

**INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE
OCCIDENTE**

BASES DE DATOS PARA APOYAR LA TOMA DE DECISIONES



ITESO

Universidad Jesuita
de Guadalajara

PRÁCTICA 5
APLICANDO MINERÍA DE DATOS

Presenta

IE706937 Lilia Arceli Lobato Martínez

Profesor: Victor Ortega

Fecha: 25/11/2021

Índice

Introducción	3
Propósito	3
Análisis de la fuente de datos	4
Información básica sobre Yu-Gi-Oh!	4
Juego de Cartas Coleccionables Yu-Gi-Oh!	4
Reglas del juego	4
Cartas	5
Árbol de decisión	6
Set yu-gi-oh_small	6
Implementación	6
Análisis	7
Árbol de decisión	8
Eficiencia	9
Predicción	10
Set yu-gi-oh_big	11
Implementación	11
Análisis	11
Árbol de decisión	12
Eficiencia	13
Predicción	14
Mejora en el modelo de predicción	15
Agrupamiento	16
Set yu-gi-oh_small	16
Implementación	16
Análisis de clusters	16
Agrupamiento del modelo con 2 clusters	17
Agrupamiento del modelo con 4 clusters	18
Agrupamiento del modelo con 7 clusters	20
Comparativa entre los árboles de decisión y el agrupamiento	21
Conclusiones y Aprendizajes	21
Bibliografía	21

Introducción

El concepto “minería de datos” se refiere al proceso de descubrimiento de diversos patrones y conocimientos a partir de un gran conjunto de datos. Es posible “minar” distintos tipos de datos, el tipo de dato a explotar depende mucho de la aplicación que quieres dar a los mismos datos. Prácticamente se puede minar cualquier tipo de datos siempre y cuando esos datos representan algún tipo de valor para el propósito que se desea. [1]

Las técnicas para realizar minería de datos varían, dependiendo del campo donde se desea analizar los datos. Sintetizando un poco, incluye técnicas como la estadística, aprendizaje máquina, reconocimiento de patrones, sistemas de BD, etc.



En este documento, nos enfocaremos en 2 tipos de modelos de clasificación: el árbol de decisión y el agrupamiento k-mean (K-Mean Clustering).


- Un árbol de decisión divide el conjunto de datos original en dos o más subconjuntos en cada paso del algoritmo, para aislar mejor las clases deseadas. Cada paso produce una división en el conjunto de datos y cada división puede representarse gráficamente como un nodo. La secuencia de nodos, es decir, la secuencia de divisiones se puede visualizar como un árbol, cuyas ramas definen una ruta de regla para aislar las clases deseadas[1], [2].
- El agrupamiento K-means se usa cuando tienes datos sin etiquetar (es decir, datos sin categorías o grupos definidos)[1], [3]. El objetivo de este algoritmo es encontrar grupos en los datos, con el número de grupos representados por la variable K. El algoritmo funciona de manera iterativa para asignar cada punto de datos a uno de los grupos K según las características que se proporcionan.

Propósito

Predecir el comportamiento de los indicadores de las áreas de negocio con base en la aplicación de diferentes algoritmos de minería de datos (data mining).

Set de datos

Para esta práctica se nos pidió que buscáramos dentro de diversas fuentes un set de datos:

9	ie706937	Lilia Arceli Lobato Martínez	Yu-Gi-Oh Normal Monster Cards	https://www.kaggle.com/rushikeshhirav/yugioh-normal-monster-cards 	Lunes 22-nov.2021 10:29 am
---	----------	------------------------------	-------------------------------	---	-------------------------------

Elegí dos set de datos referentes a de Yu-Gi-Oh!, ambos contienen la misma información en columnas pero con diferente tamaño en filas.

- El set de datos que llamaré yu-gi-oh_big contiene la información de 6449 tarjetas provenientes de la api YGOPRODECK [4] y fue utilizado como base para otros proyectos de entrenamiento de redes, reconocimiento de imágenes y predicción de movimientos.[5]
- El set de datos que llamaré yu-gi-oh_small contiene 478 tarjetas y solamente tiene la información de las tarjetas de monstruos de duelo (no hay tarjetas especiales, objetos o habilidades).

