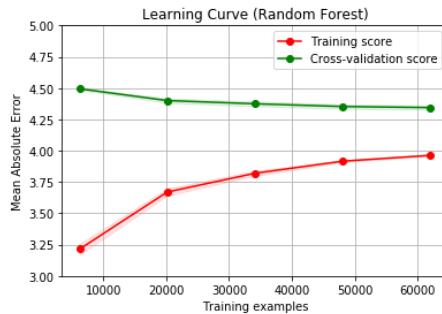
Se presento un error en el PCA con los datos nuevos porque el PCA requería 33 componentes y el nuevo conjunto de datos tenia solo 25 variables.

Al escoger estos, y probarlos con los nuevos datos, se tiene un mejor panorama del rendimiento de cada algoritmo. Así entonces se escogió solo uno, que sera el algoritmo implementado a la hora de hacer la experimentación, y fue el siguiente:

**Tabla 10: Bosque aleatorio de regresión con profundidad máxima 11 + Nuevos datos elevados al cuadrado + *ShuffleSplit*.** Seleccionado por su mejor desempeño frente a los demás. Ultimo Score Obtenido:

SCORE [MAE]	TIME [s]
4.33	8.97

**Figura 21:** Curva de aprendizaje obtenida con el modelo de árbol de decisión escogido para el año 2014



Análisis para el año 2014 semestre 2:

### 1. Datos Puros, o sin ningún tratamiento:

	Kfolds	ShuffleSplit
Biología	9.18	6.83

### 2. Datos elevados al cuadrado

	Kfolds	ShuffleSplit
Biología	9.18	6.84

### 3. Datos elevados al cubo

	Kfolds	ShuffleSplit
Biología	9.18	6.84

### 4. Datos transformados con PolynomialFeatures grado 2

	No validación cruzada
Biología	8.01

### 5. Datos transformados con PolynomialFeatures grado 3

	No Validación Cruzada
Biología	Toma demasiado tiempo

### 6. Datos transformados con PCA de 1 componente

	ShuffleSplit
Biología	7.78

Los algoritmos que presentaron el mejor desempeño son (note se que todos fueron entrenados con datos tratados con *shuffleSplit*).

**Tabla 11:** En esta tabla se puede ver las mejores implementaciones obtenidas con un bosque aleatorio en el año 2014 en el semestre 2

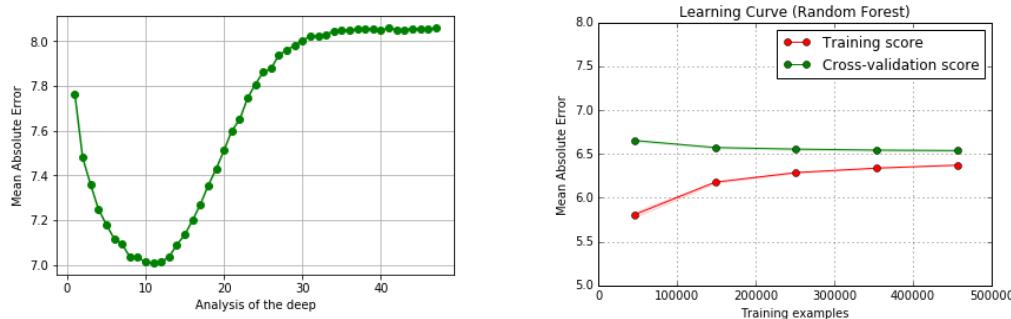
	Score/time	Score/time	Score/time	Score/time
Algoritmo	Normal	EXP 2	EXP 3	PCA 1 Componentes
Datos Originales	6.83/658.22	6.84/519.26	6.84/255.02	7.93/41.8
Datos nuevos	6.87/146.4	6.86/183.36	6.86/173.18	7.94/14.35

Al escoger estos, y probarlos con los nuevos datos, se tiene un mejor panorama del rendimiento de cada algoritmo. Ademas se realizo el análisis de la profundidad del *random forest* y se hicieron otras pruebas con los mejores algoritmos (y sus respectivos conjuntos de datos) y se escogió el mejor.

**Tabla 12: Bosque aleatorio de regresión con profundidad máxima 11 + Datos originales + *ShuffleSplit*.** Seleccionado por su mejor desempeño frente a los demás. Ultimo *Score* Obtenido:

SCORE [MAE]	TIME [s]
6.52	96.48

**Figura 22:** A la izquierda el análisis de profundidad y a la derecha la curva de aprendizaje obtenida con el modelo de bosque aleatorio escogido para el año 2014



#### 4.4.3. algoritmos seleccionados .

1. Para el año 2000 en el semestre 1:
  - a) Regresion lineal + datos originales + PCA de 35 componentes + *ShuffleSplit*
  - b) Arbol de decision con profundidad maxima de 6 + datos originales + PCA de 33 componentes + *ShuffleSplit*

- c) Bosque aleatorio con profundidad maxima de 11 + Datos nuevos elevados al cuadrado + *ShuffleSplit*

**Tabla 13:** Scores obtenidos en las pruebas para el año 2000 en el semestre 1, por cada uno de los modelos seleccionados anteriormente.

Regresión Lineal		Árbol de Decisión		Bosque Aleatorio	
SCORE [MAE]	TIME [s]	SCORE [MAE]	TIME [s]	SCORE [MAE]	TIME [s]
4.43	2.31	4.49	11.74	4.33	8.97

2. Para el año 2014 en el semestre 2:

- a) Regresión lineal + Nuevos datos elevados al cuadrado + *ShuffleSplit*
- b) Árbol de decisión con profundidad máxima de 10 + Datos originales + *ShuffleSplit*
- c) Bosque aleatorio con profundidad maxima de 11 + Datos originales + *ShuffleSplit*

**Tabla 14:** Scores obtenidos en las pruebas para el año 2014 en el semestre 2, por cada uno de los algoritmos listados anteriormente.

