Predecir si un aplicante va a las pruebas Saber 11 perderá el examen

*Paola Zárate,* [*paolazaratem@gmail.com*](mailto:paolazaratem@gmail.com), Politécnico Grancolombiano Sede Bogotá

**Abstract** El propósito de este artículo es informar el proceso realizado para predecir cuando una persona va a perder el examen de las pruebas Saber 11, los procesos realizados son: revisión de datos, selección de características y se probaron diferentes metodos de predicción como lo son Naive Bayes, regresión lineal y arboles de decisión.

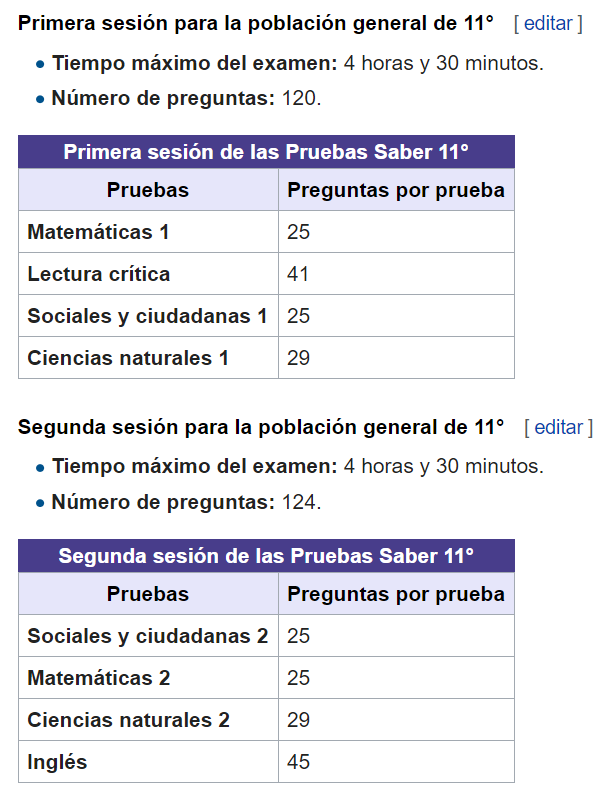
***Index Terms*—mineria de datos, prediccion,datos**

# INTRODUCCION[[1]](#footnote-0)

En 1968 se creo en Colombia el Instituto Colombiano para el Fomento de la Educación Superior (ICFES), encargado de realizar los servicios de evaluación de la educación en todos los niveles, además realiza estudios sobre los factores que inciden en la calidad de la educación.  
  
En 1980, se reglamentaron los exámenes de Estado pruebas Icfes para el ingreso a la educación superior hasta ese año la presentación del examen era voluntaria y a partir de ese año se convirtió en un requisito que cualquier estudiante que quisiera ingresar a la educación superior.  
  
En el 2009, se decidió cambiar el nombre del examen a pruebas saber y se cambio el enfoque del examen pues a partir de este momento se evalúa matemáticas, lectura critica, sociales y ciudadanías y ciencias naturales.

# PROBLEMA

El problema a resolver es identificar las características que inciden para que un aplican te a las pruebas saber 11 pierda el examen, para esto es importante tener claro que el examen consta de dos sesiones y las siguientes pruebas:



Tomado de [1]

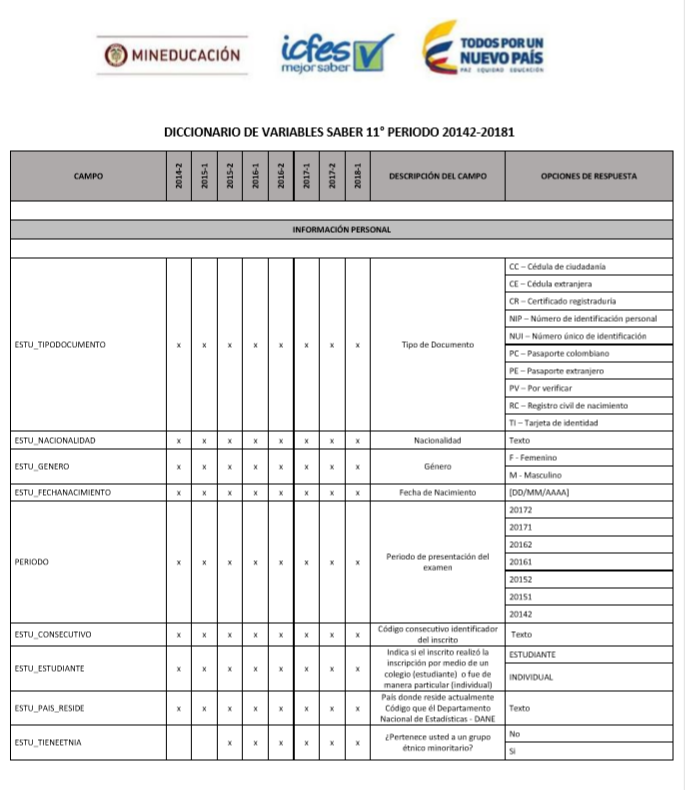
El ICFES considera: puntaje Muy bajo entre 0 y 30 puntos en cada prueba del examen o bajo entre 31 puntos y 50 puntos, como lo muestra la siguiente tabla:



Tomado de [2]

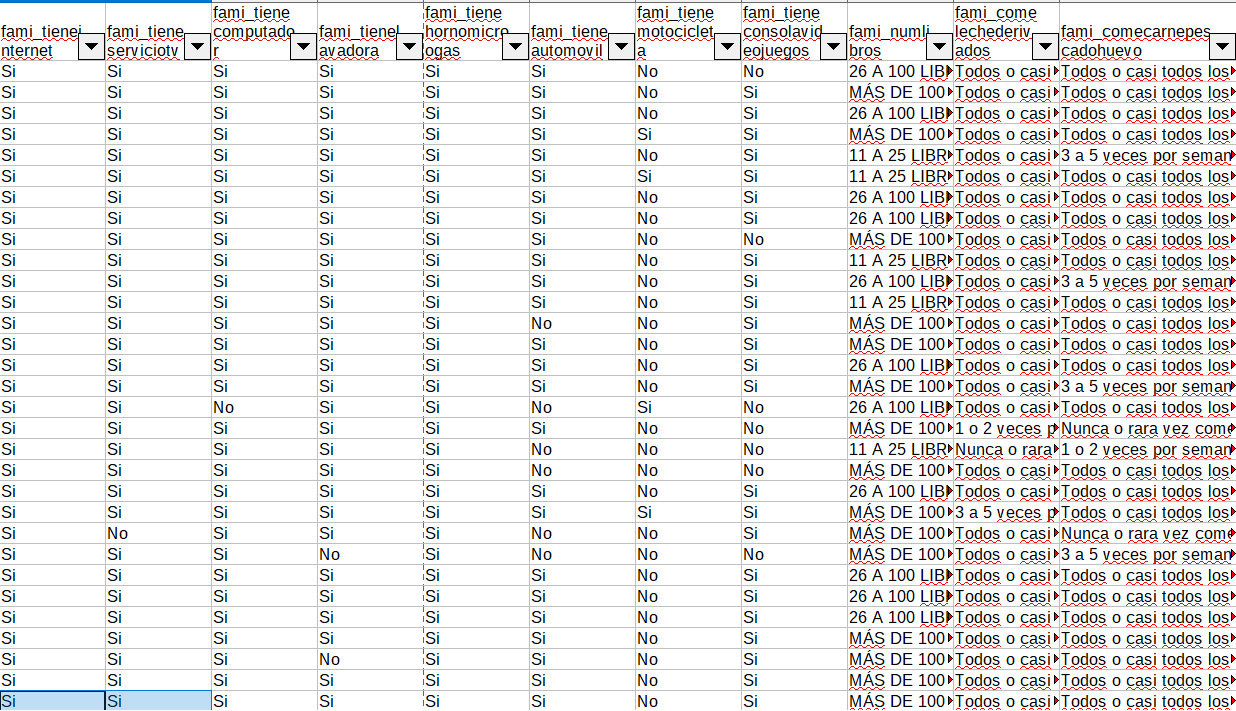
# Datos

Como necesitamos saber cuando una persona va perder el examen de las pruebas Saber 11, nuestro principal recurso para el proyecto son los datos para lo cual se obtuvieron de datos abiertos gobiernos en linea ([www.datos.gov.co](http://www.datos.gov.co)), se descargaron los resultados de los exámenes presentados en el 2018-1 con un total de 19,8K y 87 columnas [3].  
  
El diccionario de datos[2] de las 87 columnas se encontró documentado en la página del ICFES (Instituto Colombiano para la Evaluación de la Educación) se encuentra en la referencia mencionada, se realiza una especificación si dicho campo aplica para el año presentado, la descripción del campo, convención y posibles valores.



