Análisis Dropout utilizando técnicas de aprendizaje supervisado y exploración de diferentes vectores de características

Autor: Leonardo Bravo Rain
Curso: Minería y Aprendizaje Automático de Datos
Fecha: 1° Semestre 2019.

Aclaración: Se denomina dataset normal al vector de características obtenido desde la data “bruta” y no al dataset obtenido desde grafo.

Resumen análisis realizado

En el presente análisis, se compararon diferentes vectores de características los cuales se utilizaron para clasificar si un estudiante realizaría Dropout (clasificación binaria). Para esto se utilizó un vector de características obtenido directamente desde la base de datos utilizadas, un vector de características obtenido desde un grafo que representa la información de la base de datos y un vector de características obtenido a partir de mezclar los vectores anteriores.

Para la clasificación se entrenaron 5 modelos diferentes, los cuales corresponden a Support Vector Machine, Decision Tree, Rain Forest, KNN y Logistic Regression.

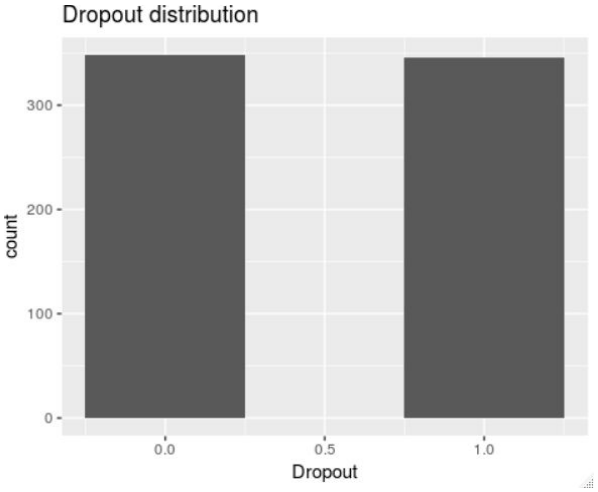
A partir de los valores de las diferentes métricas utilizadas, se tiene que el vector de características obtenido a partir de la mezcla de vectores de características normal y grafo, es quien obtiene mejores resultados en la clasificación de Dropout, teniéndose valores de accuracy de 93,3 [%], kappa de 85,2 [%] y F1 score de 94,8 [%].

Análisis de vector de características utilizando data normal

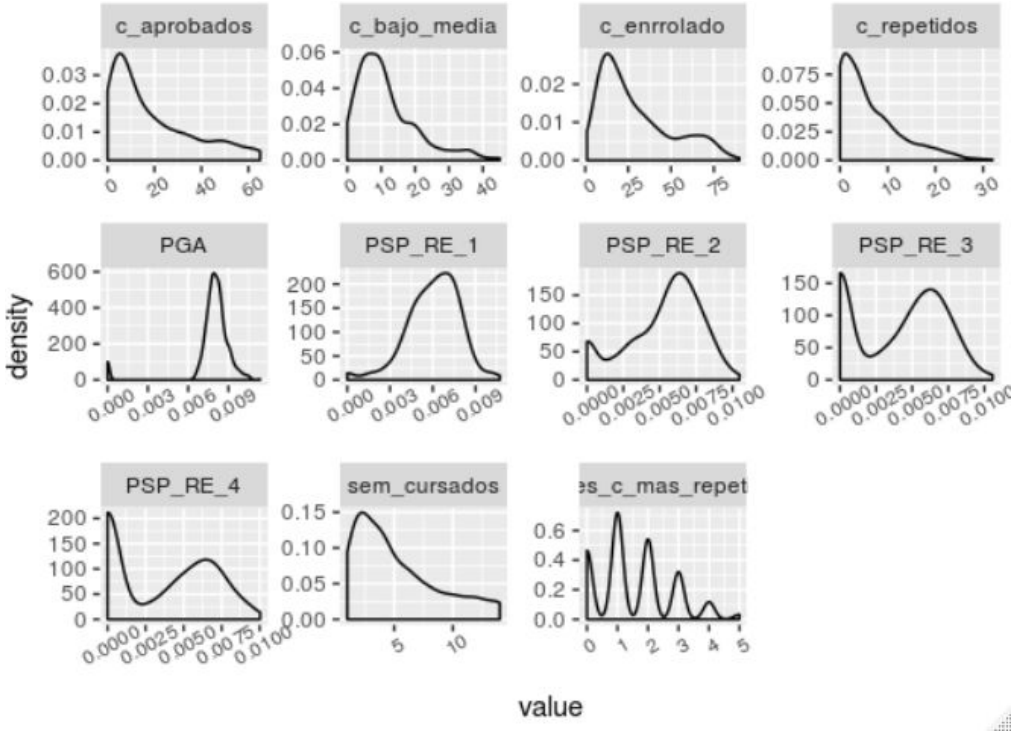
Se analiza dataset “dataset_procesado.csv”, en donde se tienen 24 variables independientes y 1 variable dependiente (Dropout), con 694 instancias.