Regresión Lineal		Árbol de Decisión		Bosque Aleatorio	
SCORE [MAE]	TIME [s]	SCORE [MAE]	TIME [s]	SCORE [MAE]	TIME [s]
6.77	11.03	6.61	32.92	6.52	96.48

#### 4.5. EXPERIMENTOS FINALES

Finalmente, con los algoritmos propuestos y las maneras de tratamiento de datos escogidas, se realizaran los experimentos con todos los demás archivos, de esta manera se obtendrán resultados que servirán para concluir que tan buenos o malos se comportan

los modelos obtenidos. Todo este proceso se debe realizar dos veces, debido al cambio que implemento el ICFES en el examen SABER11 en el segundo semestre del año 2014.

Se realizaron dos tipos de experimentos:

- **Predicción individual:** El primer experimento consistió en entrenar y validar los modelos con los datos del mismo archivo, con ayuda de los procesos de *cross-validation* y *shufflesplit*, tal y como se realizaron las pruebas. Ejemplo: se entrena el modelo con los datos del año 2000 semestre 1 y se valida con ayuda del *cross-validation* y el *shuffle-split* con los mismos datos del año 2000 semestre 1.
- **Predicción inter-semestral:** El segundo experimento consistió en entrenar y validar los modelos con datos de archivos distintos, en otras palabras, el modelo se entrenara con la totalidad de datos de un semestre y se validara con la totalidad de datos del siguiente semestre. Ejemplo: se entrena el modelo con los datos del primer semestre del año 2000 y se valida el modelo con los datos del segundo semestre del año 2000

El segundo experimento se considera mas cercano a una situación real de uso, por lo tanto se le dará mas de protagonismo en los resultados finales. Las predicciones se hicieron cuando fueron posibles, ya que por los cambios tan frecuentes que poseen los archivos como se vio en la [Figura 4], se imposibilito la implementación de los modelos en algunos casos.

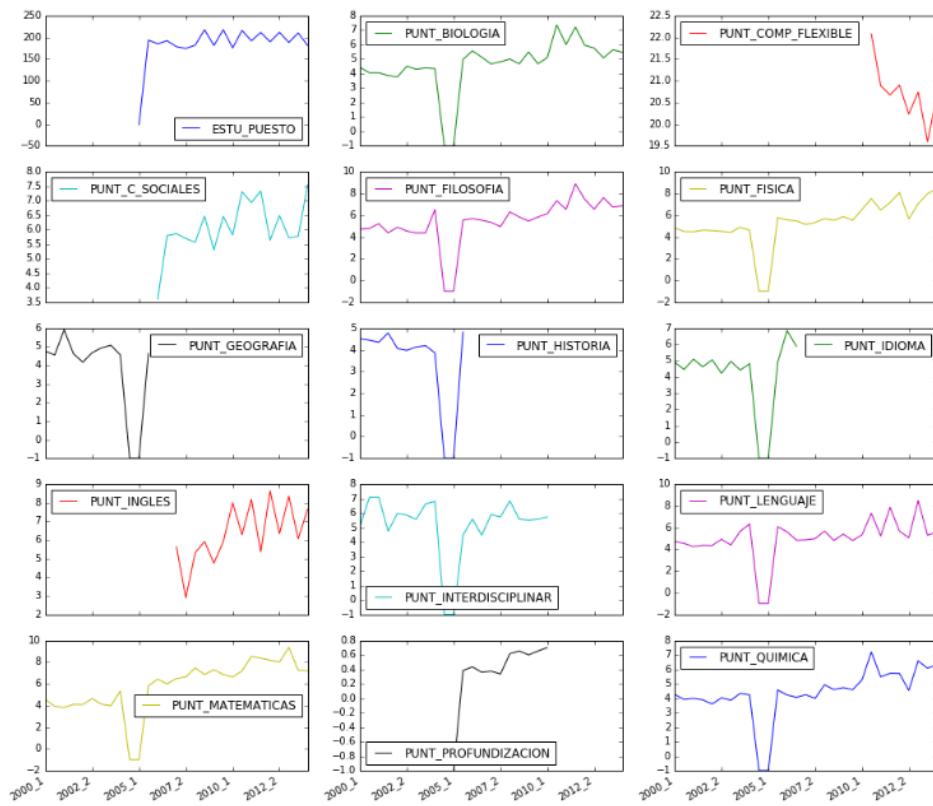
### **Predicción Individual:**

Se pueden apreciar las predicciones obtenidas a lo largo de los primeros 14 años con los tres modelos, regresión lineal[Figura 23], árbol de decisión[Figura 24] y bosque aleatorio[Figura 25], en ese orden.