Análisis de la fuente de datos

Antes de comenzar la práctica, decidí buscar y entender cómo funcionan las cartas para tener un punto de comparación con los resultados que esté generando y pueda concluir información relevante.

Información básica sobre Yu-Gi-Oh!

Yu-Gi-Oh! Es una serie de manga japonesa sobre juegos escrita e ilustrada por Kazuki Takahashi. La trama sigue la historia de un chico llamado Yugi Mutou, que resuelve el antiguo Rompecabezas del Milenio. Yugi despierta un alter-ego del juego dentro de su cuerpo que resuelve sus conflictos utilizando varios juegos.[6]

La mayoría de las encarnaciones de la franquicia incluye el juego ficticio de cartas coleccionables conocido como Duel Monsters, en el que cada jugador utiliza cartas para "batirse en duelo" en una batalla simulada de "monstruos" de fantasía.

Juego de Cartas Coleccionables Yu-Gi-Oh!

Es un juego de cartas coleccionables japonés desarrollado y publicado por Konami. Se basa en el juego ficticio de Duel Monsters.

Fue nombrado el juego de cartas coleccionables más vendido del mundo por el Guinness World Records, habiendo vendido más de 22.000 millones de cartas en todo el mundo. A partir de enero de 2021, se estima que el juego ha vendido alrededor de 35 mil millones de cartas en todo el mundo y ha recaudado más de 1 billón de yenes (9,64 mil millones de dólares).

Reglas del juego

Los jugadores roban cartas de sus respectivos mazos y se turnan para jugar cartas en "el campo". Cada jugador utiliza un mazo que contiene de cuarenta a sesenta cartas, y un "mazo extra" opcional de hasta quince cartas. También hay un mazo lateral opcional de quince cartas, que permite a los jugadores intercambiar cartas de su mazo principal y/o del mazo extra entre partidas. [7]

Los jugadores están restringidos a tres cartas por mazo y deben seguir la lista de cartas prohibidas/limitadas, que restringe las cartas seleccionadas por Konami a dos, uno o cero. Cada jugador comienza con 8000 "puntos de vida", y el objetivo principal del juego es utilizar los ataques de los monstruos y los hechizos para reducir los puntos de vida del oponente.

El juego termina al alcanzar una de las siguientes condiciones:

- Los puntos de vida de un jugador llegan a cero. Si ambos jugadores llegan a cero puntos de vida al mismo tiempo, la partida termina en empate.
- Un jugador pierde si tiene que robar una carta, pero no tiene más cartas para robar en el mazo principal.
- Algunas cartas tienen condiciones especiales que desencadenan una victoria o una pérdida automática cuando se cumplen sus condiciones.

Cartas

El juego tiene 3 tipos de cartas: Cartas de monstruo, de hechizo (antes mágicas) y de trampa. Cada tipo de carta tiene un atributo y una raza/subtipo.



Cada carta contiene la siguiente información:

- **Nombre:** El nombre de la carta.
- **Tipo:** El tipo de carta (Monstruo Normal, Monstruo de Efecto, Carta Trampa, etc.)
- **Nivel:** Nivel de Invocación
- **Raza:** La raza de la carta (Guerrero, Dragón, Hechicero, etc.),
- **Atributo:** El atributo de la carta (Agua, Fuego, Viento, etc.)
- **ATK:** Los puntos de ataque de la carta
- **DEF:** Los puntos de defensa de la carta