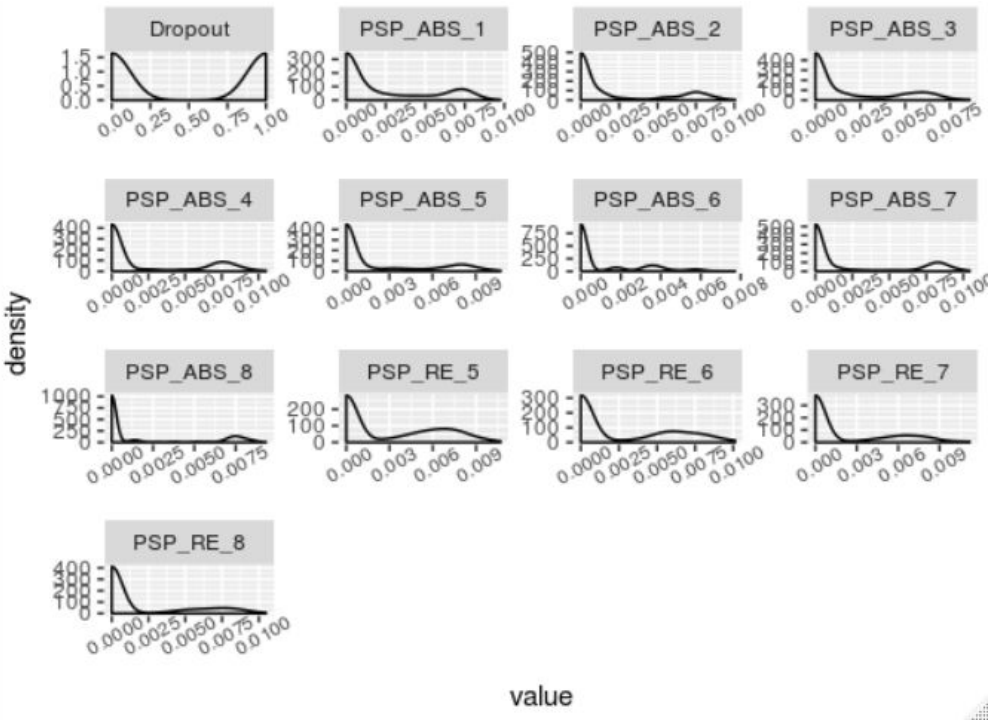
Análisis exploratorio de datos

Analizando distribución de Dropout, se tiene una distribución de clases balanceada:

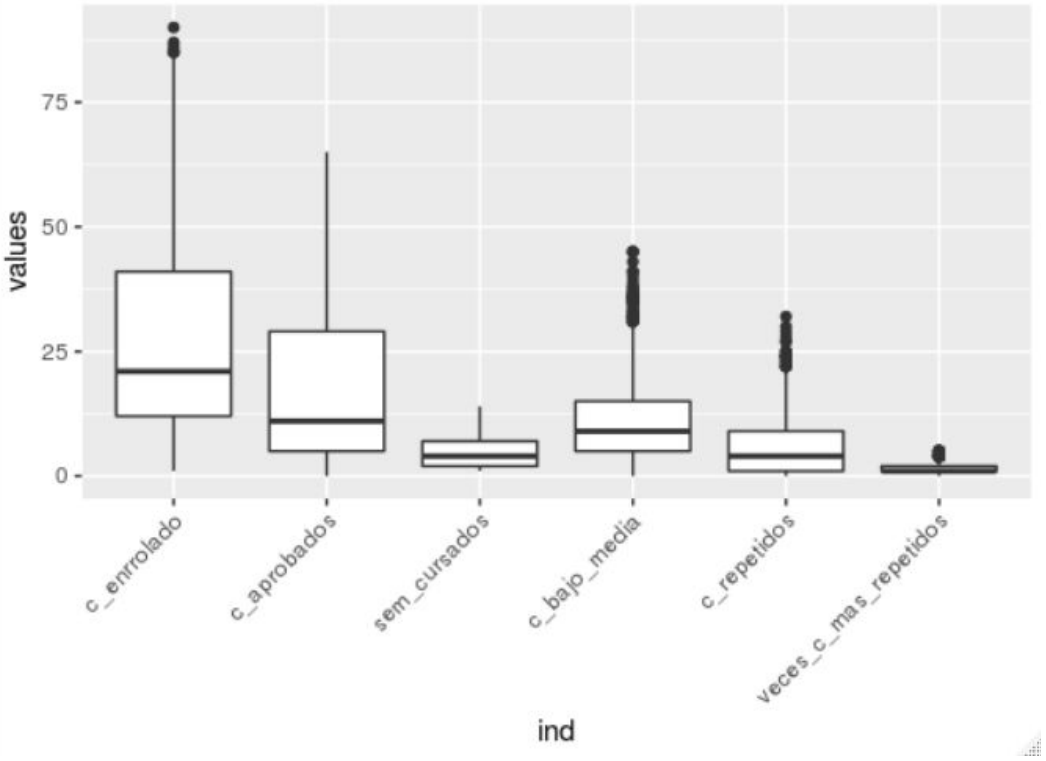


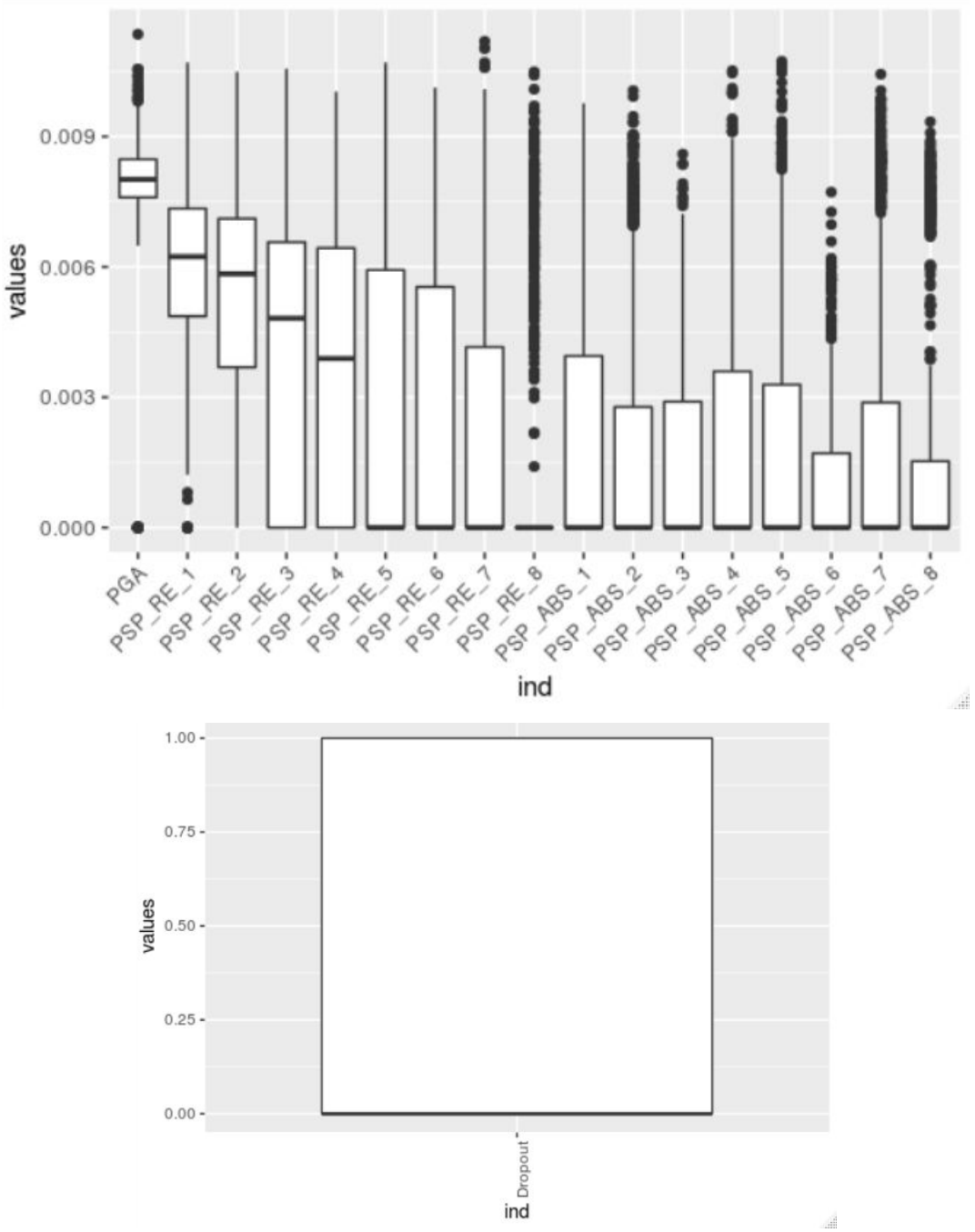
La distribuciones para las variables independientes se presentan a continuación, en donde se observa que existen distribuciones similares a las normales para algunas variables, y distribuciones alejadas de la normalidad para otras. Pese a esto, se utilizarán estas variables para el desarrollo de los modelos.





