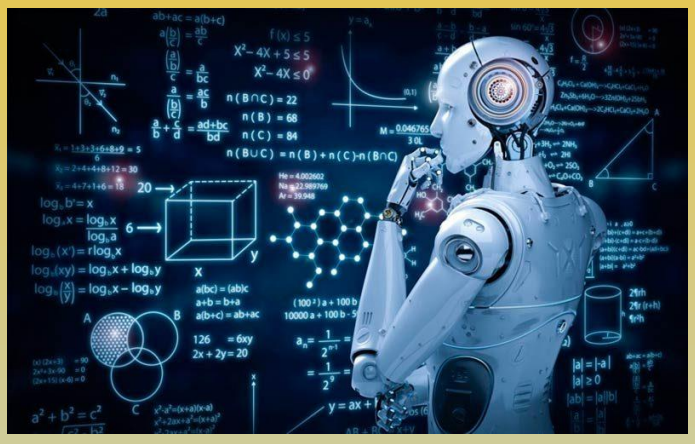
Facultad de Ingeniería



Desempeño Académico 2019

Analítica de Datos Julio Godoy Del Campo

PROYECTO ANALÍTICA DE DATO

El DataSet tiene por objetivo es identificar las variables predictoras del éxito y/o riesgo académico que mejor explica un algoritmo predictivo de la deserción académica dada una preclasificación de estudiantado preestablecida. Por consiguiente, el propósito es obtener una clasificación que identifique aquella combinación de variables que explica en mayor medida a la preclasificación del estudiantado.

Edgardo Cabrera - Loreto Mora
Analítica de Datos

iit UdeC

DESCRIPCIÓN DATASET

Introducción

La información del DataSet disponible para el proyecto "Desempeño Académico" es de carácter histórico (2000-2017) sobre la condición de ingreso del estudiante y su respectivo desempeño académico durante el proceso formativo de las carreras dictadas por la Facultad de Ingeniería de la Universidad de Concepción (UdeC). La data se encuentra contenida en 5 archivos, que se describen a continuación:

Admisión; contiene información referente a la condiciones de ingreso del estudiante a la carrera, tales como; puntajes PSU (ponderado y desglosado por matemática, lenguaje y ciencia), ponderación NEM, el tipo de colegio de egreso (particular, subvencionado, municipalizado), dato demográfico (sexo),entre otros.

Alumnos inscritos por asignaturas (Alu-Insc-Asi); dada la dimensión de la data está dividida en tres pac y contiene información referente a las asignaturas cursadas por los alumnos por semestre con sus respectiva identificación de la asignatura así también las evaluaciones.

Asignaturas complementarias; información a la oferta de asignaturas complementaria para la facultad de Ingeniería, tal como; nombre de la asignatura, unidad académica a la que pertenece, duración, créditos asociados, horas semanales (teóricas, laboratorio y practicas), entre otros.

Rendimiento; contiene información cruzada entre la data contenida en el DataSet Admisión con el DataSet Situación.

Situación; contiene información referente a la situación académica actual del estudiante, es decir, si es alumno regular, baja académica, baja por no inscripción, suspensión de estudio autorizada, autorizado para titularse, titulado, etc).

En relación a los indicadores presentes en cada DataSet se encuentran:

Tabla 1: Variables contenidas en DataSet-Admisión y DataSet - Alu-Insc-Asi

DataSet - Admisión	DataSet - Alu-Insc-Asi
Matrícula	Código per
Ponderación psu Matemática	Código Alumno
Ponderación psu Lenguaje	Código Carrera
Ponderación psu Ciencias	Código pac
Ponderación NEM	Código Asignatura
PSU ponderado	Asignatura Sección
Preferencia de postulación Carrera	Nota
Código Carrera	Nota Tipo
Nombre Carrera	Frecuencia
Colegio de Egreso	Nombre asignatura
Cohorte	Nombre unidad académica
Estado	Unidad académica Id
Semestre Estado	
Sexo	

Fuente: Elaboración propia

Tabla 2: Variables contenidas en DataSet- Asignaturas Complementarias y DataSet - Situación

DataSet - Asignaturas Complementarias	DataSet- Situación
Código Asignatura	Rut
Nombre Asignatura	Matricula
Depto[1]	Cohorte
Nombre_1[2]	Carrera
Duración	Situación Nombre
Créditos	Periodo
Horas Teóricas	Fecha
Horas Prácticas	
Horas Laboratorio	
Días Terreno	
Horas Trabajo Académico	
Modalidad	
Plev[3]	
Créditos Transferibles	
Fecha Creación	
Periodo	
Agno	

Fuente: Elaboración propia

Tabla 3: Variables contenidas en DataSet- Rendimiento

DataSet - Rendimiento	
Rut	Puntaje Ranking
Matrícula	Puntaje Ponderado
Cohorte	Puntaje NEM
Género	Situación Actual
Quintil	Puntaje Matemáticas
Carrera	Puntaje Lenguaje
Grupo Dependencia	Nota 7
Promedio Notas	Promedio Curricular
Vía Ingreso	Promedio Ponderado
Preferencia	

Fuente: Elaboración propia

[1] Departamento al que pertenece la asignatura
[2] Nombre de la asignatura dictada
[3] Asignatura modalidad escuela de verano

Objetivo Estudio

El objetivo del presente DataSet es identificar los principales predictores de éxito y riesgo académico que mejor se ajustan a la efectiva predicción de deserción académica. Por consiguiente, el propósito del estudio se centrará en determinar una clasificación, de forma tal que se logre identificar cuál combinación de variables explica, en mayor medida, la clasificación asociada

Descripción de datos

A continuación, se procede a la descripción de la DataSet, la que se muestra en figuras 1 a 5. Adicionalmente se presenta el análisis de los tipos de variables contenidas en las distintas DataSet, la que se muestra en tabla 1 a 5.