**Extracción de características:**

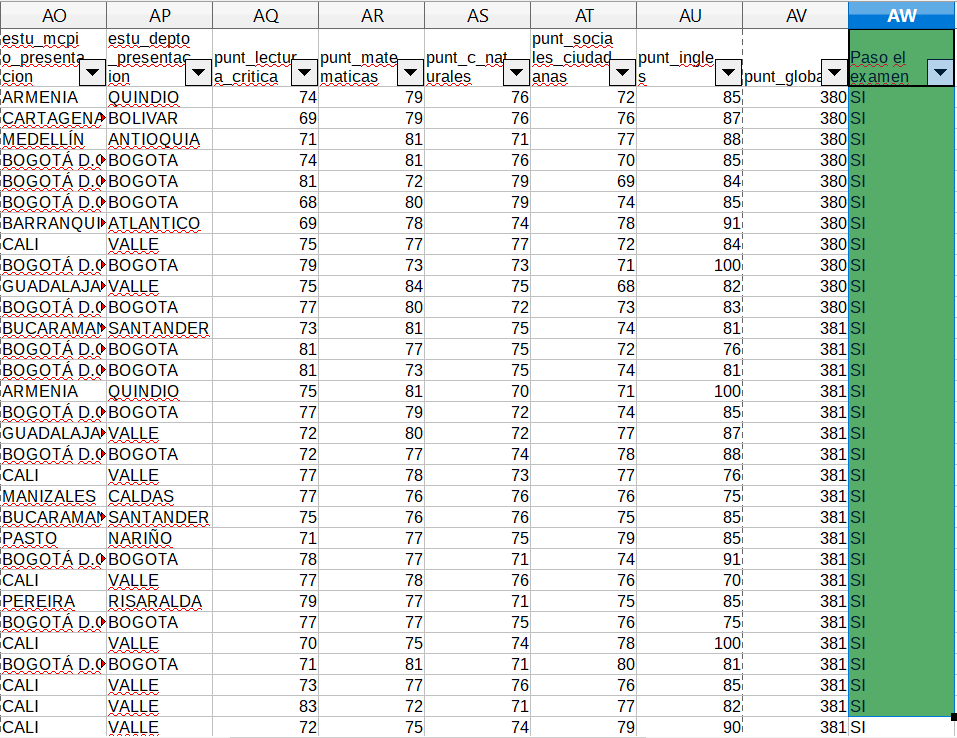
De los 87 columnas se seleccionaron 48 columnas que contienen la información relevante para la predicción los cuales están relacionados con el ambiente familiar, objetos que tienen en el hogar y puntaje en cada una secciones del examen.



