UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA
SEDE VIÑA DEL MAR

Taller 2: Aprendizaje Supervisado – Modelos de Clasificación
Ciencia de Datos

Profesor: Gabriel Jara
Primer Semestre 2024

El objetivo del Taller 2 es desarrollar un modelo de clasificación que pronostique la titulación de estudiantes de la Sede Viña del Mar de la USM.

Se le provee el set de datos "Taller_2_Titulacion_DatosTaller" en archivo de texto plano separado por "," que contiene 77 campo, sobre poco más de 3,000 registros. Se presenta al final de este documento tabla con descripción general de estos atributos.

Se trata de un levantamiento de datos de estudiantes ingresados a la USM, Sede Viña del Mar entre los años 2004 y 2009, y su estado de titulación al año 2014. No existe en el set de datos atributos que permitan individualizar personas ni instituciones fuera de la USM.

El último atributo en el set de datos corresponde al estado de titulación al año 2014, pudiendo ser valor "SÍ" o "NO" según sea el caso (Sí es que sí se titula).

Se suplementa además un segundo archivo "Taller_2_Titulacion_Evaluación" que contiene los datos de 568 estudiantes con los mismos atributos, pero sin el último que corresponde a estado de titulación.

Ambos archivos comparten un atributo Id numérico correlativo-incremental, que es el primer atributo, y podrá verificar que los Id son únicos entre ambos archivos. Por ejemplo, si en el primer set de datos no está el Id 2 es porque debe estar en el set de datos para evaluación.

El objetivo del Taller 2 es desarrollar un modelo de clasificación que permita clasificar correctamente a la mayor cantidad posible de observaciones de la muestra de evaluación. Los estudiantes de la asignatura Ciencia de Datos pueden realizar el taller en parejas o individualmente.

ATENCIÓN: SIGA ESTAS INSTRUCCIÓN POR FAVOR

- El informe que deberá presentar será un Jupyter Notebook, el que deberá consumir los datos y generar los resultados. El nombre del archivo debe contener el nombre de los participantes, y además deben estar esos nombres declarados en la primera celda.

- Al momento de cargar su informe a Aula se le solicita que cada celda venga ejecutada, es decir que no sea necesario volver a ejecutar una celda para poder leer sus resultados (tenga presente que sí podría ser ejecutado y debe funcionar).
- Comprima su Jupyter Notebook junto al archivo csv que se solicita generar en una actividad, ambos en un archivo zip, rar o 7z. Suba al Aula el archivo comprimido.
- No debe cargar a Aula los datos que consume su informe (muestras), puede asumir que el profesor ya los tiene.
- Prepare su informe organizado en cuatro Actividades (instrucciones más abajo), incluya títulos y celdas específicas para cada actividad solicitada (una o varias celdas por actividad). Puede usar los Jupyter Notebook entregados en clase como ejemplo del cómo se debiera ver su informe.
- **Súbalo a Aula antes del domingo 16 de junio a las 23:59.**

TALLER 2

Actividad 1: Preprocesamiento (10 puntos)

Realice las actividades que considere necesarias para la preparación de los datos. Comente brevemente alguna de estas actividades, que considere más relevante.

Actividad 2: Estrategia de Validación (10 puntos)

Seleccione una de las dos estrategias de validación vistas en clase, y organice su set de datos de acuerdo con la estrategia que haya elegido. Tenga presente que la etapa de testeo la puede considerar como ya cubierta por el set de datos de evaluación que se le ha provisto, de modo que lo que realmente le importa es organizar el entrenamiento y validación que usted hará. Comente brevemente sobre la estrategia que ha seleccionado, y procure ser consistente en su código Python con lo que declara, así como en usar la estrategia declarada en la siguiente actividad.

Actividad 3: Selección de modelos e hiperparámetros (20 puntos)

Identifique al menos dos tipos de modelo de aprendizaje adecuados para este tipo de problema. Para cada uno de los tipos de modelo que identificó, señale al menos dos hiperparámetros, o características del modelo, que pueda configurar para producir distintas variantes del modelo.

Ejecute un proceso de búsqueda del mejor clasificador usando los dos tipos de modelo que identificó previamente, y con al menos dos parámetros o variantes por cada uno. Para un mismo tipo de modelo puede usar la herramienta de búsqueda exhaustiva vista en clase o simplemente organizar un par de experimentos con distinta configuración. Lo importante es que acumule al menos cuatro experimentos, con dos tipos de modelo y dos variantes por cada uno.