Data_Admisión

Ilustración 1: Análisis descriptivo Data_Admisión

	Matricula	Matematica	Lenguage	Ciencias	NEM	Preferencia	Cod. Carrera	Cohorte	Estado	Sem. Estado
count	1.208900e+04	12089.000000	12089.000000	12089.000000	12088.000000	12089.000000	12089.000000	12089.000000	12089.00000	1.206900e+04
mean	3.934845e+09	661.686161	608.703119	610.508479	657.070069	0.542642	3309.789064	2010.939780	5.17396	2.013510e+06
std	4.025436e+06	87.114888	89.989444	92.474841	99.151633	0.972328	24.513405	4.003371	9.84183	3.814711e+03
min	3.925421e+09	0.000000	0.000000	0.000000	0.000000	0.000000	3180.000000	2004.000000	1.00000	2.004100e+06
25%	3.930928e+09	628.000000	569.000000	578.000000	620.000000	0.000000	3309.000000	2007.000000	1.00000	2.011100e+06
50%	3.934971e+09	663.000000	613.000000	619.000000	662.000000	0.000000	3313.000000	2011.000000	2.00000	2.014200e+06
75%	3.937976e+09	702.000000	657.000000	657.000000	723.000000	1.000000	3318.000000	2014.000000	5.00000	2.017100e+06
max	3.941294e+09	850.000000	837.000000	850.000000	826.000000	9.000000	3327.000000	2017.000000	99.00000	2.017200e+06

Fuente: Elaboración propia con lenguaje Python

Tabla 4: Descripción variable del Data_Admisión

Tipo variables		Admisión	
Matricula	int64	Carrera	object
Matemática	int64	Colegio_Egreso	object
Lenguage	int64	Cohorte	int64
Ciencias	int64	Estado	int64
NEM	float64	Sem. Estado	float64
Preferencia	int64	Sexo	object
Ponderado	object	**dtype**	**object**
Cod. Carrera	int64		

Fuente: Elaboración propia

Data_ Alumnos inscritos por asignaturas (Alu-Insc-Asi) - PAC-1, PAC-2 y PAC_3

Ilustración 2: Análisis descriptivo Data_ Alumnos inscritos por asignaturas (Alu-Insc-Asi); PAC-1.

	crr_Cod	pac_Cod	asi_Cod	asi_Secc	not_Tip	Freq	uac_Id
count	64573.000000	6.457000e+04	64570.000000	64570.000000	64570.000000	61296.000000	64567.0
mean	3314.318337	2.010442e+06	560122.691048	2.050534	0.024408	1.312810	100.0
std	22.501328	2.027956e+03	125246.017950	4.455260	0.211325	0.666201	0.0
min	100.000000	2.007100e+06	101230.000000	0.000000	0.000000	1.000000	100.0
25%	3310.000000	2.009100e+06	520141.000000	0.000000	0.000000	1.000000	100.0
50%	3313.000000	2.010230e+06	521230.000000	1.000000	0.000000	1.000000	100.0
75%	3318.000000	2.012100e+06	530023.000000	3.000000	0.000000	1.000000	100.0
max	3327.000000	2.017110e+06	999328.000000	99.000000	2.000000	8.000000	100.0

Fuente: Elaboración propia con lenguaje Python

Ilustración 3: Análisis descriptivo Data_ Alumnos inscritos por asignaturas (Alu-Insc-Asi); PAC-2.

	crr_Cod	pac_Cod	asi_Cod	asi_Secc	not_Tip	Freq	uac_Id
count	57407.000000	5.739500e+04	57395.000000	57395.000000	57395.000000	53276.000000	57383.0
mean	3303.099657	2.014304e+06	548465.656939	2.640927	0.293005	1.525058	100.0
std	59.190640	1.290943e+03	104127.591264	5.540408	0.681619	1.005639	0.0
min	100.000000	2.009100e+06	101101.000000	0.000000	0.000000	1.000000	100.0
25%	3309.000000	2.013200e+06	520145.000000	1.000000	0.000000	1.000000	100.0
50%	3313.000000	2.014120e+06	525147.000000	2.000000	0.000000	1.000000	100.0
75%	3317.000000	2.015120e+06	530023.000000	3.000000	0.000000	2.000000	100.0
max	3327.000000	2.017110e+06	999335.000000	99.000000	2.000000	11.000000	100.0

Fuente: Elaboración propia con lenguaje Python

Ilustración 4: Análisis descriptivo Data_ Alumnos inscritos por asignaturas (Alu-Insc-Asi); PAC-3.

	crr_Cod	pac_Cod	asi_Cod	asi_Secc	not_Tip	Freq	uac_Id
count	21525.000000	2.152300e+04	21523.000000	21523.000000	21523.000000	20582.000000	21521.0
mean	3304.112706	2.016244e+06	534294.268922	3.375784	1.492682	1.416723	100.0
std	46.927795	6.764187e+02	70007.478522	5.281583	0.850709	0.893665	0.0
min	100.000000	2.011110e+06	101513.000000	0.000000	0.000000	1.000000	100.0
25%	3309.000000	2.016110e+06	510226.000000	1.000000	1.000000	1.000000	100.0
50%	3312.000000	2.016120e+06	525147.000000	2.000000	2.000000	1.000000	100.0
75%	3317.000000	2.017100e+06	530023.000000	5.000000	2.000000	2.000000	100.0
max	3327.000000	2.017110e+06	999335.000000	99.000000	2.000000	13.000000	100.0

Fuente: Elaboración propia con lenguaje Python

Tabla 5: Tipo variable Data_ Alumnos inscritos por asignaturas; PAC-1-2-3

Tipo variables		PAC-1; PAC-2; PAC-3	
per_Cod	object	per_Cod	object
alu_Cod	object	alu_Cod	object
crr_Cod	int64	crr_Cod	int64
pac_Cod	float64	pac_Cod	float64
asi_Cod	float64	asi_Cod	float64
asi_Secc	float64	asi_Secc	float64
Nota	object		

Fuente: Elaboración propia

Data_Asignaturas Complementarias

Ilustración 5: Análisis descriptivo Data_ Asignaturas complementarias.