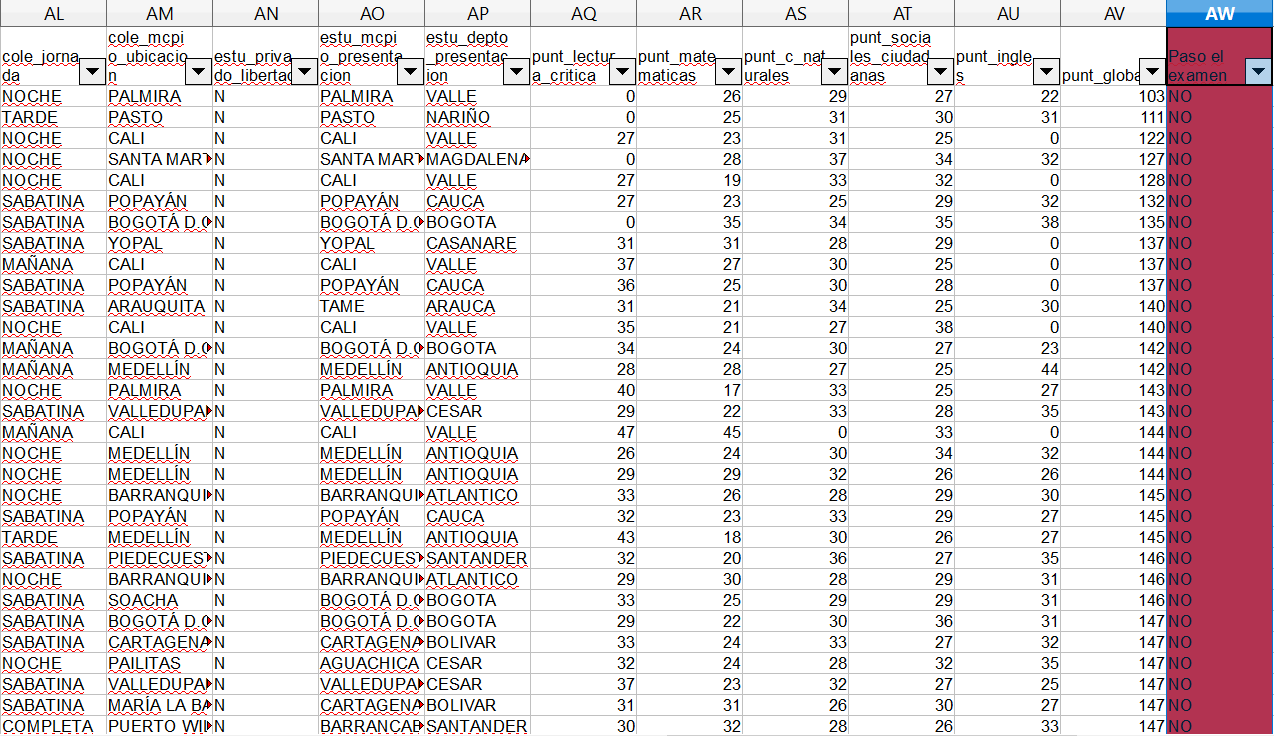
**Categorización:**

De los 19,8K registros de la data proporcionada por datos en linea seleccione 2000 registros, los cuales fueron categorizados en “SI paso pruebas saber 11” y “NO paso pruebas saber 11” distribuidos equitativamente para no generar resultados inconsistentes.