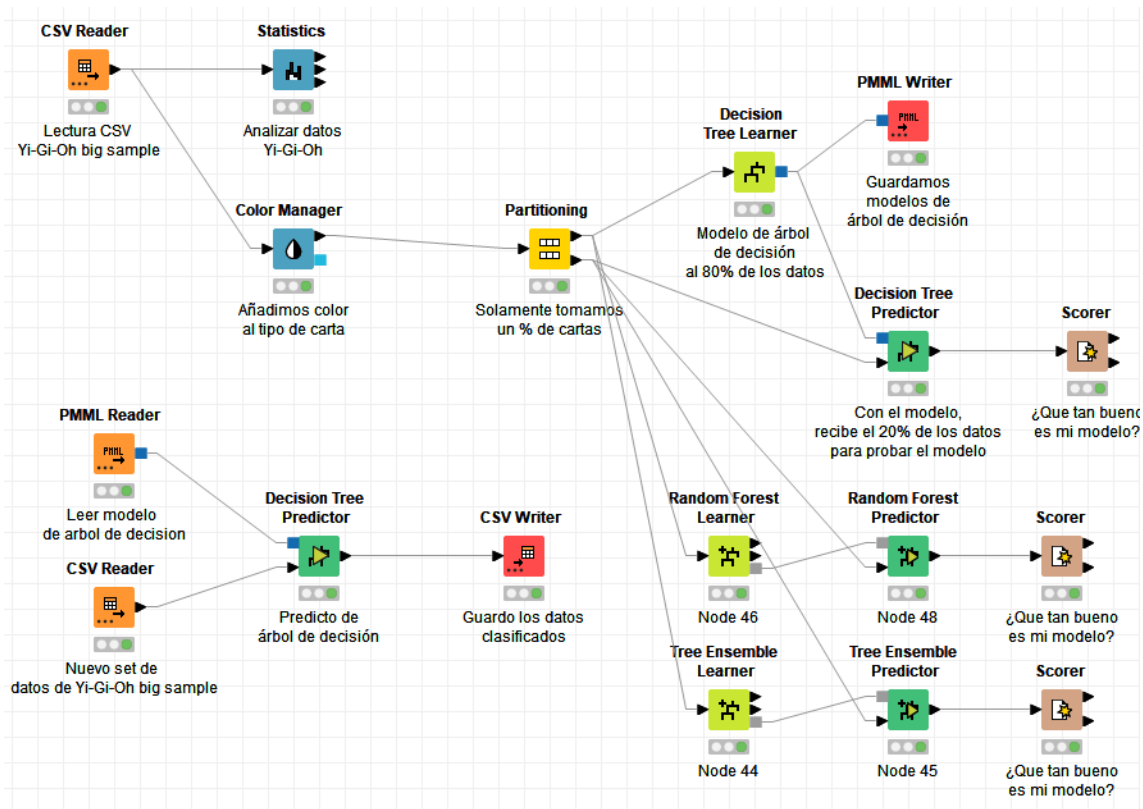
Árbol de decisión

Un árbol de decisión es un diagrama en forma de árbol que muestra la probabilidad estadística o determina un curso de acción. Muestra los pasos que se deben tomar y cómo las diferentes elecciones podrían afectar todo el proceso para llegar a un resultado.

Set yu-gi-oh_small

Implementación

Leemos el set de datos, lo particionamos y usamos el 80% de los datos para generar un árbol de decisión. Este árbol lo usamos para predecir el resultado tomando el 20% restante de la partición y comparamos los datos predichos contra los originales para obtener la eficiencia del modelo.



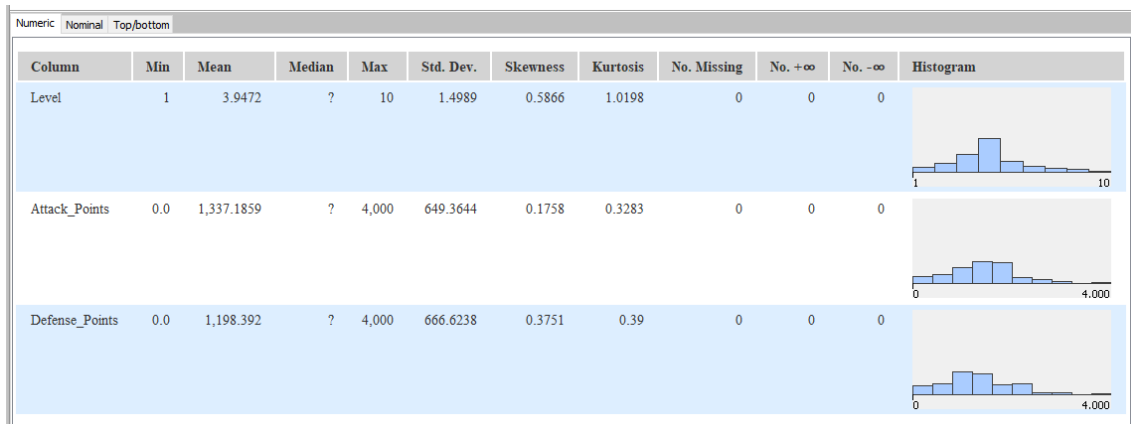
Usamos la columna de atributo ya que absolutamente todo tipo de carta contiene esta información y constituye uno de los elementos principales en las estrategias de Yu-Gi-Oh! Cada carta de monstruo puede aumentar su ataque o defensa basándose en sus atributos (Divino, Agua, Fuego, Luz, Oscuridad, Tierra y Viento) ya que ciertas combinaciones de atributos son favorables. [8]

Otro detalle importante es que los jugadores buscan tener masos balanceados ya que permite tener ataques con un valor de daño alto y formar estrategias sólidas que maximicen su defensa.