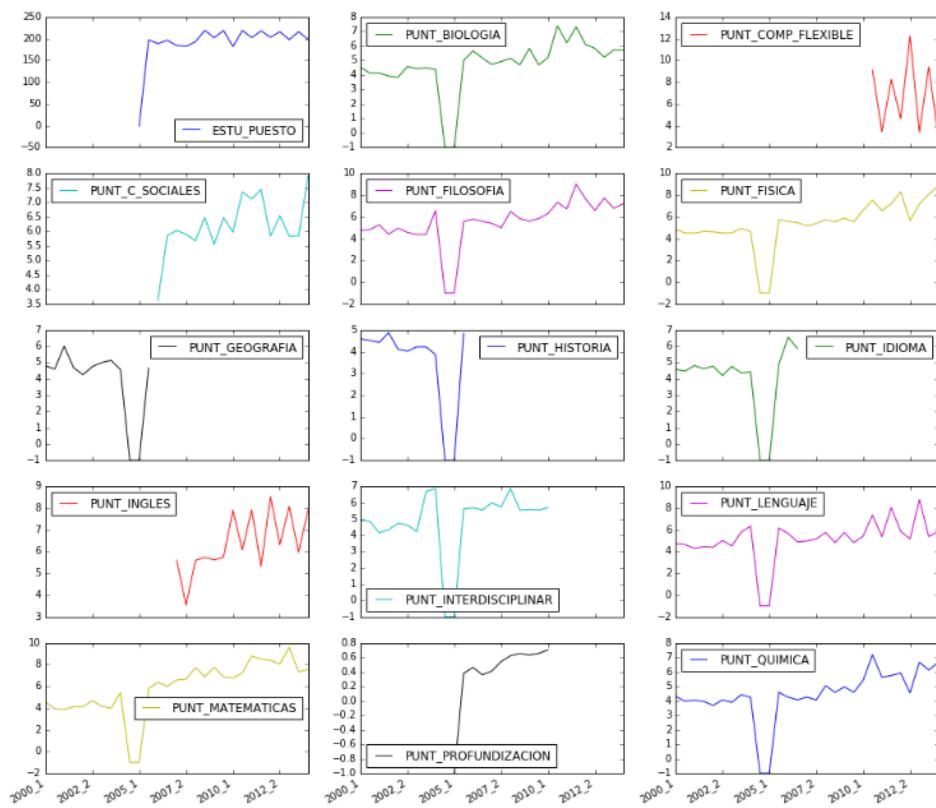
### **Predicción Individual años posteriores al 2014 semestre 2**

Para la segunda parte de la predicción individual se pueden apreciar las predicciones

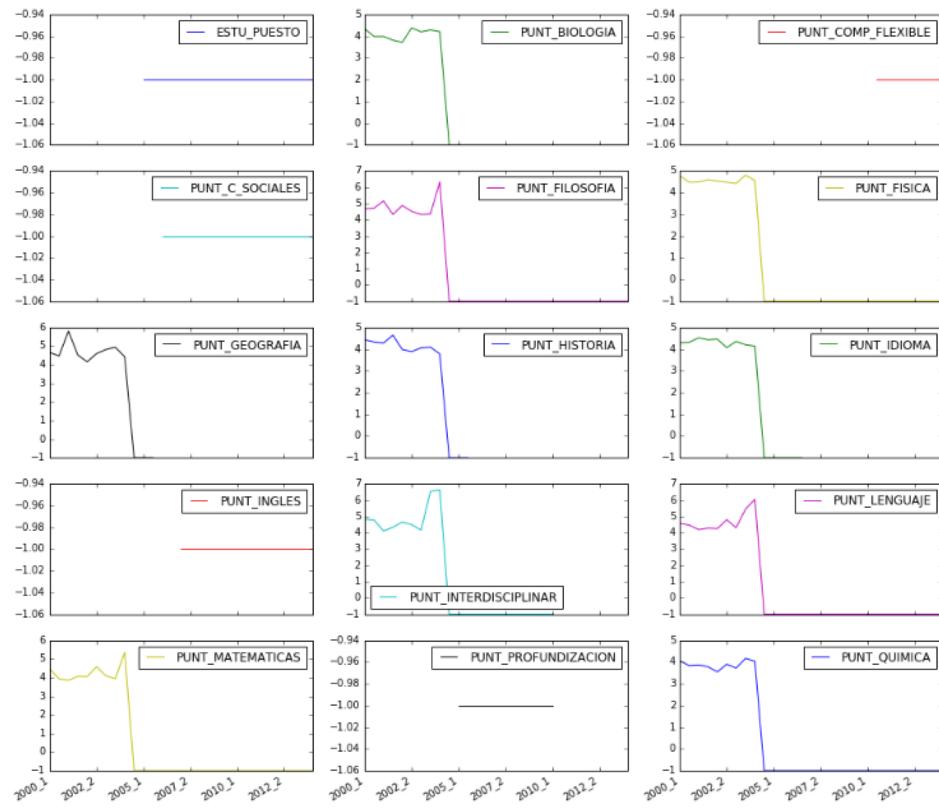
**Figura 23:** Scores obtenidos con el modelo de regresión lineal para los primeros 14 años



**Figura 24:** Scores obtenidos con el modelo de árbol de decisión para los primeros 14 años

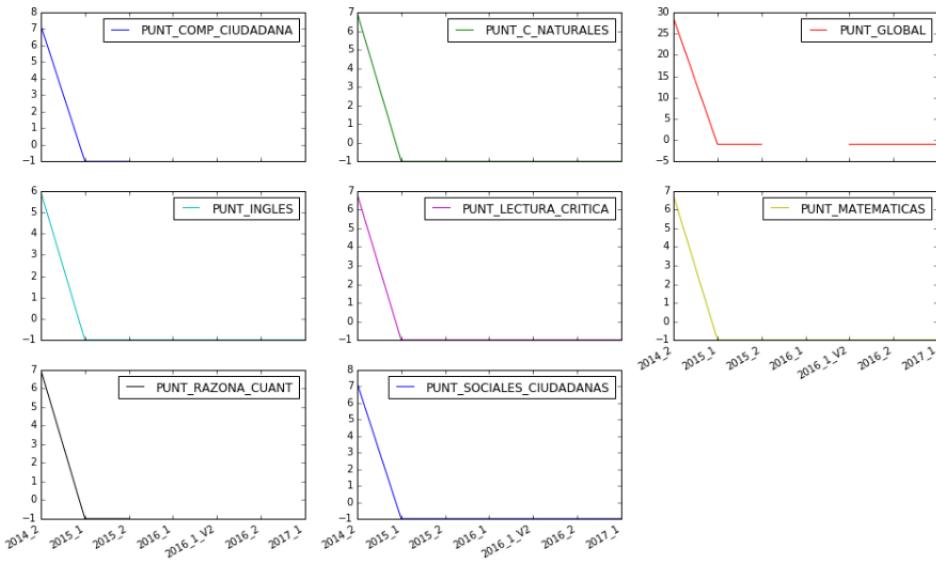


**Figura 25:** Scores obtenidos con el modelo del Bosque Aleatorio para los primeros 14 años



obtenidas a lo largo de los años posteriores al 2014 semestre 2 con los tres modelos, regresión lineal[Figura 26], árbol de decisión[Figura 27] y bosque aleatorio[Figura 28], en ese orden.