Categoría *“SI paso pruebas saber 11”*

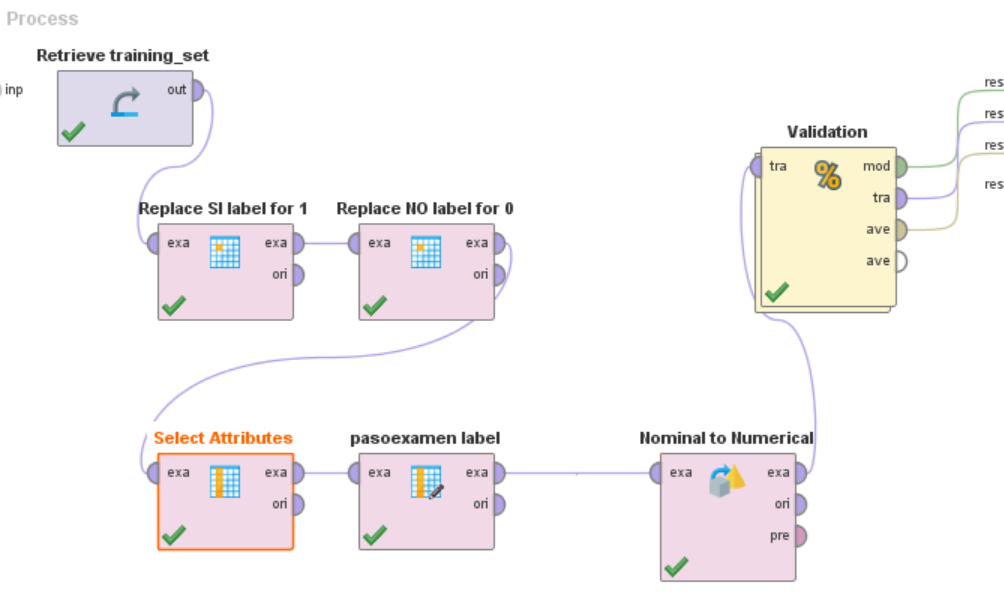


Categoría *“NO paso pruebas saber 11”*



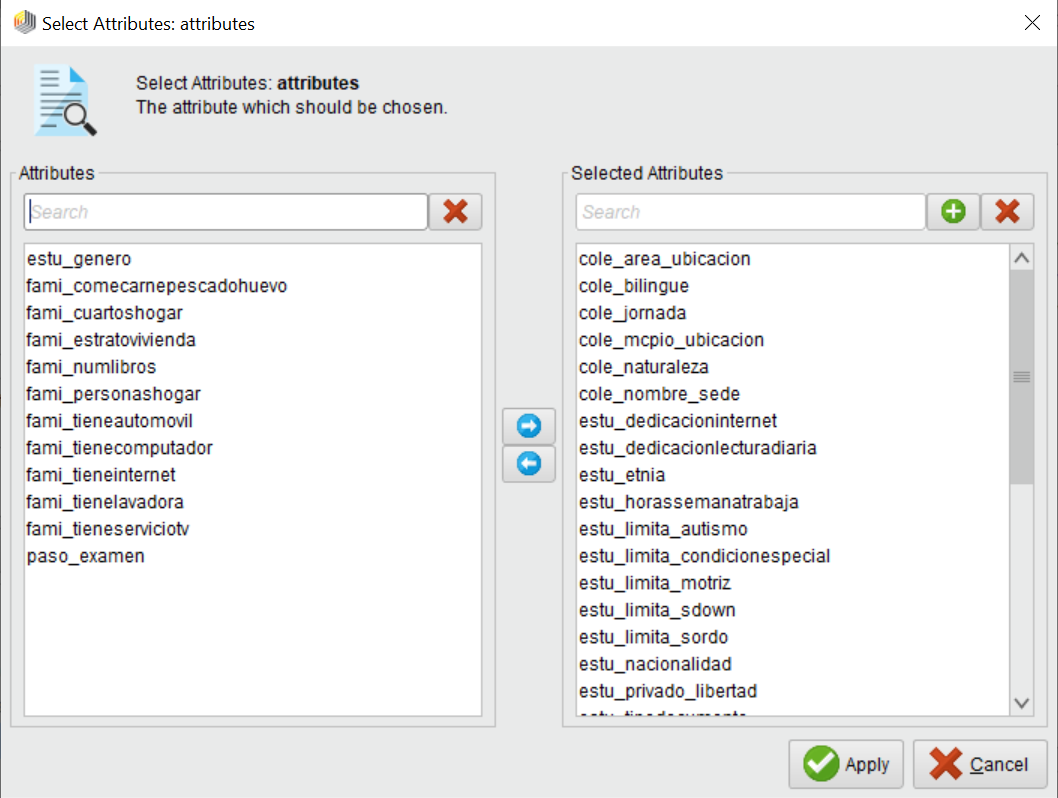
# PROCESAMIENTO

Se uso la herramienta rapidminer para el procesamiento y clasificación de los datos.



Los pasos realizados para el preprocesamiento de los datos fueron:

* Reemplazar los label por valores numéricos:
  + SI es igual a 1
  + No es igual a 0
* Seleccionar las características con las cuales se va a realizar la predicción, que están realicionadas con objetos que tiene la familia, estrato social, genero, come carne pescado huevo.

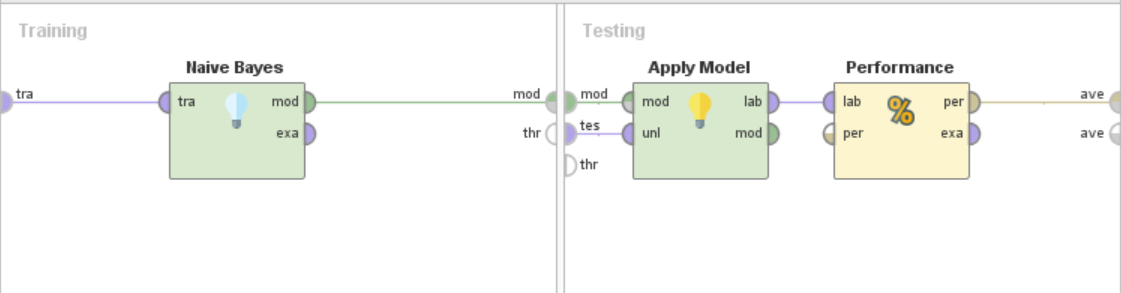


* Convertir los datos de texto a numérico.

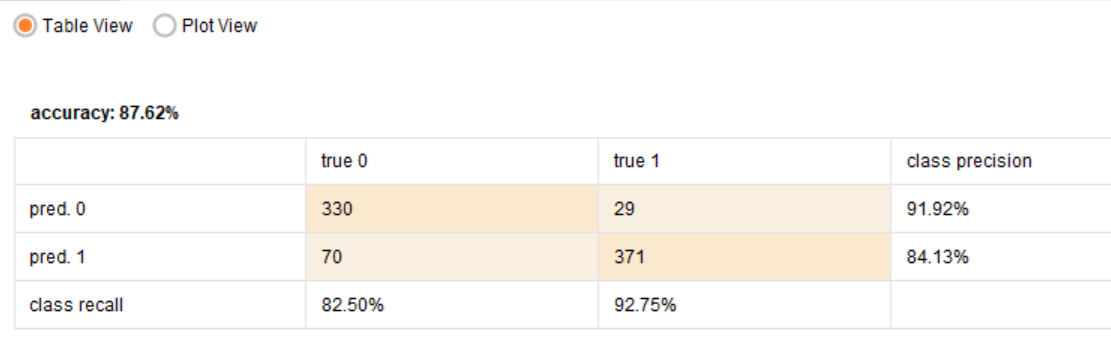
# Predicción

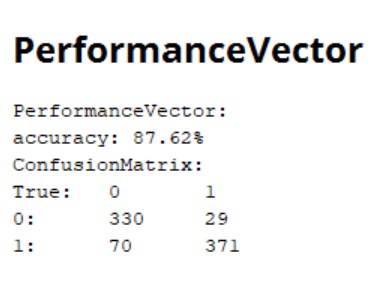
Como necesitamos saber cuando una persona va perder el examen de las pruebas Saber 11, nuestro principal recurso para el proyecto son los datos para lo cual se obtuvieron de datos abiertos gobiernos en linea [(www.datos.gov.co](http://www.datos.gov.co)), se descargaron los resultados de los exámenes presentados en el 2018-1 con un total de 19,8K y 87 columnas[3]. Para la predicción se uso las técnicas Naive Bayes, regresión lineal y arboles de decisión.