Análisis

Cuando vemos la distribución de nivel, puntos de ataque y defensa vemos que es una curva de bayes con distribución normal.

Me hace bastante sentido que las cartas están pensadas de esta forma ya que las cartas se adquieran por sobres de 5 cartas aleatorias. Con la distribución actual, se garantiza que los jugadores van a tener un mazo balanceado, pocas cartas con valores extremadamente bajos o altos.



Cuando analizamos los atributos nominales nos concentramos en el de atributos y podemos comprobar que todas nuestras tarjetas caen en alguno de los 7 tipos de categorías.

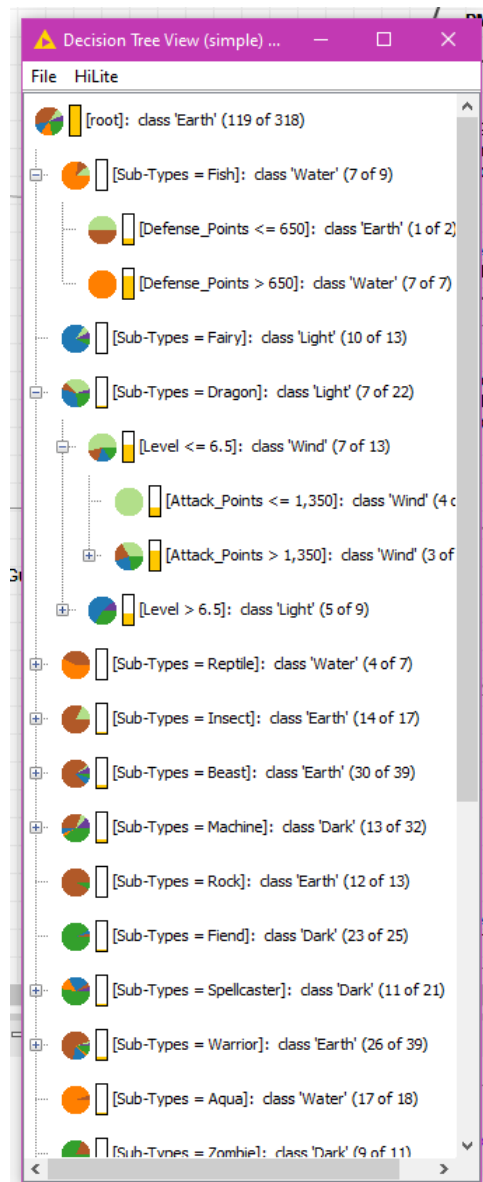
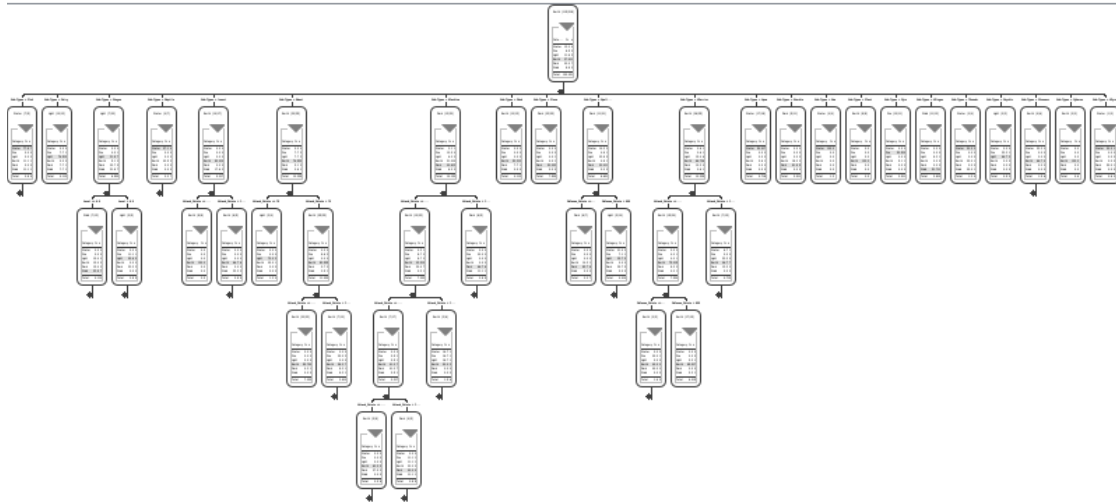