	CREDITOS	HORAS_TEORICAS	HORAS_PRACTICAS	HORAS_LABORATORIO	DIAS_TERRENO	HORAS_TRABAJO_ACADEMICO	CREDITOS_TRANSFERIBLES
count	198.000000	198.000000	198.000000	198.000000	198.0	198.000000	20.0
mean	1.964646	1.070707	1.893939	0.065657	0.0	0.404040	0.0
std	1.019487	1.194340	1.797873	0.416332	0.0	1.399196	0.0
min	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.0
25%	1.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.0
50%	2.000000	1.000000	2.000000	0.000000	0.0	0.000000	0.0
75%	2.000000	2.000000	2.000000	0.000000	0.0	0.000000	0.0
max	5.000000	8.000000	14.000000	4.000000	0.0	9.000000	0.0

Fuente: Elaboración propia con lenguaje Python

Tabla 6: Tipo variables Data_ Asignaturas complementarias.

Tipo variables		Asignaturas complementarias	
codigo_asignatura	int64	dias_terreno	int64
nombre	object	horas_trabajo_academico	int64
depto.	int64	modalidad	object
nombre_1	object	plev	object
duracion	object	creditos_transferibles	float64
creditos	int64	fecha_creacion	object
horas_teoricas	int64	periodo	int64
horas_practicas	int64	agno	int64
horas_laboratorio	int64	**dtype:**	**object**

Fuente: Elaboración propia

Data_Rendimiento

Ilustración 6: Análisis descriptivo Data_ Rendimiento.

	Rut	Matrícula	Cohorte	Quintil	Preferencia	Puntaje Ranking	Puntaje Ponderado	Puntaje NEM	Puntaje Matemáticas	Puntaje Lenguaje
count	1.363600e+04	1.363500e+04	13635.000000	13635.000000	13550.000000	4791.000000	11998.000000	12949.000000	12949.000000	12949.000000
mean	5.471566e+07	3.933541e+09	2009.593766	1.601173	0.472325	707.574828	6231.920820	663.141555	678.423739	617.081396
std	1.750367e+06	5.061623e+06	5.054920	1.771309	0.918284	88.911672	18453.085676	77.260276	60.925598	67.943378
min	1.176500e+04	3.922701e+09	2000.000000	0.000000	0.000000	0.000000	0.000000	0.000000	439.000000	305.000000
25%	5.341491e+07	3.928971e+09	2005.000000	0.000000	0.000000	648.000000	620.000000	620.000000	637.000000	571.000000
50%	5.484682e+07	3.933927e+09	2010.000000	1.000000	0.000000	714.000000	658.000000	667.000000	673.000000	616.000000
75%	5.609119e+07	3.937954e+09	2014.000000	3.000000	1.000000	775.000000	699.000000	723.000000	715.000000	662.000000
max	6.274044e+07	3.941294e+09	2017.000000	7.000000	9.000000	850.000000	81630.000000	826.000000	850.000000	837.000000

Fuente: Elaboración propia con lenguaje Python

Tabla 7:Tipo Variable Data_ Rendimiento.

Tipo variables		Asignaturas complementarias	
Rut	int64	Puntaje Ranking	float64
Matrícula	float64	Puntaje Ponderado	float64
Cohorte	float64	Puntaje NEM	float64
Género	object	Situación Actual	object
Quintil	float64	Puntaje Matemáticas	float64
Carrera	object	Puntaje Lenguaje	float64
Grupo Dependencia	object	Nota 7	object
Promedio Notas	object	Promedio Curricular	object
Vía Ingreso	object	Promedio Ponderado	object
Preferencia	float64	**dtype**	**object**

Fuente: Elaboración propia

Data_Situación

Ilustración 7: Análisis descriptivo Data_ Situación.

	Rut	Matricula	Cohorte	Periodo
count	1.635000e+05	1.635000e+05	163500.000000	163500.000000
mean	5.437633e+07	3.932391e+09	2008.440569	201112.540526
std	1.586436e+06	4.680564e+06	4.674637	442.240329
min	3.712916e+07	3.922701e+09	2000.000000	200001.000000
25%	5.328228e+07	3.928923e+09	2005.000000	200801.000000
50%	5.447600e+07	3.932924e+09	2009.000000	201202.000000
75%	5.554449e+07	3.935978e+09	2012.000000	201501.000000
max	6.274044e+07	3.941294e+09	2017.000000	201702.000000

Fuente: Elaboración propia con lenguaje Python

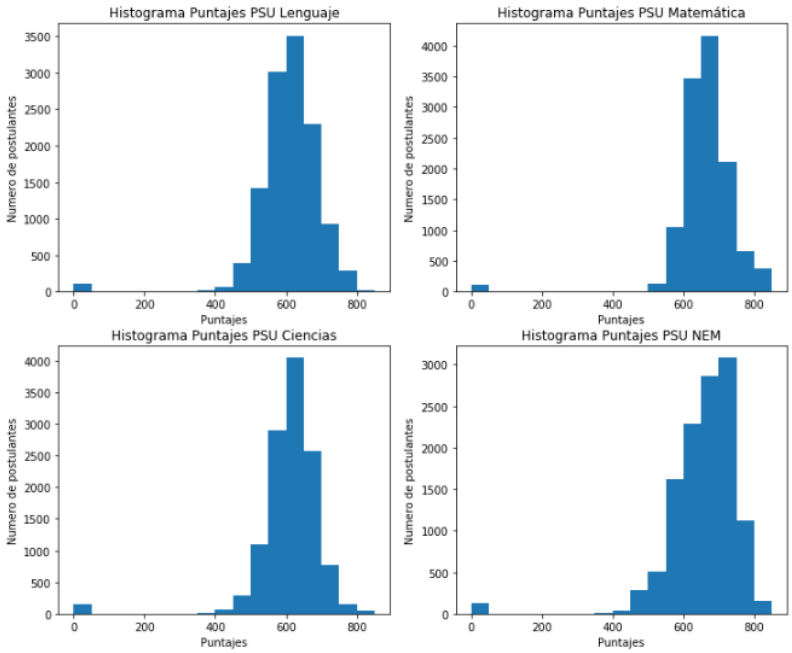
Tabla 8:Tipo variable Data_ Situación.