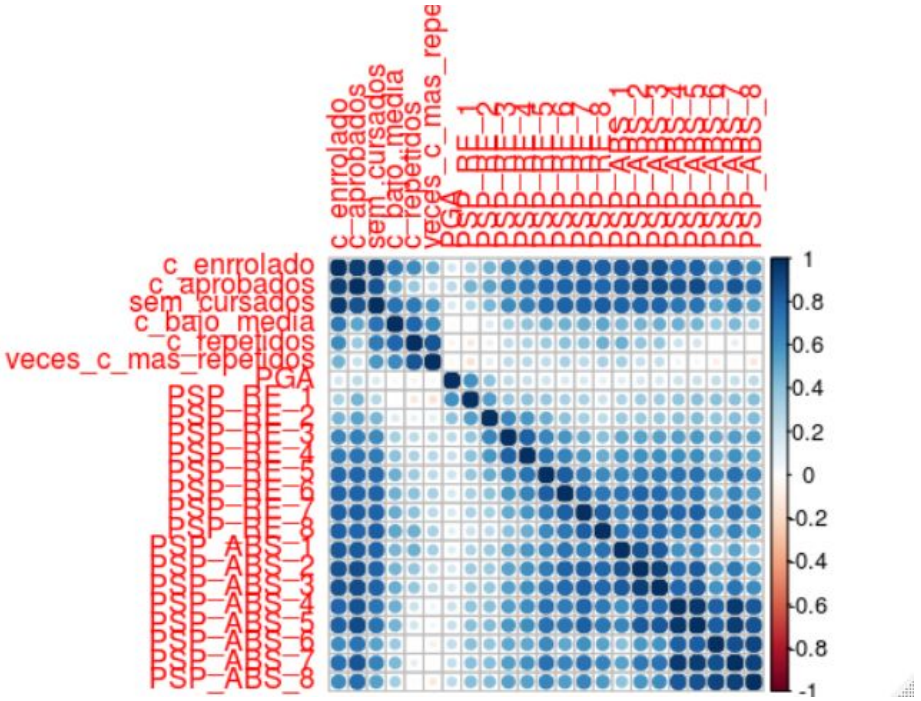
Los promedios y las distribuciones de los datos son presentados a continuación, en donde se observa la presencia de varias variables con datos outliers. Se propone realizar análisis de outliers para una posible extensión de investigación, sin embargo en este análisis se consideran los outliers para el proceso de desarrollo de los modelos.



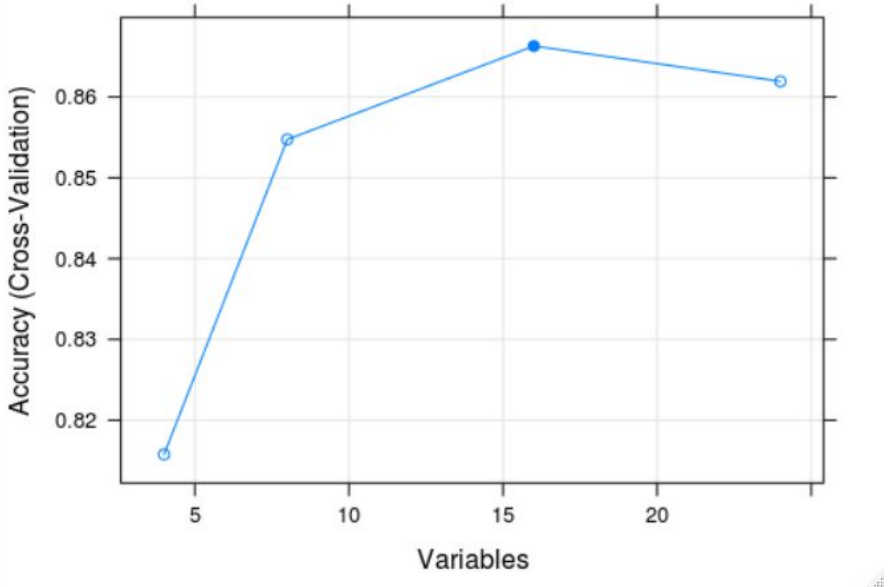


Feature selection

La matriz de correlación entre las variables independientes se presenta en la siguiente figura, en donde se observan las diferentes relaciones entre las variables:



Se utiliza algoritmo RFE para seleccionar aquellas variables más relevantes en el dataset, obteniéndose que el máximo valor de accuracy se obtiene utilizando 16 variables, tal como se presenta en la siguiente figura:



En donde las variables mas relevantes corresponden a "c_mas_repetido", "c_aprobados", "c_enrrolado", ‘sem_cursados", "PSP_ABS_8", "PSP_ABS_1", "c_repetidos", "PSP_RE_1", "PGA", "PSP_RE_4", "PSP_ABS_6", "PSP_RE_3", “c_bajo_media", "PSP_RE_2", “PSP_ABS_7" y "veces_c_mas_repetidos".

Desarrollo de modelos de clasificación

Se separa el dataset, en donde el 70 [%] de las instancias son para training y 30 [%] son para testing.