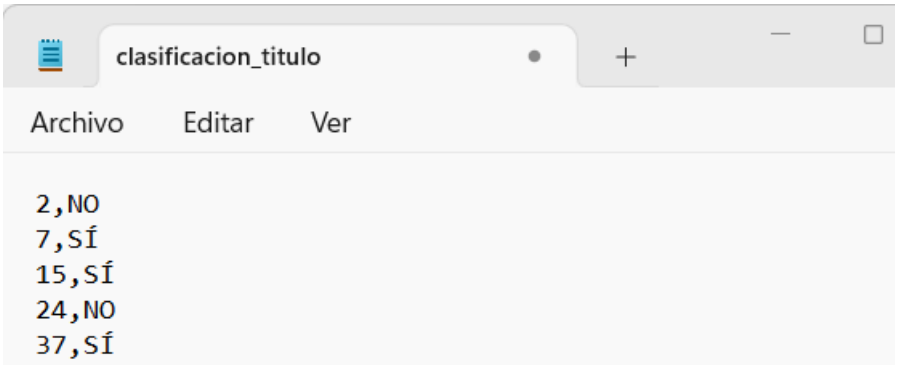
Usando los resultados obtenidos seleccione el modelo con mejor expectativa de ofrecer un buen resultado.

Actividad 4: Evaluación sobre casos nuevos (60 puntos)

Cargue el set de datos de evaluación desde su Jupyter Notebook y genere la predicción sobre las 568 observaciones. Utilice para ello el mejor modelo que haya encontrado en las actividades previas.

Desde su Jupyter Notebook genere un archivo denominado: “clasificación_título.txt”, donde se señale el Id de cada registro en la evaluación (respete el Id que ya tiene cada registro) y un “SÍ” o “NO” con que lo clasificó el modelo predictivo.

Así se vería el archivo que se solicita comprimir junto a su Jupyter Notebook y entrega en Aula. Note que no tiene encabezado.



EVALUACIÓN

Rúbrica Actividades 1 a 3:

La siguiente rúbrica se aplica a cada una de las tres primeras Actividades.

0%	No responde y/o lo que realiza no es lo solicitado.
25%	Realiza parcialmente la actividad solicitada, incompleta o con errores, sin describir las decisiones que toma ni proponer conclusiones.
50%	Realiza casi completamente la actividad solicitada, pero omitiendo algún aspecto relevante, y/o comete un error importante al tomar decisiones de diseño y ejecución de sus experimentos, consistencia con actividad anterior, interpretar resultados u obtener conclusiones.
75%	Realiza de manera básica la actividad solicitada, con errores poco relevantes en el diseño y ejecución de los experimentos, que no afectan significativamente la interpretación de resultados y conclusiones que declara.
100%	Realiza la actividad solicitada en forma correcta, describiendo y justificando las decisiones que toma, manteniendo consistencia con actividades anteriores e incluye análisis e interpretación de sus resultados que aportan a las conclusiones que declara.

Criterio de Evaluación de la Actividad 4:

Está actividad será evaluada en base al éxito en encontrar predecir la titulación ante nuevos casos. La métrica en que se evaluará será Accuracy.

La escala con que se asignará el puntaje es:

- Si su accuracy es igual o menor a 60%, su accuracy será el porcentaje que obtenga de los puntos de la actividad.
- Si su accuracy es mayor a 75% y logra el mayor accuracy entre todas las entregas, obtiene 100% de los puntos de la actividad.
- Si ningún trabajo supera 75%, este será el valor que marcará 100 puntos en la actividad.
- En cualquier otro caso recibirá puntaje proporcional al accuracy ajustado entre 60% y MÁXIMO(mejor accuracy encontrado, 75%).