Tipo variables		Situación	
Rut	int64	Situation Nombre	object
Matricula	int64	Periodo	int64
Cohorte	int64	Fecha	object
Carrera	object	**dtype**	object

Fuente: Elaboración propia

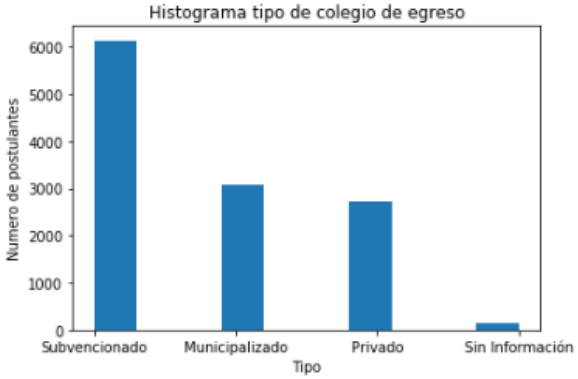
De acuerdo a las figuras 1 y 2, la información contenida en los presentes histogramas muestra la dispersión de los puntajes psu por las distintas pruebas y según tipo de colegio de egreso.

Figura: 1: Histograma Puntaje PSU por Disciplina.



Fuente: Elaboración propia con lenguaje Python

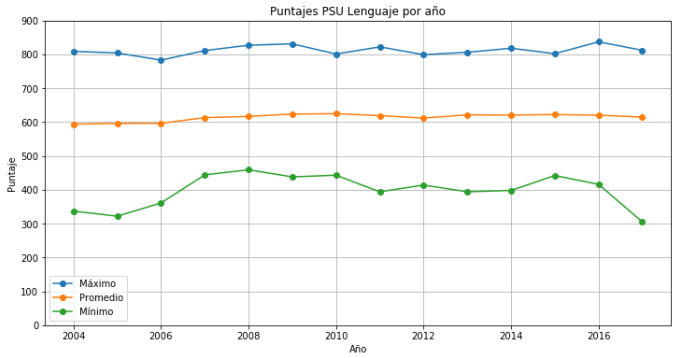
Figura: 2: Número de postulantes por Tipo Colegio de Egreso.



Fuente: Elaboración propia con lenguaje Python

Por otro lado, la siguiente gráfica contenidas en las figuras 3 a la 6 representa la información contenida en relación al puntaje psu_promedio, psu_máximo y psu_mínimo por año para los puntajes de lenguaje, matemática, ciencia y NEM.

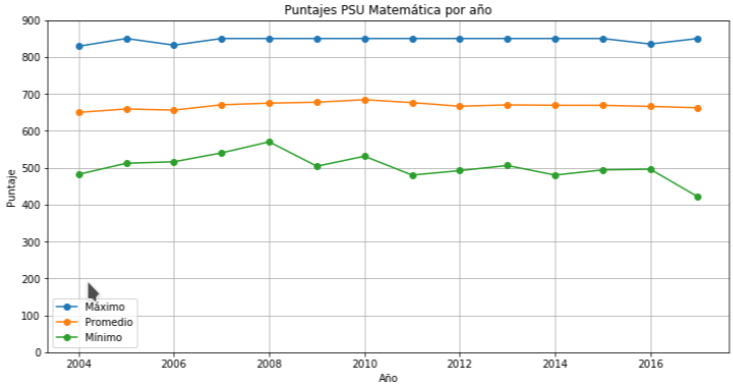
Figura: 3: Gráfico Puntaje PSU Lenguaje por Año



Fuente: Elaboración propia con lenguaje Python

En la visualización de la figura 4 se puede observar la presencia de valore inválidos (puntajes PSU iguales a cero), para los cuales hace evidente la aplicación de técnicas de imputación de datos que más se ajuste al análisis.

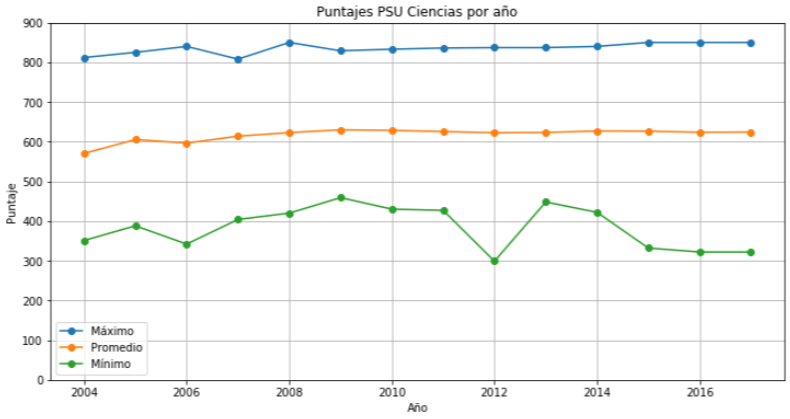
Figura: 4: Gráfico Puntaje PSU Matemática por Año



Fuente: Elaboración propia con lenguaje Python

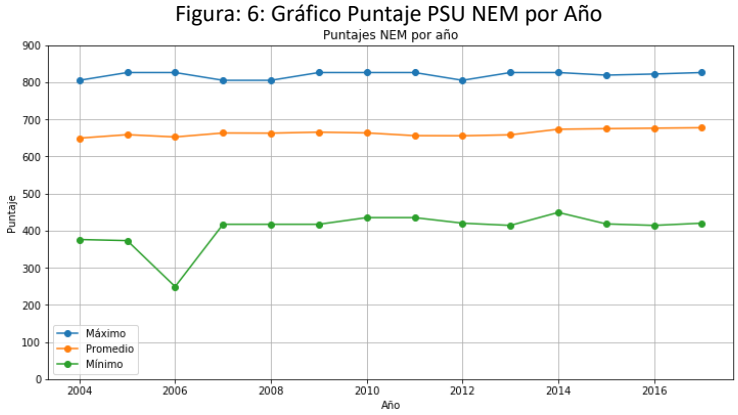
Mientras que para las siguientes visualizaciones gráfica presentes en las figuras 5 a la 6 en relación a los puntajes por año estos valores ceros fueron eliminados.