**Figura 26:** Scores obtenidos con el modelo de regresión lineal para los años posteriores al 2014 semestre 2



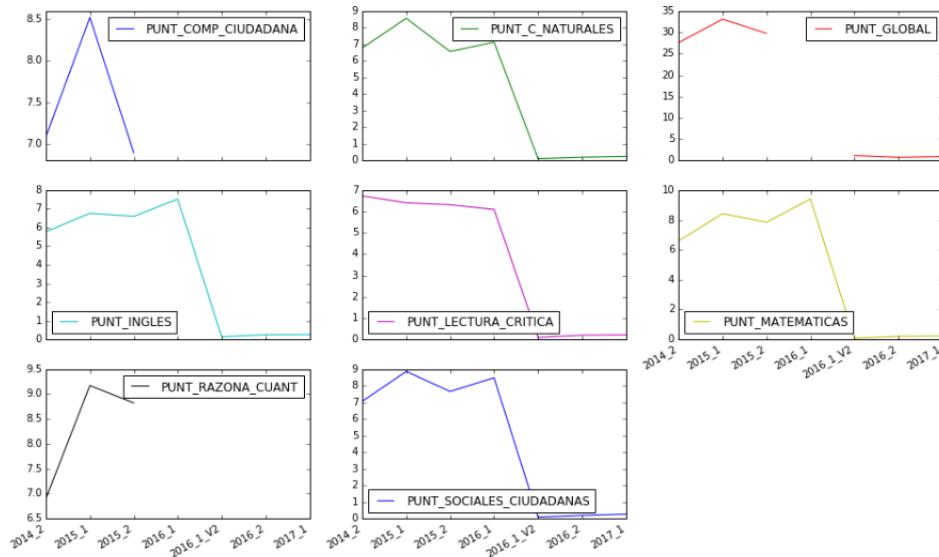
### Prediccion Inter-Semestral primeros 14 años:

Se pueden apreciar las predicciones obtenidas a lo largo de los primeros 14 años con los tres modelos, regresión lineal[Figura 29], árbol de decisión[Figura 30] y bosque aleatorio[Figura 31], en ese orden. Entrenados con un archivo y validados con el siguiente, por esta razón el año 2000 semestre 1 no aparece en las figuras.

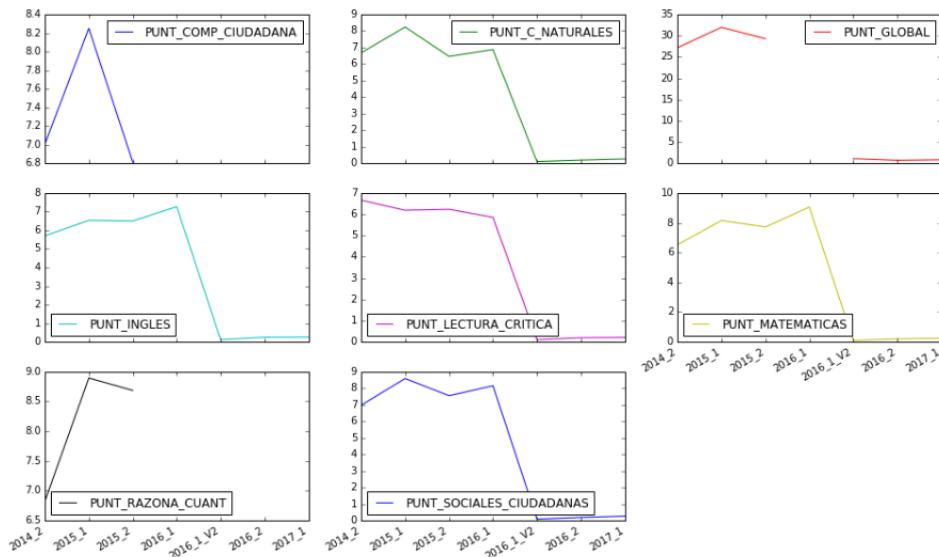
### Predicción inter-semestral años posteriores al 2014 semestre 2

Para la segunda parte de la predicción inter-semestral se pueden apreciar las predicciones obtenidas a lo largo de los años posteriores al 2014 con los tres modelos, regresión lineal[Figura 32], árbol de decisión[Figura 33] y bosque aleatorio[Figura 34], en ese orden. Entrenados con un archivo y validados con el siguiente, por esta razón el año 2014 semestre 2 no aparece en las figuras.