Se desarrollan 5 modelos:

1) Support Vector Machine
2) Decision Tree
3) Random Forest
4) K-Nearest Neighbors
5) Logistic Regression

Cada uno de los modelos fueron entrenados utilizado las siguientes condiciones:

- Cross-Validation con k = 10, y se repite 3 veces.
- Preprocesamiento de centrado y escalado.

Los resultados son presentados en la siguiente tabla, en donde se observa que el modelo rain forest obtiene el máximo valor de accuracy, kappa y f1 score (los valores mostrados son obtenidos comparando las predicciones sobre datos de testing y las clases de los datos reales) :

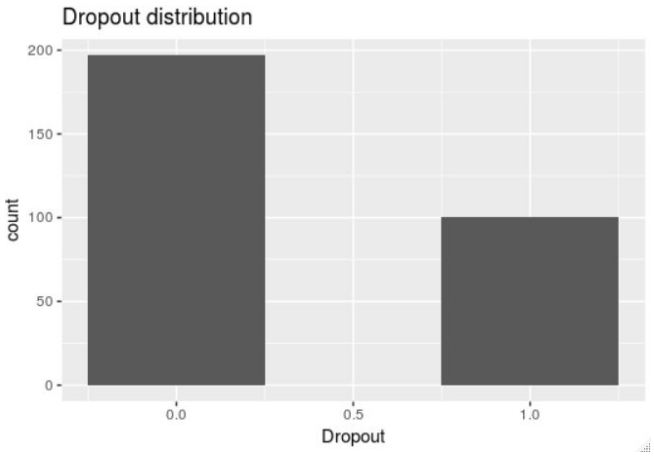
	model	accuracy	kappa	recall	precision	f1_score
1	svm linear	0,855	0,710	0,846	0,863	0,854
2	decision tree	0,821	0,643	0,731	0,894	0,804
3	knn	0,802	0,604	0,788	0,812	0,800
4	rain forest	0,879	0,759	0,837	0,916	0,874
5	logistic regression	0,821	0,643	0,779	0,853	0,814

Análisis de vector de características utilizando data desde grafo

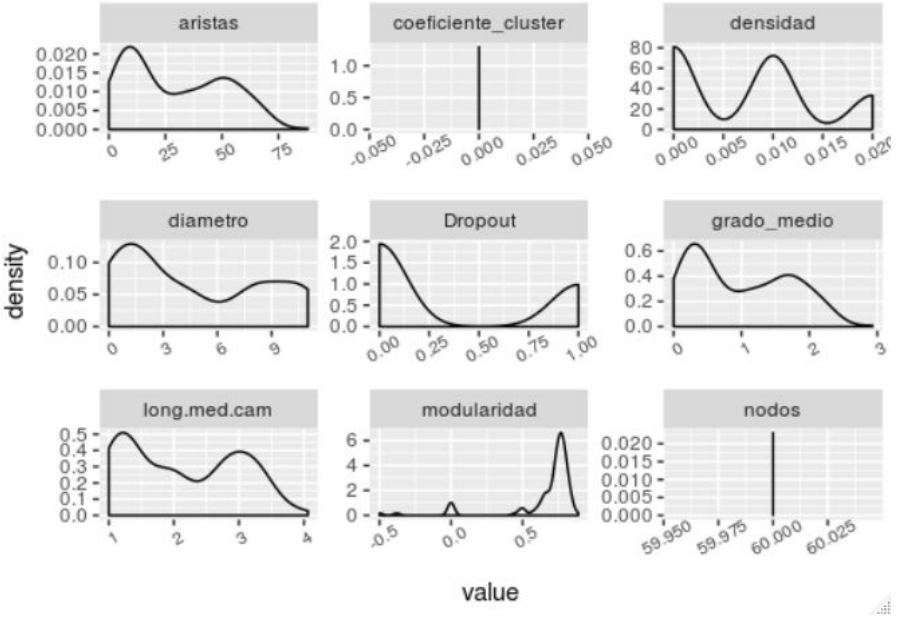
Se analiza dataset “graph_data.csv”, en donde se tiene 297, con 8 variables independientes y 1 variable dependiente.

Análisis exploratorio de datos

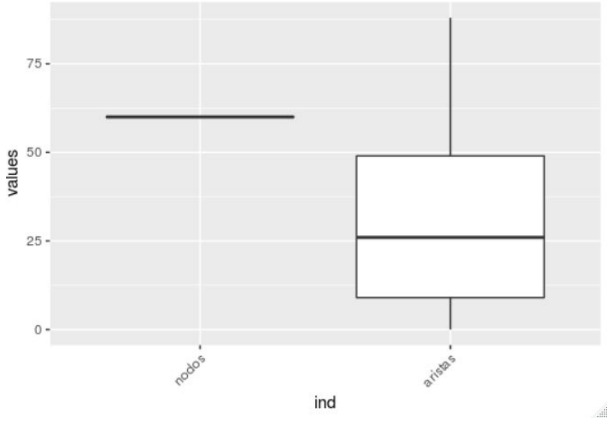
Se analiza la distribución de las clases, en donde se tiene datos desbalanceados, ya que hay más instancias de la clase 0. Esto podría influir en los resultados obtenidos, por lo que se propone realizar análisis en un trabajo posterior. En este análisis se realiza no se considera este factor.