Figura: 5: Gráfico Puntaje PSU Ciencias por Año



Fuente: Elaboración propia con lenguaje Python

Figura: 6: Gráfico Puntaje PSU NEM por Año



Fuente: Elaboración propia con lenguaje Python

Predicción

Dada la clasificación entregada a priori*; "bueno estudiante", "muy buen estudiante" y "excelente estudiante"*, entendiéndose por *buen estudiante* aquel que se Titula, por *muy buen estudiante* aquel que se Titula a lo más en T+1 años (T= duración carrera declarada) y por *excelente estudiante* aquel que se Titula en tiempo T. Procederemos a clasificar en base a lo anterior y observaremos que variables que tienen más peso en dicha clasificación. Para tales efectos, como primeros pasos tentativos de exploración de los datos procederemos a aplicar aprendizaje supervisado a través de algoritmos de árbol de decisión con el objetivo de comenzar a describir a través de los datos entregados en la clasificación predefinida. Adicionalmente consideramos que el algoritmo es útil y claro para identificar cuáles son las características relevantes para la definición *"buen estudiante", "muy buen estudiante" y "excelente estudiante"*.

ANÁLISIS Y MODELO DEL DATASET

Procedimiento

En el siguiente apartado procederemos a describir el proceso de exploración, limpieza y análisis, modelamiento y visualización de la data. El objetivo es presentar los experimentos realizados, los resultados obtenidos, aprendizajes (dificultades encontradas y cómo se superaron) y finalmente las conclusiones obtenidas.

Recapitulando, la exploración realizada de la primera etapa de este informe, el presente DataSet pertenece a las carreras dictadas por la Facultad de Ingeniería de la Universidad de Concepción (UdeC), la cual tiene un carácter histórico (2000-2017) y contiene información asociada a la condición de ingreso del estudiante y su respectivo desempeño académico. El DataSet está contenida en 5 archivos denominados: admisión, alumnos inscritos por asignaturas, asignaturas complementarias, rendimiento y situación. La base de dato previamente mencionada tiene por objetivo servir de insumo para verificar la clasificación entregada a priori*; "bueno estudiante", "muy buen estudiante" y "excelente estudiante"*, e identificar cuál combinación de variables explica, en mayor medida, la clasificación asociada.

Para tales efectos, se decide trabajar con los archivos admisión, rendimiento y situación, dado que estos concentran la principal información para poder determinar tanto el perfil de egreso e ingreso del estudiante, como las variables que se encuentran correlacionadas entre ambos perfiles.

Por consiguiente, para determinar el perfil de egreso del estudiante trabajamos con el archivo Situaciones_MV_1.csv, con el cual se desarrolló en los siguientes pasos:

Duración en la carrera por alumno: una vez cargado el historial de situaciones para cada alumno, se obtuvo el año y semestre de la primera situación registrada y la última situación registrada, en conjunto a su año y semestre respectivo. Con ambos años, se obtiene la duración en la carrera de cada alumno, procedimiento presentado en ilustración 8.

Ilustración 8: Duración en la Carrera por Alumno

Rut	Matricula		Carrera	Situacion Nombre	Año inicio	Semestre inicio	Año fin	Semestre fin	Años
37129156	3932621423	90012	INGENIERIA CIVIL MECANICA	BAJA ACADEMICA	2009	1	2009	2	1.0
	3937539525	154963	INGENIERIA CIVIL INFORMATICA	BAJA POR NO INSCRIPCION	2015	1	2015	1	0.5
45059390	3924701423	15138	INGENIERIA CIVIL ELECTRONICA	TITULADO	2002	1	2005	2	4.0
45611602	3926922499	22671	INGENIERIA CIVIL - PLAN COMUN	SUSPENSION DE ESTUDIOS SOLICITADA	2003	1	2003	2	1.0
46046835	3928171421	30911	INGENIERIA CIVIL MECANICA	BAJA POR NO INSCRIPCION	2004	1	2004	2	1.0

Fuente: Elaboración propia con lenguaje Python

Filtrado: luego del procedimiento anterior, se filtran los estudiantes que se encuentran todavía estudiando, se han cambiado de carrera o han suspendido temporalmente los estudios, debido a que estas situaciones curriculares no pueden ser utilizadas en el entrenamiento del modelo, las que corresponden a los siguientes estados:

- AUTORIZADO PARA INSCRIBIR
- CAMBIO DE CARRERA
- BAJA NO ACADEMICA (DEFUNCION)
- SUSPENSION TRANSITORIA
- SUSPENSION DE ESTUDIOS SOLICITADA

Creación columna "Clasificación": finalmente, se crea la columna de *Clasificación* con el objetivo de clasificar al alumno de acuerdo a la clasificación ya entregada (*Buen estudiante, muy buen estudiante, excelente estudiante)*. Sin embargo, se crea y agrega una nueva "etiqueta de clasificación de estudiante" denominada como *mal estudiante*, la cual abarca todos aquellos estudiantes que no finalizan la carrera. Procedimiento presentado en ilustración 9.