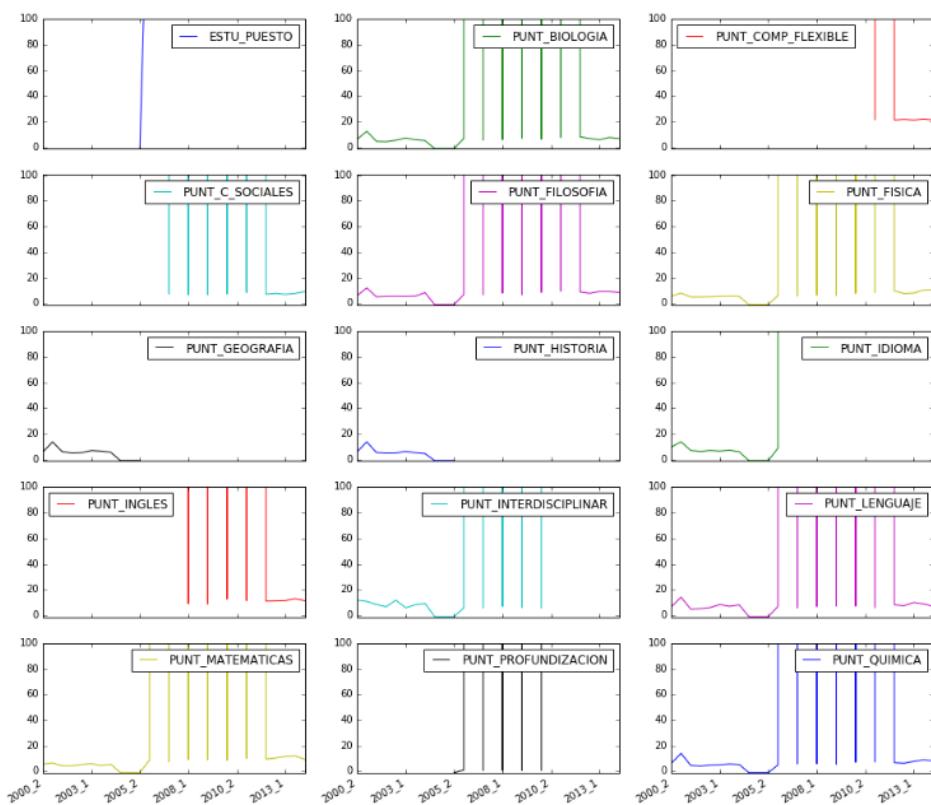
**Figura 27:** Scores obtenidos con el modelo de árbol de decisión para los años posteriores al 2014 semestre 2



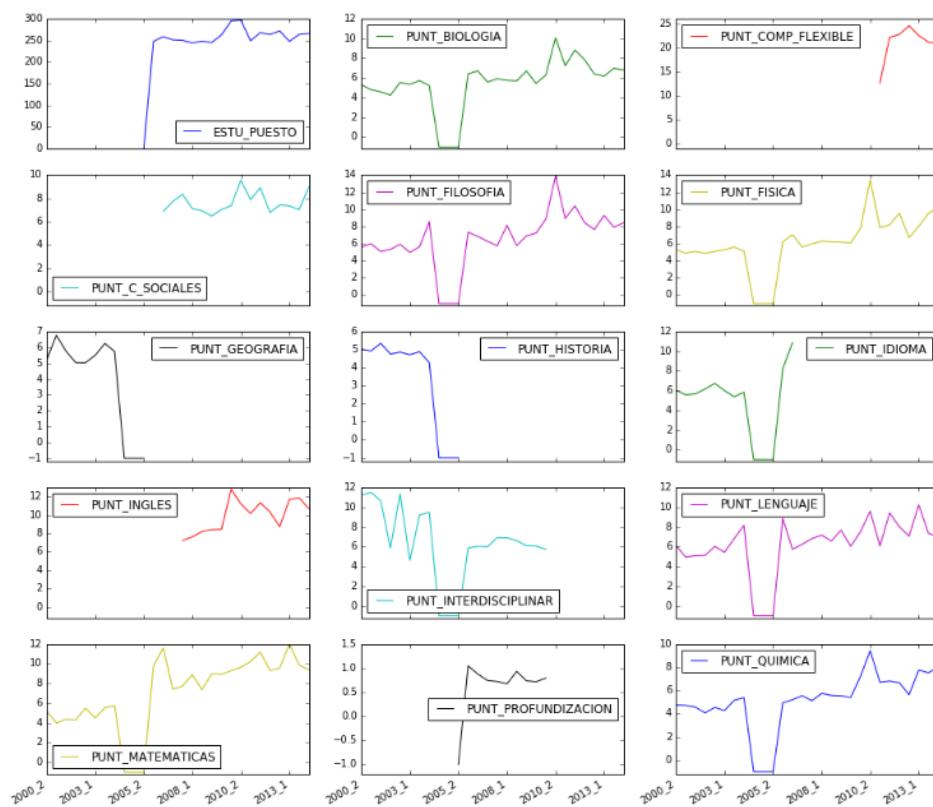
**Figura 28:** Scores obtenidos con el modelo del Bosque Aleatorio para los años posteriores al 2014 semestre 2



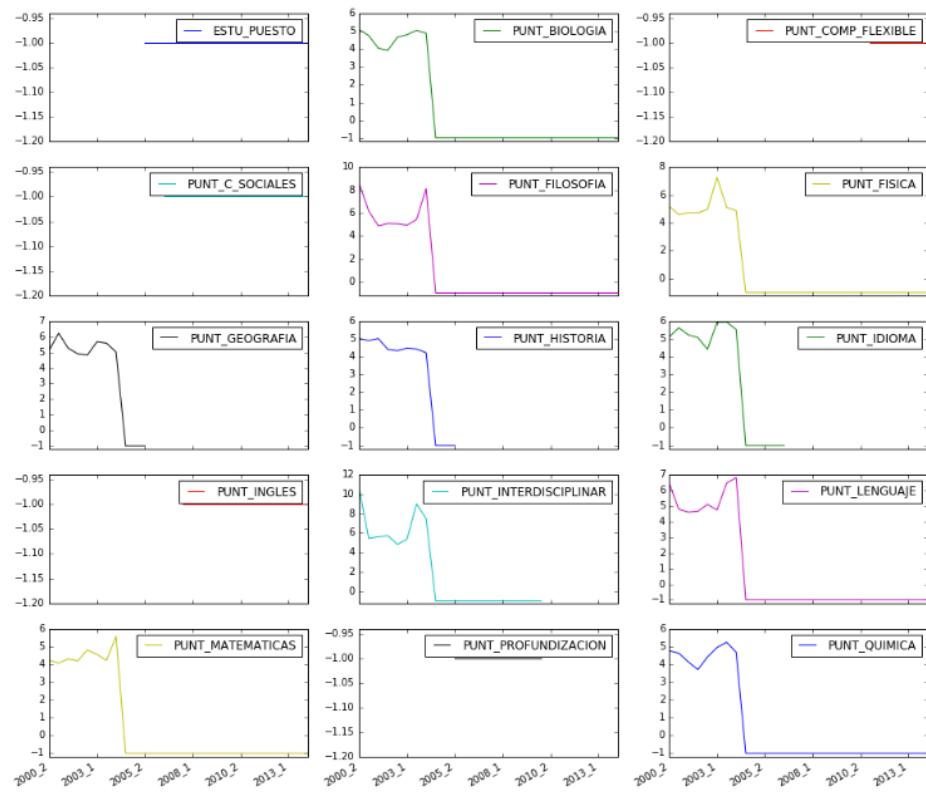
**Figura 29:** Scores obtenidos con el modelo de regresión lineal para los primeros 14 años



