Universidad de Valparaíso
Facultad de Ingeniería
Escuela de Ingeniería Informática

Bases de Datos III

Tarea Nº2 (20%)

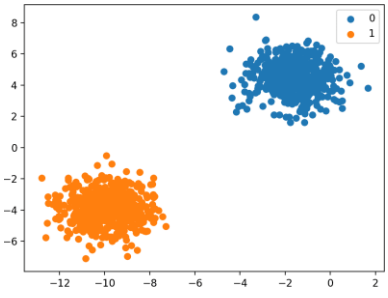
Prof: Ana Aguilera Faraco
Ayudante: Camila Araya

Octubre 2023

Instrucciones:

- La tarea Nº2 es grupal (máximo 3 integrantes, si su grupo de proyecto es de 4 integrantes pueden dividirse y formar grupos de 2 integrantes), y en caso de copia se aplicarán las sanciones correspondientes.
- El nombre del archivo debe ser "T2-NombreApellido.ipynb", debe contener el nombre y apellido de todos los integrantes separado por un guion alto ("-"). En caso de no cumplir este formato se descontarán 1 décimas en la nota final.
- Puntaje total: 100 puntos. Nota 4,0: 60 puntos.

En el aprendizaje automático, la clasificación se refiere a un problema de modelado predictivo en el que se predice una etiqueta de clase para los datos de entrada.



Según las estadísticas ¿Es posible predecir si un Pokémon es legendario? Existe el dataset de pokémon que contiene todos los pokémon en videojuegos que es la siguiente pokemonb.net. El conjunto de datos contiene un total de datos (800, 12) contiene 12 columnas y 800 filas. Se buscará predecir en base a las etiquetas *Type 1, Type 2, HP, Attack, Defense, Sp. Atk, Sp. Def, Speed y Generation* el resultado ***Legendary***.

Para la tarea debe considerar lo siguiente:

1. Conjunto de datos:
    a. Debe trabajar el siguiente Dataset
    b. Describir el dataset.
    c. Para el análisis debe transformar el dataset y tratar las variables categóricas, por ejemplo, definir valores legendarios para falso (0) y para verdadero (1).

	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
795	796	Diancie	Rock	Fairy	50	100	150	100	150	50	6	True
796	797	Mega Diancie	Rock	Fairy	50	160	110	160	110	110	6	True
797	798	Hoopa Confined	Psychic	Ghost	80	110	60	150	130	70	6	True
798	799	Hoopa Unbound	Psychic	Dark	80	160	60	170	130	80	6	True
799	800	Volcanion	Fire	Water	80	110	120	130	90	70	6	True

Antes

	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	1	45	49	49	65	65	45	1	0
1	1	1	60	62	63	80	80	60	1	0
2	1	1	80	82	83	100	100	80	1	0
3	1	1	80	100	123	122	120	80	1	0
4	2	0	39	52	43	60	50	65	1	0

Después

2. Para el problema de Clasificación debe aplicar lo siguiente:
    a. Explorar el dataset y manejar los datos (tratar con nulos y categóricos).

b. Aplicar reducción de características aplicando dos técnicas: una selección de características y una reducción de dimensionalidad (aplicar PCA).
c. Aplicar 3 algoritmos de clasificación como por ejemplo KNeighbors Classifier, AdaBoost Classifier, RandomForest Classifier, Logistic regression, Gaussian NB, Gradient Boosting, Decision Tree, entre otros. Use uno de cada familia estudiado en clases, obligatoriamente debe incluir algún metaalgoritmo. Para su elección puede utilizar para seleccionar los modelos posiblemente óptimos mediante el análisis de curva ROC.
d. Aplicar estudio de hiper parámetros en el primer clasificador a utilizar
e. Creación de los 3 modelos, entrenamiento y validación. La variable dependiente y_train debe ser Legendary la cual se busca predecir.
f. Aplicar curvas ROC, evaluar resultados y presentar reporte de clasificación junto a matriz de confusión y sus métricas de evaluación.
g. Obtener predicciones.