Ilustración 9: Creación columna "Clasificación"

Rut	Matricula		Carrera	Situacion Nombre	Año inicio	Semestre inicio	Año fin	Semestre fin	Años	Clasificacion
37129156	3932621423	90012	INGENIERIA CIVIL MECANICA	BAJA ACADEMICA	2009	1	2009	2	1.0	Mal estudiante
	3937539525	154963	INGENIERIA CIVIL INFORMATICA	BAJA POR NO INSCRIPCION	2015	1	2015	1	0.5	Mal estudiante
45059390	3924701423	15138	INGENIERIA CIVIL ELECTRONICA	TITULADO	2002	1	2005	2	4.0	Excelente estudiante
46046835	3928171421	30911	INGENIERIA CIVIL MECANICA	BAJA POR NO INSCRIPCION	2004	1	2004	2	1.0	Mal estudiante
46738142	3924924508	8018	INGENIERIA CIVIL METALURGICA	INSCRIPCION NO VALIDADA	2001	1	2003	1	2.5	Mal estudiante

Fuente: Elaboración propia con lenguaje Python

Para determinar el perfil de ingreso del estudiante trabajamos con el archivo admision.csv, con el cual se desarrolló en el siguiente paso:

Filtrado: se filtran los datos de admisión, por matrículas, para aquello estudiantes ya clasificados en la data anterior. Procedimiento presentado en ilustración 10.

Ilustración 10: Tabla Admisión filtrada

	Cohorte	Matricula	Matematica	Lenguage	Ciencias	NEM	Colegio_Egreso	Carrera
4	2014	3937959540	611	618	587	672.0	Subvencionado	Ingeniería Civil Electrónica
16	2014	3937563493	630	625	619	601.0	Subvencionado	Ingeniería Civil Metalúrgica
20	2014	3937972398	734	646	714	706.0	Subvencionado	Ingeniería Civil Química
25	2016	3939933868	573	531	578	547.0	Subvencionado	Ingeniería Civil en Telecomunicaciones
44	2014	3937563574	659	613	647	717.0	Privado	Ingeniería Civil Química

Fuente: Elaboración propia con lenguaje Python

Luego procedemos a trabajar el archivo rendimiento_MV.csv con información del rendimiento académico asociado a cada estudiante. Con este archivo se desarrolló en los siguientes pasos:

Filtro: para efectos del análisis se filtran los datos de rendimiento de la misma forma que los archivos anteriores, es decir, por Rut y matrícula. Sin embargo, al explorar los datos se detectó que algunas columnas tenían valores Nan y en particular la columna Colegio_Egreso, había un ítem etiquetado como "SIN INFORMACIÓN".

Para tales efectos en el caso de los Nan, al ser una cantidad menor en las columnas de interés los eliminamos. Para el caso de "SIN INFORMACIÓN", procedemos a calcular que porcentaje tiene la etiqueta "SIN INFORMACIÓN" del total de los datos. Procedimiento presentado en ilustración 11.

Ilustración 11: Creación columna "Clasificación"

Colegio_Egreso	Rut	Matricula	Quintil	Via Ingreso	Promedio Curricular	Colegio_Egreso	Carrera
False	8857	8857	8857	8857	8857	8857	8857
True	615	615	615	615	615	615	615

Fuente: Elaboración propia con lenguaje Python

Como se observa en la ilustración 11 el total de la data es 9472 y la etiqueta evaluada contiene 615 datos, por lo tanto, el porcentaje de la etiqueta sin información perteneciente a la columna Colegio_Egreso solo representa un 6,5% del total de la data. A pesar de ello, para asegurarnos de no perder información, entonces remplazamos el dato de acuerdo a la moda asociada al quintil. Para este caso resulto ser el Quintil 0 asociado a la etiqueta "PARTICULAR SUBVENCIONADO", etiqueta que se utiliza para el remplazo.

Una vez terminada la limpieza, ajuste y análisis de los datos procedemos a juntar las tablas de rendimiento y admisión preprocesadas. Recordemos que a estas tablas se eliminaron los Nan, por lo tanto, esto implico filtrar la tabla de situación de los estudiantes para utilizar los datos de aquellos estudiantes que no hayan sido eliminados en las dos tablas anteriormente preprocesadas. Los resultados se presentan en las ilustraciones 12 y 13.

Ilustración 12: Atributos de Estudiantes

Rut	Quintil	Via Ingreso	Colegio_Egreso	Matematica	Lenguage	Ciencias	NEM	Carrera
37129156	0	10	2	0	0	0	0.0	10
37129156	0	10	2	0	0	0	0.0	9
46738142	0	6	1	693	624	0	536.0	11
46738142	0	6	1	693	624	0	536.0	4
47877958	0	16	2	673	582	696	723.0	11

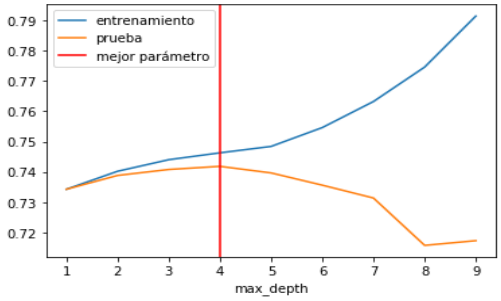
Ilustración 13: Etiqueta o Clasificación para cada Estudiante

```
Rut
37129156     Mal estudiante
37129156     Mal estudiante
46738142     Mal estudiante
46738142    Buen estudiante
47877958    Buen estudiante
Name: Clasificacion, dtype: object
```

Fuente: Elaboración propia con lenguaje Python

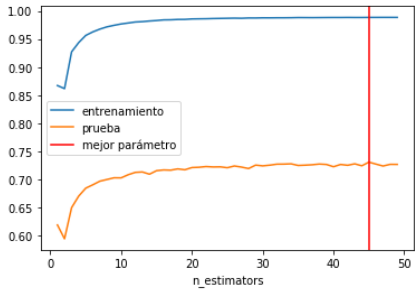
Luego de la unión anteriormente descrita, se procede a buscar los hiperparámetros para optimizar los modelos seleccionados dado el conjunto de datos. En otras palabras, se aplica validación cruzada a los modelos de clasificación seleccionados. Los que puedes observarse en las figuras 7, 8 y 9.