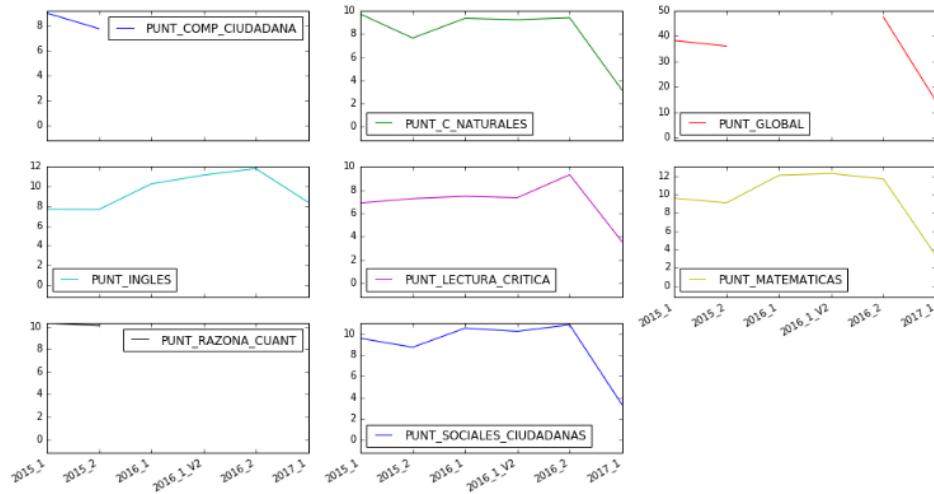
**Figura 30:** Scores obtenidos con el modelo de árbol de decisión para los primeros 14 años



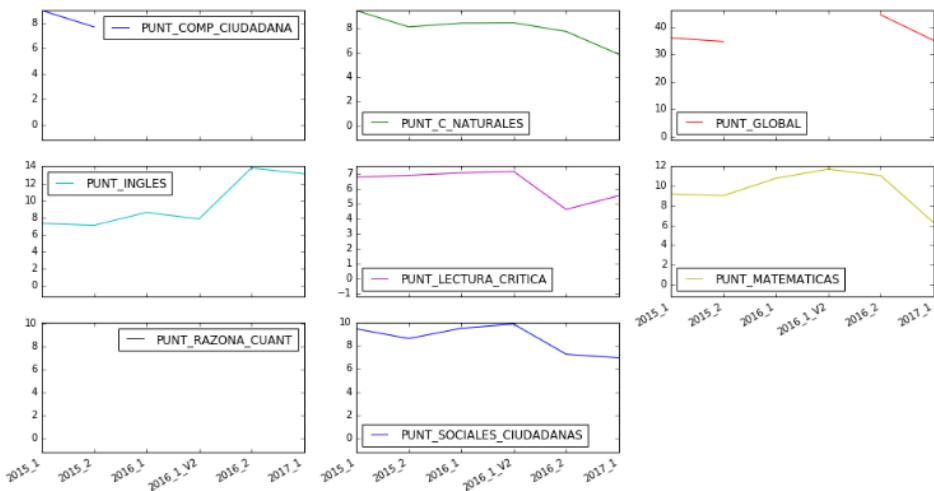
**Figura 31:** Scores obtenidos con el modelo de árbol de decisión para los primeros 14 años



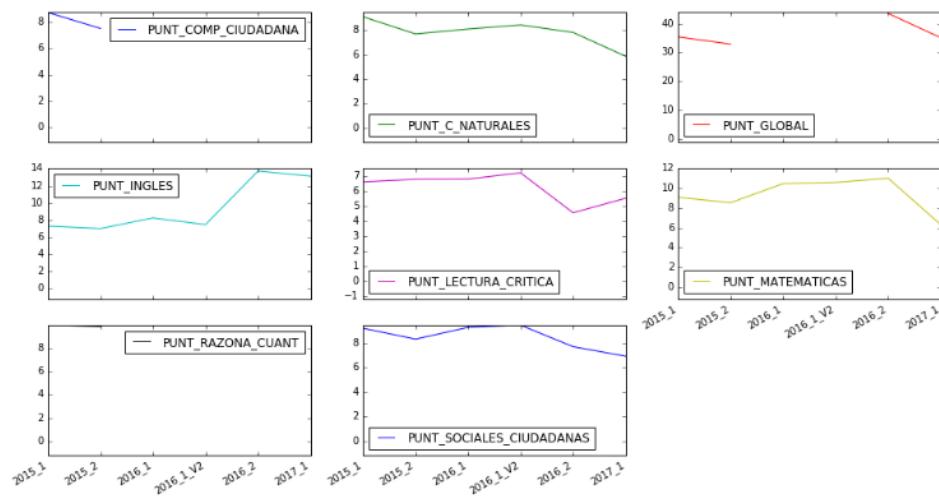
**Figura 32:** Scores obtenidos con el modelo de regresión lineal para los años posteriores al año 2014