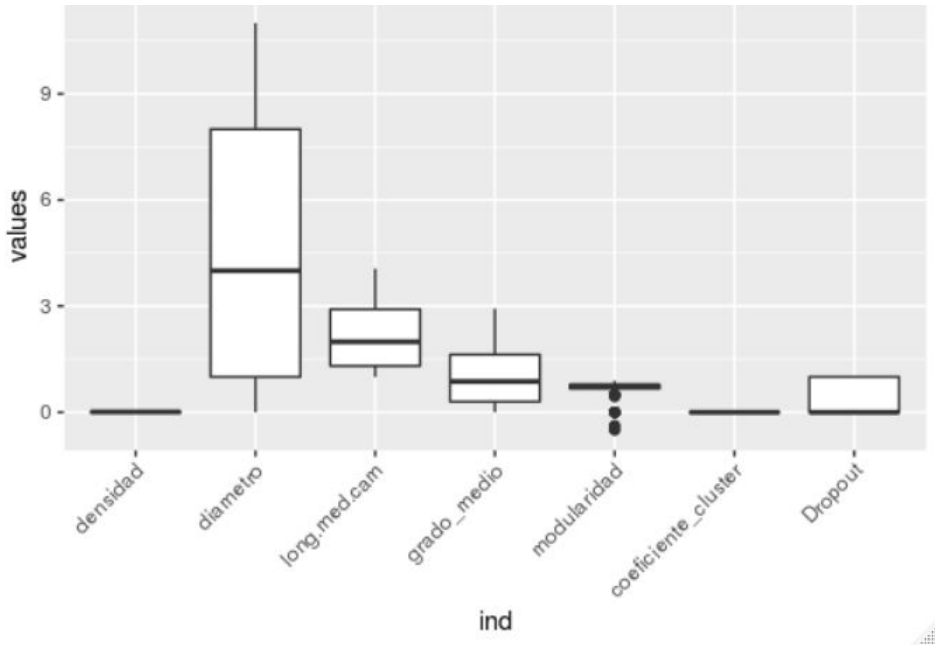


Las distribuciones de las variables independientes se presentan en la siguiente figura, en donde se destaca que las variables “coeficiente_cluster” y “nodos” corresponden a un único valor para todas las instancias. Dado lo anterior, estas variables se eliminan del análisis.



Los promedios y los datos outliers se presentan en la siguiente figura, en donde se observa que existen outliers solo para la variable modularidad. También se vuelve destacar el valor constante de las variables “coeficiente_cluster” y “nodos”:

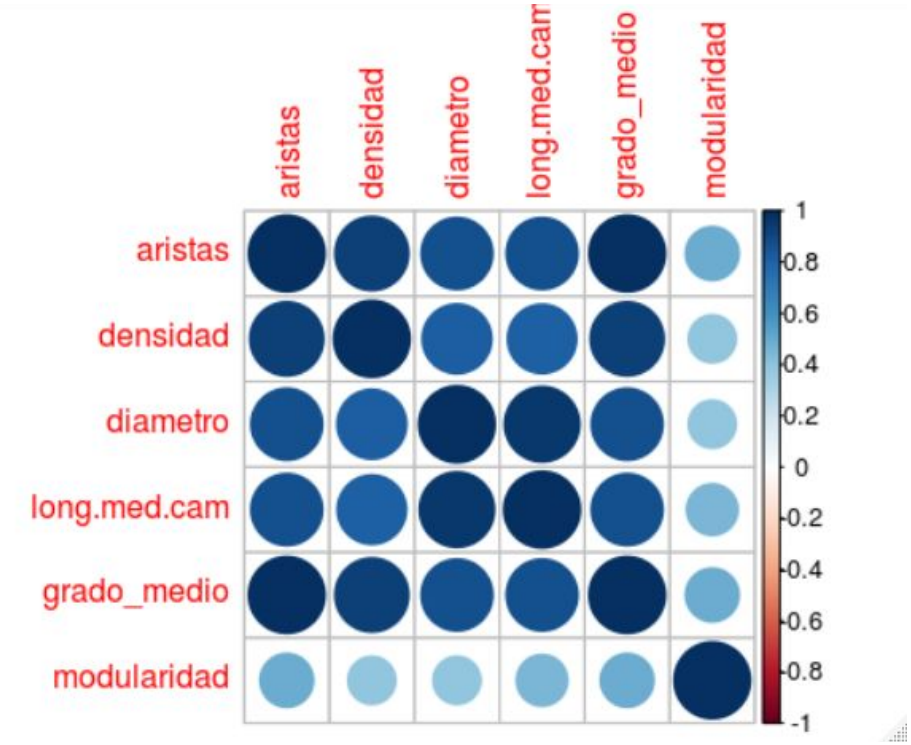




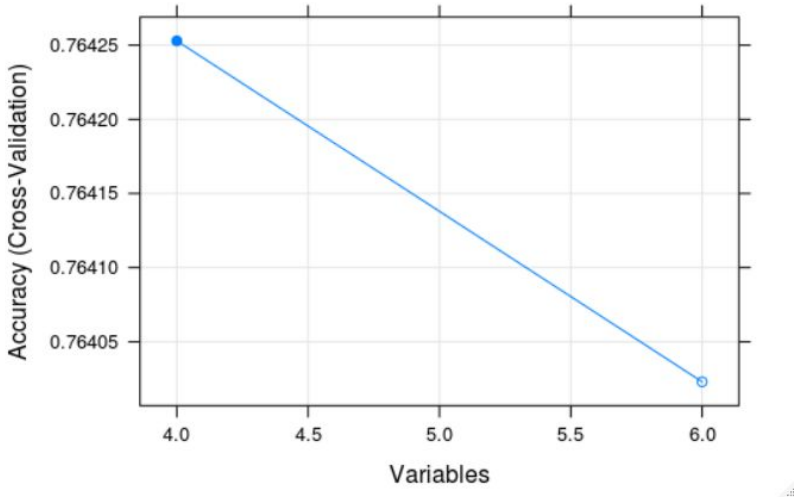
Para el caso de la variable “long.med.cam” se cambian los 47 valores con valor NaN por 0.

Feature selection

Las correlaciones entre las variables independientes se presentan en la siguiente figura, en donde se observa gran correlación entre diferentes variables:



Se utiliza algoritmo RFE para seleccionar variables más relevantes, mostrándose los resultados en la siguiente figura, en donde se observa que la mayor accuracy se obtiene para 4 variables, las cuales corresponden a "diametro", "modularidad", "long.med.cam", “aristas":



Desarrollo de modelos de clasificación

Se divide dataset en 70 [%] para training y 30 [%] para testing. Se entrenan mismos modelos que para el dataset anterior, utilizando las mismas condiciones descritas anteriormente. Los resultados son presentados en la siguiente tabla, en donde se observa que el modelo knn es el que obtiene una mayor accuracy, kappa y f1 score:

	model	accuracy	kappa	recall	precision	f1_score
1	svm linear	0,742	0,328	0,949	0,737	0,830
2	decision tree	0,753	0,398	0,898	0,768	0,828
3	knn	0,787	0,475	0,932	0,786	0,853
4	rain forest	0,708	0,335	0,797	0,770	0,783
5	logistic regression	0,730	0,318	0,915	0,740	0,818

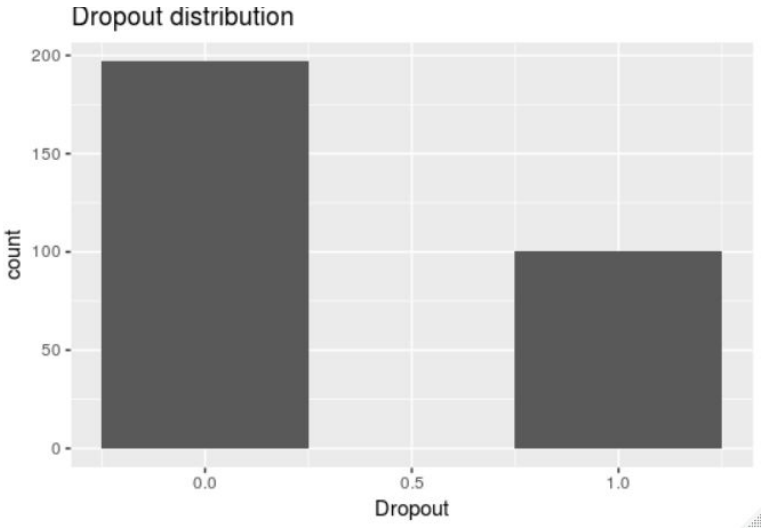
Dado que se entrena modelo utilizando dataset desbalanceado, los valores de accuracy no son confiables. Por lo que un buen parámetro se considera el valor de f1 score.

Análisis de vector de características utilizando data desde data normal y desde grafo

Se analiza dataset obtenido al unir ambos datasets anteriores, el cual se denomia “merge_feature_vector.csv”, en donde se tienen 30 variables independientes y 1 dependiente, utilizando 297 instancias.