Figura: 7: Gráficos de Validación Cruzada con Hiperparámetro - Profundidad Decision Tree



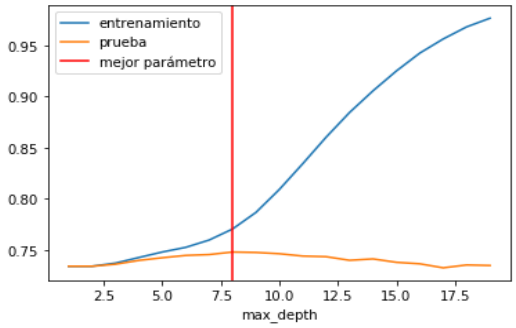
Fuente: Elaboración propia con lenguaje Python

Figura: 8: Gráficos de Validación Cruzada con Hiperparámetro – Profundidad Random Forest



Fuente: Elaboración propia con lenguaje Python

Figura: 9: Gráficos de Validación Cruzada con Hiperparámetro - Número de Estimadores



Fuente: Elaboración propia con lenguaje Python

Posteriormente se crean los clasificadores con los hiperparámetros obtenidos anteriormente para los algoritmos de Decision Tree y Random Forest. Cabe mencionar que no se consideró el algoritmo KNN debido a que este NO tiene el atributo "feature_importances" o uno similar, el cual nos entrega la importancia o relevancia de los atributos que mejor explican el algoritmo de clasificación. Los resultados se presentan en la ilustración 14.

Ilustración 14: Tabla de Importancia de cada Atributo en el Algoritmo

	Quintil	Via Ingreso	Colegio_Egreso	Matematica	Lenguage	Ciencias	NEM	Carrera
Decision Tree	0.128533	0.024239	0.000000	0.380666	0.006905	0.000000	0.210944	0.248712
Random Forest	0.080492	0.063107	0.033126	0.202628	0.065815	0.148978	0.209722	0.196132

Fuente: Elaboración propia con lenguaje Python

Finalmente calculamos el "accuracy", es decir la precisión de los algoritmos elegidos con el fin de garantizar que la información sea lo más correcta posible. Los resultados se presentan en la ilustración 15.

Ilustración 15: Accuracy Algoritmo Decision Tree y Random Forest

	accuracy
Decision Tree	0.745055
Random Forest	0.758310

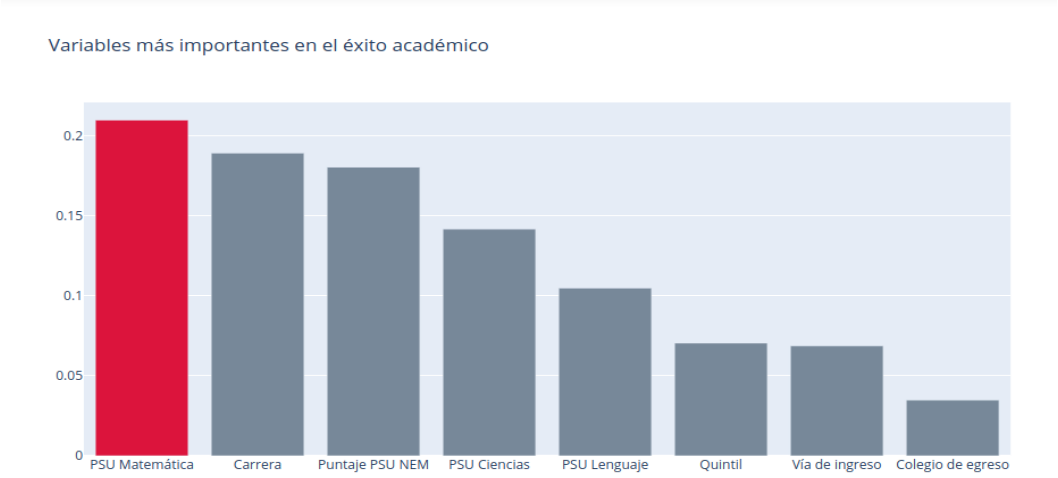
Fuente: Elaboración propia con lenguaje Python

CONCLUSIÓN

De acuerdo al desarrollo presentado, se concluye que el mejor algoritmo para obtener la importancia de cada una de las variables que explican el modelo es el Random Forest, lo que se evidencia con el accuracy de un 76%.

Por otro lado, las variables que más explican a la clasificación de estudiantes son: en primer lugar, la prueba de selección universitaria de matemática, luego le sigue la carrera a la que ingresa el alumno y en tercer lugar en puntaje correspondiente al NEM, lo que se puede visualizar en la figura 10.

Figura: 10: Gráfico – Variables que Mejor Explican Clasificación



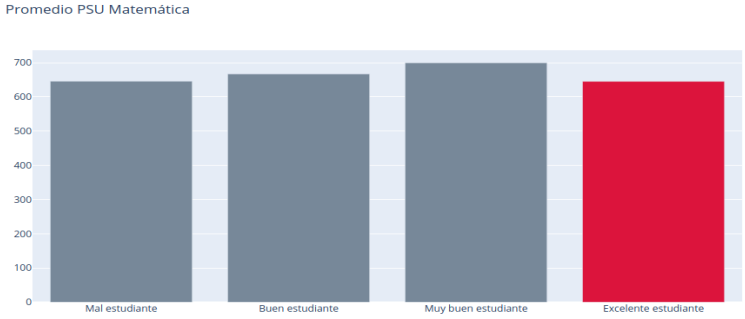
Fuente: Elaboración propia con lenguaje Python

De acuerdo a la figura 10 la variable Colegio_Egreso tiene baja incidencia en la clasificación asignada al éxito o fracaso académico del estudiante, lo cual se contrapone con los paradigmas convencionales asociadas a las actuales brechas en relación a la calidad e equidad de la educación secundaria entregada a nivel nacional por los diferentes tipos de instituciones existentes en el país. Entre las posibles explicaciones a este hallazgo puede ser que las pruebas de selección universitaria realmente no se correlaciona con los contenidos entregados por las instituciones de educación superior, dicho en otras palabras, existe la posibilidad que basta con que un alumno realice una buena preparación para la prueba de matemática, a través de un ente externo experto en la materia como lo son los pre-universitarios y con esos conocimientos nivele los posibles vacíos provenientes de las instituciones secundarias más precarizadas en contenidos.