**Figura 33:** Scores obtenidos con el modelo de regresión lineal para los años posteriores al año 2014



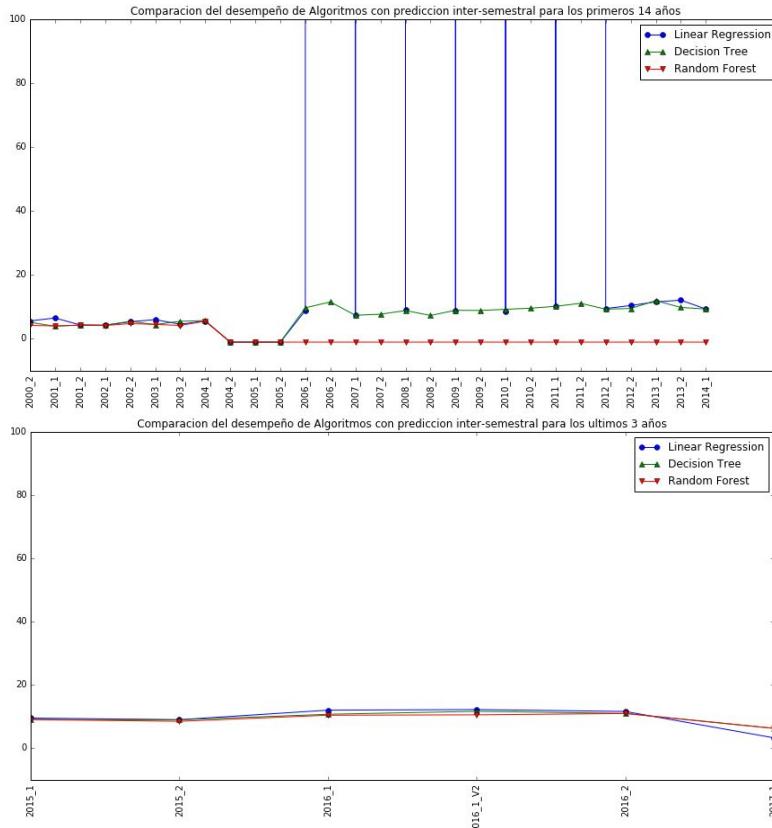
**Figura 34:** Scores obtenidos con el modelo de regresión lineal para los años posteriores al año 2014



## 5. EVALUACION DE RESULTADOS

- El catalogo de datos se realizo y ayudo a ver mas claramente los datos de los que se disponían para la investigación, y cuales eran mas factibles para su posterior uso de acuerdo al cronograma del proyecto.
- Con los datos escogidos en el catalogo, se realizaron las divisiones o segmentaciones de acuerdo a dos ámbitos, uno socio-económico por cantidad de salarios mínimos mensuales legales vigentes que ingresaban al capital familiar y otro geográfico por departamento de residencia.
- Se obtuvo información sobre la influencia de las variables en la predicción [Figura 15] y se evidencio que el departamento en donde viva el estudiante no tiene una influencia significativa en la predicción, pero variables económicas como los ingresos familiares, nivel de sisben entre otras, si afectan en mayor medida en el desempeño del estudiante.
- Se escogió una métrica de error para medir los resultados de los modelos, específicamente se escogió el MAE *Mean Absolute Error* o Error Absoluto Medio, ya que esta medida estará en las mismas unidades que la variable a predecir, y el tiempo se midió en segundos.
- Se plantearon 2 tipos de tareas de *Machine Learning* para predecir el puntaje en cada área del conocimiento, utilizando estrategias de tratamiento de datos como PCA y *Polynomial Features*, y los 3 algoritmos de regresión escogidos (regresión lineal, arboles de regresión y bosques aleatorios para regresión).
- Las tareas planteadas se implementaron y los resultados que se obtuvieron se pueden apreciar en las figuras 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, y 34, que muestran los resultados de los modelos que obtuvieron los mejores resultados en los experimentos.
- Se compararon los resultados obtenidos de cada modelo con el experimento de predicción inter-semestral y se pudo evidenciar que la regresión lineal genero errores demasiado grandes, por encima de 100 en varios años, y el modelo del bosque

**Figura 35:** Comparacion de desempeños, predicción inter-semestral para matemáticas. Se puede ver que el algoritmo mas estable o que mejor se comporta a lo largo de todos los años es el árbol de decisión o *decision tree*. Y se puede evidenciar los pésimos desempeños de la regresión lineal en varios años a partir del 2006 semestre 1 y el bosque aleatorio después del año 2004 semestre 2.



aleatorio no se pudo implementar en varios años porque no cumplía con los requisitos de datos de los modelos. Por otro lado el árbol de decisión fue el que mejores resultados dio al haberse podido implementar en la mayoría de años y obteniendo resultados aceptables [Figura 35].

## **6. CONCLUSIONES**

La herramienta propuesta funciona de manera aceptable con uno de los modelos propuestos, pero no se recomienda su uso en un ámbito mas real y/o comercial, ya que no se considera que sea lo suficientemente bueno porque existen otros aspectos que se pueden llegar a relacionar y afectar el rendimiento académico de los estudiantes.

En cuanto al desempeño, el modelo de regresión lineal y los bosques aleatorios fueron los peores de los 3 utilizados, ya que para los datos de un mismo semestre el modelo funcionaba bien, pero en un escenario un poco mas real, que es validando el modelo con los datos del siguiente semestre, se obtiene un desempeño muy bajo. Estos bajos desempeños se deben a la gran variación en la cantidad de columnas que poseen los archivos. Los arboles de decisión obtuvieron desempeños aceptables en los dos tipos de experimentos realizados, con un error promedio de 7.13 para la validación intersemestral, lo cual se sitúa por debajo de 10 puntos de error y es considerado un error bajo.