Anexo: Descripción de Atributos

1	Id	Identificador del estudiante en el Taller 3
2	MAT_1SEM_PROM	Promedio Matemáticas primer semestre. Si está vacío es que no rindió una matemática el primer semestre (podría haberla desinscrito)
3	FIS_1SEM_PROM	Promedio Física primer semestre
4	ING_1SEM_PROM	Promedio Inglés primer semestre
5	ACTF_1SEM_A	Actividad Formatica Aprobada (Educación Física)
6	ACTF_1SEM_R	Actividad Formatica Reprobada (Educación Física)
7	OTRANS_1SEM_A	Otras asignaturas Aprobadas (cuenta de asignaturas)
8	OTRANS_1SEM_R	Otras asignaturas Reprobadas
9	OTRANS_1SEM_PROM	Promedio otras asignaturas
10	ESP_1SEM_A	Asignaturas especialidad Aprobadas
11	ESP_1SEM_R	Asignaturas especialidad Reprobadas
12	ESP_1SEM_PROM	Promedio asignaturas de especialidad
13	INS_1SEM	Asignaturas inscritas primer semestre (cuenta de asignaturas)
14	PROM_1SEM	Promedio primer semestre
15	MAT_2SEM_PROM	SE REPITE LOS MISMOS ATRIBUTOS PARA EL REGISTRO CURRICULAR DEL SEGUNDO SEMESTRE
16	FIS_2SEM_PROM	
17	ING_2SEM_PROM	
18	ACTF_2SEM_A	
19	ACTF_2SEM_R	
20	OTRANS_2SEM_A	
21	OTRANS_2SEM_R	
22	OTRANS_2SEM_PROM	
23	ESP_2SEM_A	
24	ESP_2SEM_R	
25	ESP_2SEM_PROM	

26	INS_2SEM	
27	PROM_2SEM	
28	INSC_AnO	Asignaturas Inscritas primer año
29	PROM_AnO	Promedio del primer año
30	GENERO	Sexo del o la estudiante
31	NOM_CARRERA	Carrera que cursa
32	TIPO_CARRERA	Técnico o Ingeniería
33	nombre_nacionalidad	Es chileno o extrajeto
34	descripcion_situacion_egreso_postulante	Situación de esgreso de Enseñanza Media (EM)
35	nombre_comuna_EM	Comuna en que estudio EM
36	nombre_provincia_EM	Provincia en que estudio EM
37	nombre_region_EM	Región en que estudio EM
38	nombre_dependencia_EM	Dependencia del establecimiento de EM
39	nombre_grupo_dependencia_EM	Grupo de dependencia del establecimiento de EM
40	descripcion_rama_educacional_EM	Rama educacional (Científico-Humanista, Técnico-profesional, otros)
41	nombre_secretaria_admision	Dónde rindió prueba de selección
42	numero_estado_civil	Estado civil al momento de postular
43	descripcion_trabajo_remunerado	TODOS ESTOS CAMPOS SON DE LA ENCUESTA DE SITUACIÓN FAMILIAR SOCIOECONÓMICA. SON AUTORREPORTADOS POR EL POSTULANTE, NO VERIFICADOS. EL NOMBRE DE CADA CAMPO ES SUFICIENTE DESCRIPCIÓN (Y NO TENGO MÁS QUE ESO TAMPOCO).
44	numero_horario_trabajo	
45	descripcion_proseguir_estudios	
46	descripcion_jefe_familia	
47	descripcion_fuente_financiamiento_estudio_superior	
48	descripcion_fuente_financiamiento_estudio_superior2	
49	inferior_ingreso_bruto_fam	
50	superior_ingreso_bruto_fam	
51	nombre_cobertura_salud	
52	descripcion_viven_padres	
53	descripcion_nivel_educacion_padre	
54	descripcion_nivel_educacion_madre	
55	descripcion_situacion_ocupacional_padre	
56	descripcion_situacion_ocupacional_madre	
57	descripcion_tipo_organismo_trabajan_padre	
58	descripcion_tipo_organismo_trabajan_madre	
59	descripcion_ocupacion_principal_padre	
60	descripcion_ocupacion_principal_madre	
61	descripcion_rama_actividad_padre	
62	descripcion_rama_actividad_madre	

63	horas_promedio_trabajo	
64	cantidad_personas_grupo_familiar	
65	cuantos_trabajan_grupo_familiar	
66	cuantos_estudian_grupo_pre_basica	
67	cuantos_estudian_grupo_basica	
68	cuantos_estudian_grupo_media_1_3	
69	cuantos_estudian_grupo_media_4	
70	cuantos_estudian_grupo_superior	
71	cuantos_estudian_grupo_otras	
72	ptje_nem	Puntaje Notas EM
73	ptje_leng	Puntaje Prueba de Lenguaje
74	ptje_mate	Puntaje Prueba de Matemáticas
75	ptje_hycs	Puntaje Prueba de Historia y Ciencias Sociales (vacío si no la rinde)
76	ptje_cien	Puntaje Ciencias (vacío si no la rinde)
77	Rotulo_Titulado	Sí estaba o No titulado al año 2014