Cuando se realizo la evaluación de Naive Bayes con el training set:

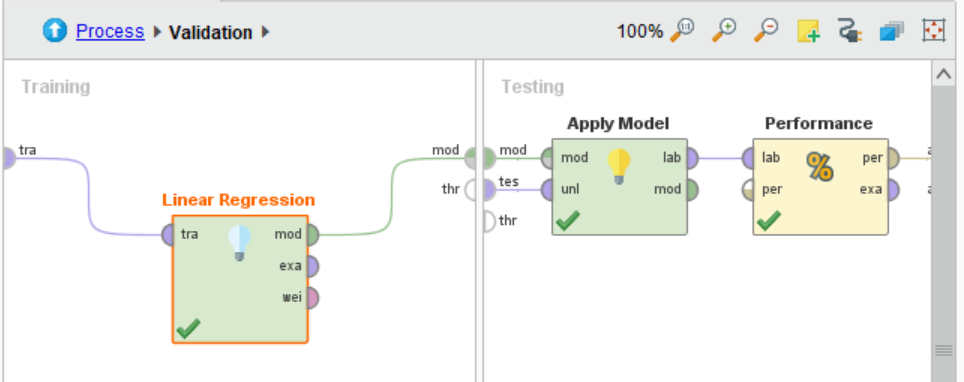


El accuraccy obtenido es de **87.62%**

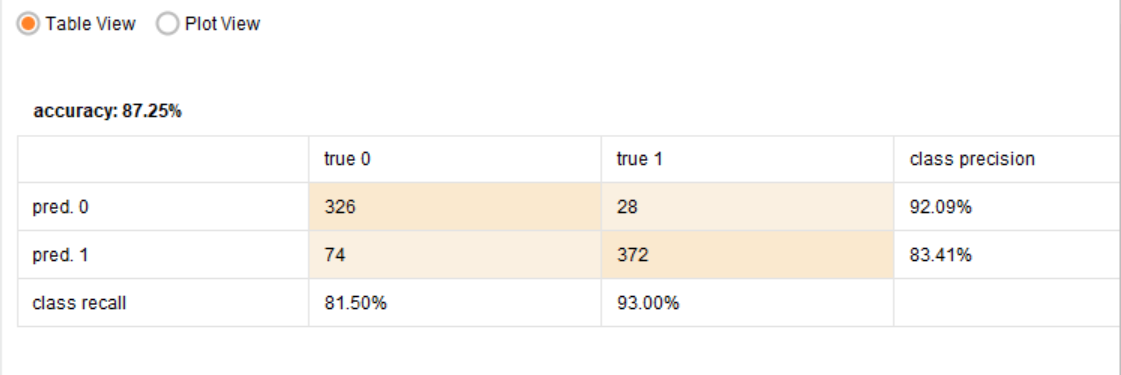


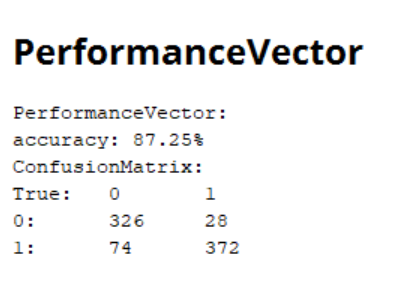


Prueba con el método de regresión lineal

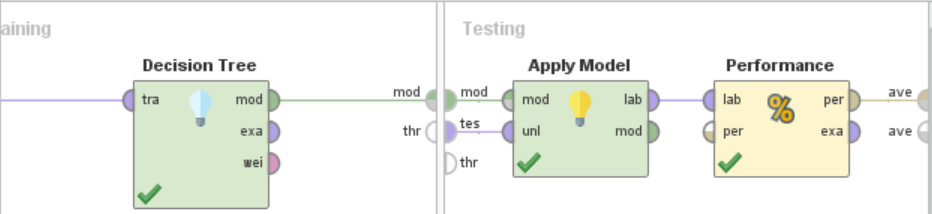


El accuraccy obtenido es de **87.25%**

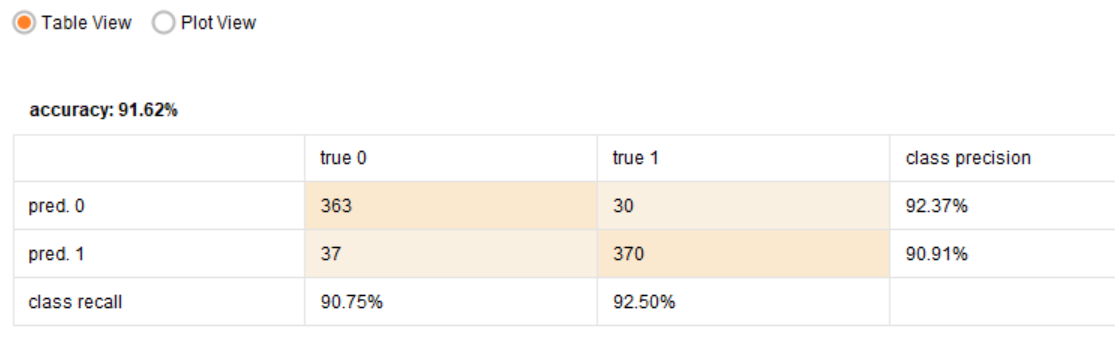




Prueba con arboles de decisión

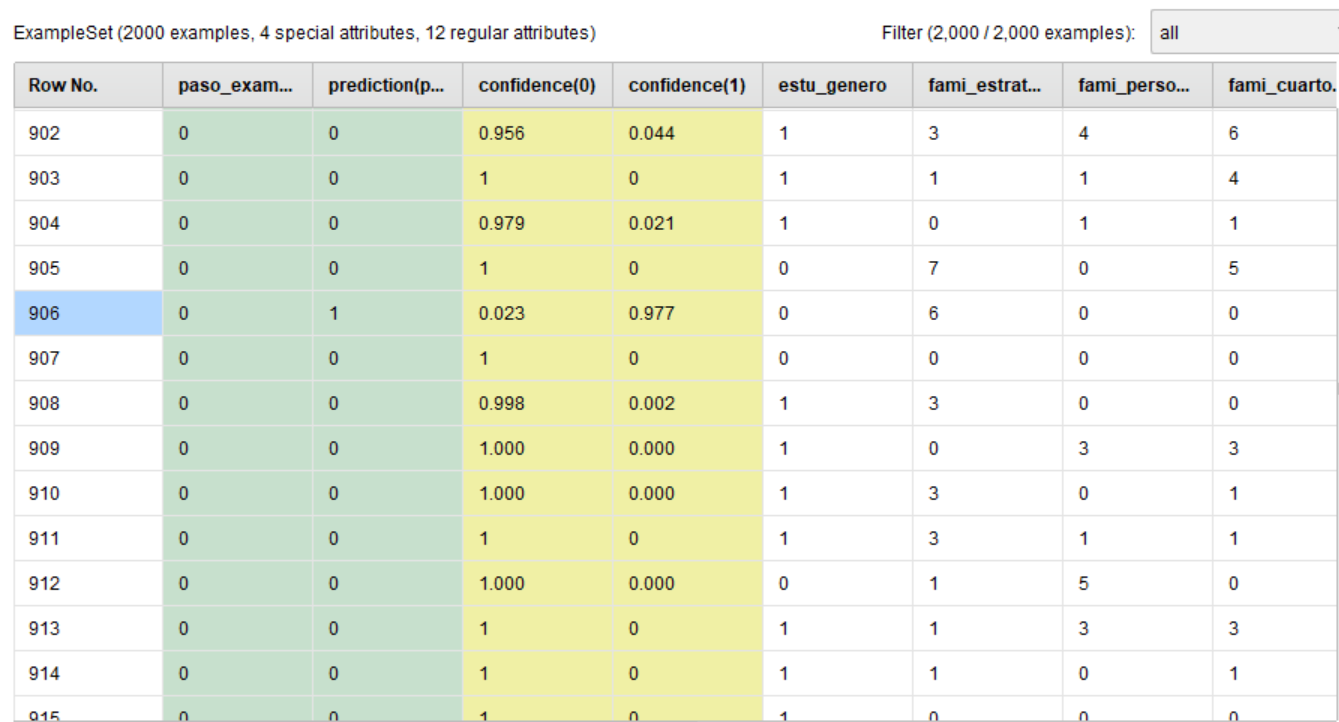


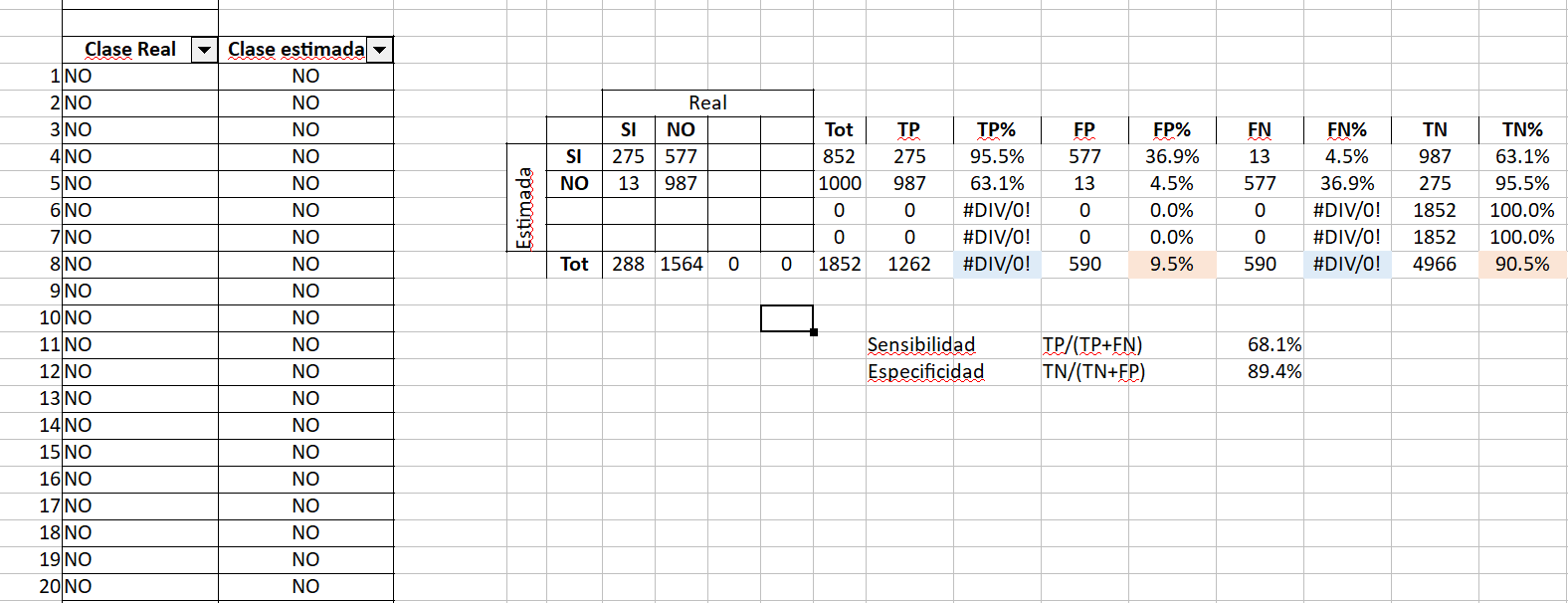
El accuraccy obtenido es de **91.62%**



Matriz de confusión

Se realizó la validación de los valores reales y esperados para obtener matriz de confusión de Naive Bayes