Ademas se pudo observar que variables influían un poco mas en el desempeño de los estudiantes, gracias al análisis de características realizado. Las variables que mas influyen son de tipo económico tales como los ingresos familiares, el nivel del sisben que tiene que ver con la pobreza de la familia, el valor de la pensión del colegio; y otras variables como la edad del estudiante, la carrera deseada, la universidad deseada, la educación de los padres y en menor medida la región geográfica en donde residen.

Los datos proporcionados por el ICFES son muy variantes y poseen muchos vacíos y errores, si se estandariza y se mejora la recolección de datos para los años siguientes se puede llegar a conseguir mejores resultados con herramientas como la propuesta en esta investigación en trabajos futuros.

## **7. PERSPECTIVAS**

Para trabajos futuros se recomienda utilizar datos de nuevos años que proporcione el ICFES, y crear una lista estándar de variables influyentes para mejorar el desempeño en los modelos propuestos. Se propone relacionar otros datos que puedan influir en el rendimiento académico, proporcionados por el ICFES como información socio-económica de los planteles estudiantiles o los resultados de pruebas como la prueba SABER 5 o la prueba SABER 9.

Se propone evaluar nuevos modelos que incluyan algoritmos como el *Gradient Boosting* el cual es uno de los mas populares, he incluso modelos de *Deep Learning* y hacer uso de redes neuronales enfocadas a la regresión.

## BIBLIOGRAFIA

BAGNATO, Juan I. Regresión Lineal en Español con Python, [En Linea], 13 de Mayo de 2018 [Revisado Junio de 2018]. Disponible en Internet: <<http://www.aprendemachinelearning.com/regresion-lineal-en-espanol-con-python/#more-5722>>

BOWLES, Michael. Machine Learning in Python. Essential techniques for predictive analysis. Wiley, 2005.

COLEMAN, J. et al. 1996 Equality of Educational Opportunity. Washington: US Government Printing Office. [En Linea] <<http://library.sc.edu/digital/collections/eeoci.pdf>>

CONWAY,Drew; MYLES,John. Machine Learning for Hackers. Estados Unidos de América: O'Reilly Media, Inc.:2012

GOMILLA,Juan. Curso completo de Machine Learning: Data Science en Python. Udemy [En Linea] <<https://www.udemy.com/machinelearningpython/>>

ICFES, Instituto Colombiano para la Evaluación de la Educación. Determinantes del desempeño académico universitario. El caso de la Región Caribe colombiana (2014). [En Linea] <<http://www.icfes.gov.co/docman/investigadores-y-estudiantes-de-posgrado/resultados-de-investigaciones/equidad/989-determinantes-del-desempeno-academico-universitario-el-caso-region-caribe-colombiana>>

ICFES, Instituto Colombiano para la Evaluación de la Educación. [En Linea]. 21 de Junio de 2017. SABER11. Disponible en Internet: <<ftp://ftp.icfes.gov.co/>>

LA LLAVE. Estándares, evaluación y mejoramiento. [En Linea] <http://www.mineducacion.gov.co/1621/article-87448.html>>

LEY 715 de 2001. Titulo II: Sector Educación, Capítulo III: De las instituciones educativas, los rectores y los recursos. Artículo 9 y Artículo 10.4 [En Linea] <http://www>.

[mineducacion.gov.co/1621/articles-86098\\_archivo\\_pdf.pdf](https://mineducacion.gov.co/1621/articles-86098_archivo_pdf.pdf)

ROUSE, Margaret. ¿Que es Análisis de Datos?, en SearchDataCenter en español. [En Linea]. Disponible en Internet: <<https://searchdatacenter.techtarget.com/es/definicion/Analisis-de-Datos>>

SCIKIT-LEARN Developers. sklearn.decomposition.PCA [En Linea], [Revisado 26 Julio 2018]. Disponible en internet <<http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>>

SCIKIT-LEARN Developers. sklearn.preprocessing.PolynomialFeatures [En Linea], [Revisado 26 Julio 2018]. Disponible en internet <<http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html>>

SISTEMA DE APOYO PARA LA ACREDITACIÓN DE LA CALIDAD DE PROGRAMAS ACADEMICOS DE LA UNIVERSIDAD DE CALDAS, APLICANDO TÉCNICAS EN MINERÍA DE DATOS. [En Linea] <[http://repositorio.autonoma.edu.co/jspui/bitstream/11182/350/1/Msc.GyDlloSoft\\_InformeFinal\\_JuanCarlosGonzalez.pdf](http://repositorio.autonoma.edu.co/jspui/bitstream/11182/350/1/Msc.GyDlloSoft_InformeFinal_JuanCarlosGonzalez.pdf)>

WIKIPEDIA[Anónimo]. Regresión Lineal [En Linea]. Disponible en Internet: <[https://es.wikipedia.org/wiki/Regresi%C3%B3n\\_lineal](https://es.wikipedia.org/wiki/Regresi%C3%B3n_lineal)>

WIKIPEDIA[Anónimo]. Decision Tree [En Linea]. Disponible en Internet: <[https://es.wikipedia.org/wiki/%C3%81rbol\\_de\\_decisi%C3%B3n](https://es.wikipedia.org/wiki/%C3%81rbol_de_decisi%C3%B3n)>

WIKIPEDIA[Anónimo]. Random Forest [En Linea]. Disponible en Internet: <[https://es.wikipedia.org/wiki/Random\\_forest](https://es.wikipedia.org/wiki/Random_forest)>