En la misma línea, llama la atención que la variable prueba PSU de lenguaje tenga una mayor incidencia en el modelo que la variable anteriormente analizada, esto se puede deber a que la prueba en sí mide el desarrollo de habilidades cognitivas, la que tiene directa relación con las capacidades cognitivas desarrolladas, las que a su vez, guardan relación con el procesamiento de la información, es decir, en relación a mantener y distribuir la atención, percepción, memoria, resolución de problemas, comprensión, establecimientos de analogías, entre otras. En resumen, aquellos estudiantes que tenga mejores resultados en esta prueba significan que tiene un mayor desarrollo

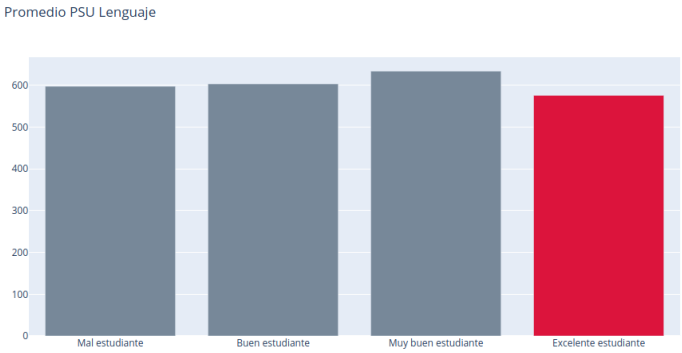
de las habilidades y capacidades cognitivas derivando en un mejor éxito académico en cualquier disciplina.

Figura: 11: Gráfico – Promedio PSU Matemática por clasificación de estudiante



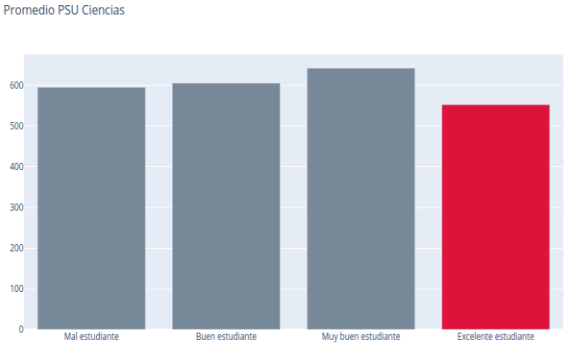
Fuente: Elaboración propia con lenguaje Python

Figura: 12: Gráfico – Promedio PSU Lenguaje por clasificación de estudiante



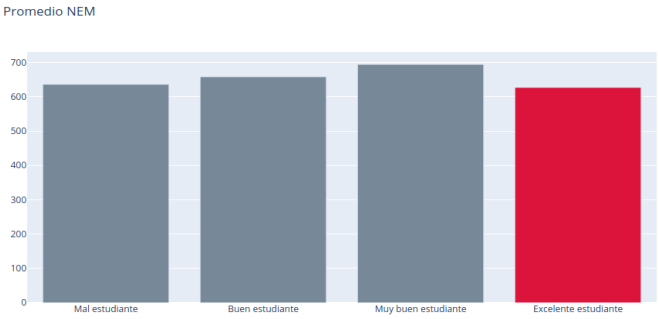
Fuente: Elaboración propia con lenguaje Python

Figura: 13: Gráfico – Promedio PSU Ciencia por clasificación de estudiante



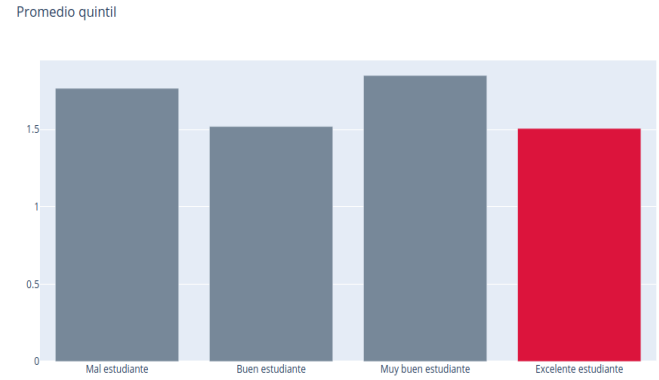
Fuente: Elaboración propia con lenguaje Python

Figura: 14: Gráfico – Promedio NEM por clasificación de estudiante



Fuente: Elaboración propia con lenguaje Python

Figura: 15: Gráfico – Promedio Quintil por clasificación de estudiante



Fuente: Elaboración propia con lenguaje Python

Otra conclusión detectada en base al análisis de los datos que se utilizaron para crear el modelo es que existe una correlación entre las pruebas de selección universitarias y la clasificación dada del éxito o fracaso académico del estudiante, exceptuando la etiqueta de *excelente estudiante*. Esto implica que este último, debiera analizarse de forma separada a las otras clasificaciones para determinar si esa etiqueta es la correcta o bien, debiera tomarse otros parámetros para asociar las clasificaciones del éxito o fracaso. El mismo análisis fue realizado con las variables NEM y Quintil. Esta conclusión se puede cotejar en las siguientes figuras 11 a la 15.