# Conclusiones

Cuando se inicio el proceso de predicción se elegieron unas características de acuerdo al criterio del experto despues con las pruebas con los metodos de predicción se decidió modificar las características quitando características como los puntajes y limitaciones físicas.  
El método de predicción con el mejor accuracy fue el de arboles de decisión con 91.62%  
Se realizo la validación de la matriz de confusión para el método de Naive Bayes para obtener la especificidad y senbilidad.

1. References
2. Secciones del examen pruebas saber 11

<https://es.wikipedia.org/wiki/ICFES#Pruebas_Saber>

1. Formulario de inscripción estudiantes saber 11

<http://www2.icfes.gov.co/docman/investigadores-y-estudiantes-de-posgrado/acceso-a-bases-de-datos/documentacion-y-diccionarios-saber-11/formularios-de-inscripcion-saber-11/5546-ac-2014-2-est/file?force-download=1>

1. Resultados del examen Saber 11 para el primer periodo del año 2018

<https://www.datos.gov.co/Educaci-n/Saber-11-2018-1/w42r-j8f9>

1. Diccionario de datos pruebas saber 11

<http://www2.icfes.gov.co/docman/investigadores-y-estudiantes-de-posgrado/acceso-a-bases-de-datos/documentacion-y-diccionarios-saber-11/diccionarios-saber-11/5527-diccionario-saber-11-2014-2-a-2018-1/file?force-download=1>

1. <https://dl-acm-org.loginbiblio.poligran.edu.co/citation.cfm?id=2460324>
2. <http://www.ncolr.org/jiol/issues/pdf/12.1.2.pdf>
3. <https://ac.els-cdn.com/S0360131516301634/1-s2.0-S0360131516301634-main.pdf?_tid=e96bb767-bd74-4387-b4a1-96f146a022c4&acdnat=1527224021_aceba1cd9f6cf88342c41bfc614fc4d0>

1. . [↑](#footnote-ref-0)