3. ¿Qué puede concluir respecto de los métodos utilizados?, ¿y cuál recomendaría utilizar?

4. Para las predicciones puede realizarlo de la siguiente manera (utilizar el modelo con mejor score):

```
poke = [2,0,39,52,43,60,50,65,1] #(Type 1,Type 2,HP,Attack,Defense,Sp.Atk,Sp.Def,Speed,Generation,Legendary)
poke = np.array(poke).reshape(1, -1)

y_pred = RF.predict(poke)

result = 'Legendary' if y_pred == True else 'Not Legendary'

print(result)
```

```
Not Legendary
```

```
poke = [15,4,150,190,191,180,95,95,6]#(Type 1,Type 2,HP,Attack,Defense,Sp.Atk,Sp.Def,Speed,Generation,Legendary)

poke = np.array(poke).reshape(1, -1)

y_pred = RF.predict(poke)

result = 'Legendary' if y_pred == True else 'Not Legendary'

print(result)
```

```
Legendary
```

Ejemplo de predicción con Random Forest

Universidad de Valparaíso
Facultad de Ingeniería
Escuela de Ingeniería Informática

Rúbrica de Evaluación

Presenta	Aspectos a evaluar	No aplica (0%)	Deficiente (30%)	Regular (60%)	Bueno (80%)	Destacado (100%)	Puntaje máximo del ítem
Calidad en la presentación del cuaderno	• Orden de código • Buena ortografía • Comentarios (buena redacción y entendibles) • Documentación (si utiliza librerías debe especificar)	No incluye estos aspectos.	Solo incluye solo comentarios y el orden de código no es adecuado.	1 aspecto no queda del todo claro y el código no tiene buen orden.	1 aspecto no queda del todo claro.	Todos los aspectos están claros.	5
Descripción DataSet	• Explica el Dataset (sobre qué tema trata) • Describe sus variables • Obtiene información del Dataset(variables,tamaño etc).	No incluye estos aspectos.	No se detalla bien y no cumple con 2 aspectos.	2 aspectos están presentes pero no detallados en profundidad.	Solo 1 aspecto no queda claro o no está presente.	Todos los aspectos están claros y detallados.	5
Primer análisis de Clasificación	• Realiza el análisis • Obtiene resultados • Obtiene métricas y temas planteados en el punto 2. • Estudio de hiper parámetros	No incluye estos aspectos.	No queda claro el análisis y solo incluye 1 aspecto.	2 aspectos están presentes pero no detallados en profundidad.	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros, son completos y consistentes.	35
Segundo análisis de Clasificación	• Realiza el análisis • Obtiene resultados • Obtiene métricas y temas planteados en el punto 2.	No incluye estos aspectos.	No queda claro el análisis y solo incluye 1 aspecto.	2 aspectos están presentes pero no detallados en profundidad.	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros, son completos y consistentes.	20
Tercer análisis de Clasificación	• Realiza el análisis • Obtiene resultados • Obtiene métricas y temas planteados en el punto 2.	No incluye estos aspectos.	No queda claro el análisis y solo incluye 1 aspecto.	2 aspectos están presentes pero no detallados en profundidad.	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros, son completos y consistentes.	20
Conclusiones	• Responde a las preguntas del punto 3. • Detalla de buena manera sus conclusiones. • Domina el tema de clasificación.	No incluye conclusiones y no domina el tema.	Responde 1 de las 2 preguntas y concluye que son deficientes o no aplican.	Responde 2 preguntas y obtiene conclusiones pero no domina bien el tema.	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros y detallados.	10
Predicciones	• Presenta predicciones • Utiliza bien las etiquetas y obtiene y_pred como resultado.	No incluye estos aspectos.	No queda claro	2 aspectos están presentes pero no detallados en profundidad.	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros y detallados.	5