Análisis exploratorio de datos

Se analiza la distribución de las clases, en donde se tiene datos desbalanceados, ya que hay más instancias de la clase 0:

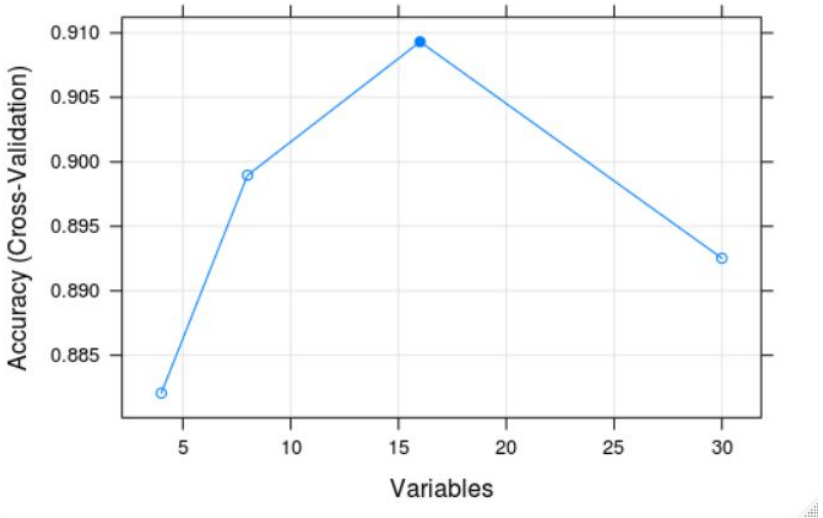


Las distribuciones de las variables independientes y sus promedios y outliers son los mismos a los datasets independientes, por lo que no se presentan.

Feature selection

Las correlaciones entre las variables independientes no se presentan debido a extensión de figura.

Se utiliza algoritmo RFE para seleccionar variables más relevantes, mostrándose los resultados en la siguiente figura, en donde se tiene que con 16 variables se obtiene la máxima accuracy:



Las 16 variables más relevantes corresponden a variables obtenidas desde el dataset de datos normal (14 variables) y el dataset de grafos (2 variables), por lo que corresponden a mezclas de variables. Estas variables son presentadas en la siguiente tabla:

N°	Variable	Fuente
1	c_aprobados	Normal
2	c_enrrollado	Normal
3	sem_cursados	Normal
4	c_mas_repetido	Normal
5	PSP_ABS_1	Normal
6	PSP_ABS_8	Normal
7	PSP_ABS_6	Normal
8	c_repetidos	Normal
9	PSP_ABS_5	Normal
10	PSP_ABS_7	Normal
11	PSP_RE_1	Normal
12	lon.med.cam	Grafo
13	diametro	Grafo
14	PSP_ABS_4	Normal
15	grado_medio	Normal
16	PGA	Normal

Desarrollo de modelos de clasificación

Se divide dataset en 70 [%] para training y 30 [%] para testing. Se entrenan mismos modelos que para el dataset anterior, utilizando las mismas condiciones descritas anteriormente.

Los resultados son presentados en la siguiente tabla, en donde se tiene que el modelo rain forest posee los máximo valores de accuracy, kappa y f1 score:

	model	accuracy	kappa	recall	precision	f1_score
1	svm linear	0,888	0,753	0,898	0,930	0,914
2	decision tree	0,910	0,805	0,898	0,964	0,930

3	knn	0,854	0,670	0,898	0,883	0,891
4	rain forest	0,933	0,852	0,932	0,965	0,948
5	logistic regression	0,876	0,730	0,881	0,929	0,904

Dado que se entrena modelo utilizando dataset desbalanceado, los valores de accuracy no son confiables. Por lo que un buen parámetro se considera el valor de f1 score.

Análisis de resultados

En la siguiente tabla se presentan los resultados obtenidos para cada vector de características, presentandose solo aquellos que obtuvieron los mejores resultados para cada caso:

Vector car.	model	accuracy	kappa	recall	precision	f1_score
Normal	rain forest	0,879	0,759	0,837	0,916	0,874
Grafo	knn	0,787	0,475	0,932	0,786	0,853
Mezcla ambos	rain forest	0,933	0,852	0,932	0,965	0,948

Desde la tabla se observa que el caso del vector de características obtenido desde la mezcla de ambos es quien obtiene mejores métricas, siendo las mas relevantes los valores de accuracy, kappa y f1 score, en donde para cada uno, estos valores son mayores comparados con los otros vectores de características.

Conclusión

A partir de los valores de las diferentes métricas utilizadas, se tiene que el vector de características obtenido a partir de la mezcla de normal y grafo, es quien obtiene mejores resultados en la clasificación de Dropout, teniéndose valores de accuracy de 93,3 [%], kappa de 85,2 [%] y F1 score de 94,8 [%].