Busqué la distribución estadística oficial [9] del total de cartas de monstruo y encontré que se similar a la que tenemos en nuestro set de datos:

- 26% OSCURIDAD
- 25% TIERRA
- 20% LUZ
- 10% AGUA
- 9% VIENTO
- 8% FUEGO
- 0.1% DIVINIDAD

Árbol de decisión

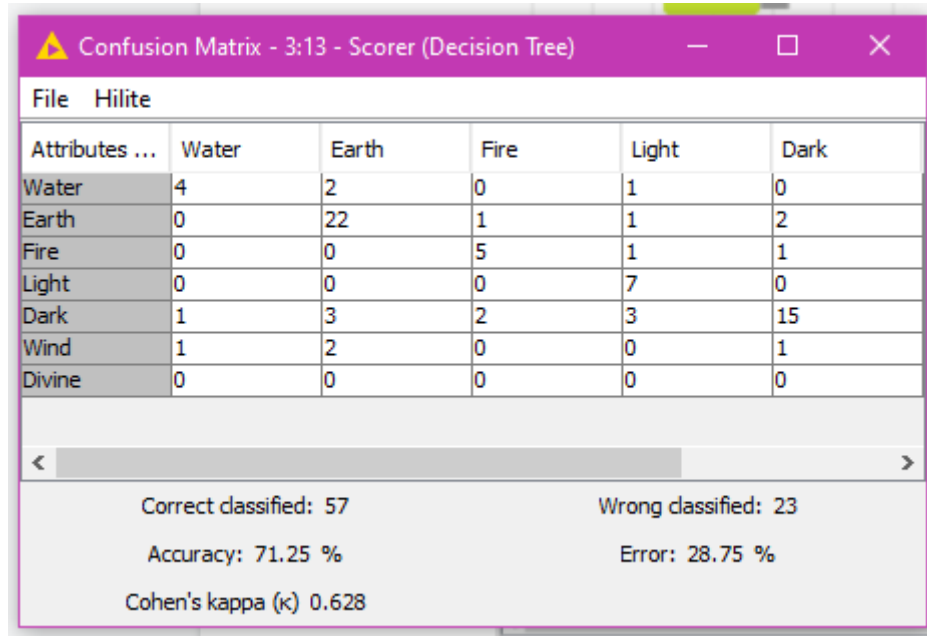
Claramente tenemos un árbol de decisión complejo, donde se parte de un Tipo de carta, se analiza los puntos de defensa o ataque y posteriormente se revisan las demás opciones.



Eficiencia

Tras utilizar el árbol de decisión con el 20% de los datos restantes, podemos calcular una matriz de confusión.

Nuestro modelo predice una respuesta correcta el 71.25% de las veces que lo intenta.



Attributes ...	Water	Earth	Fire	Light	Dark
Water	4	2	0	1	0
Earth	0	22	1	1	2
Fire	0	0	5	1	1
Light	0	0	0	7	0
Dark	1	3	2	3	15
Wind	1	2	0	0	1
Divine	0	0	0	0	0

Correct classified: 57 Wrong classified: 23
Accuracy: 71.25 % Error: 28.75 %
Cohen's kappa (κ) 0.628

Se que es mejor que elegir al azar una clase, ya que tendríamos una probabilidad del 14% de atinarle.

Honestamente no estaba segura de cual es un buen porcentaje de predicción así que busqué si existían análisis previos de predicción. Encontré un trabajo que pretende predecir el ataque y defensa de una carta pero que genera una predicción para todos los atributos de la tarjeta [10] ; revisando el punto 4.1 obtenemos estos resultados con los que podemos validar que el modelo está en un rango aceptable:

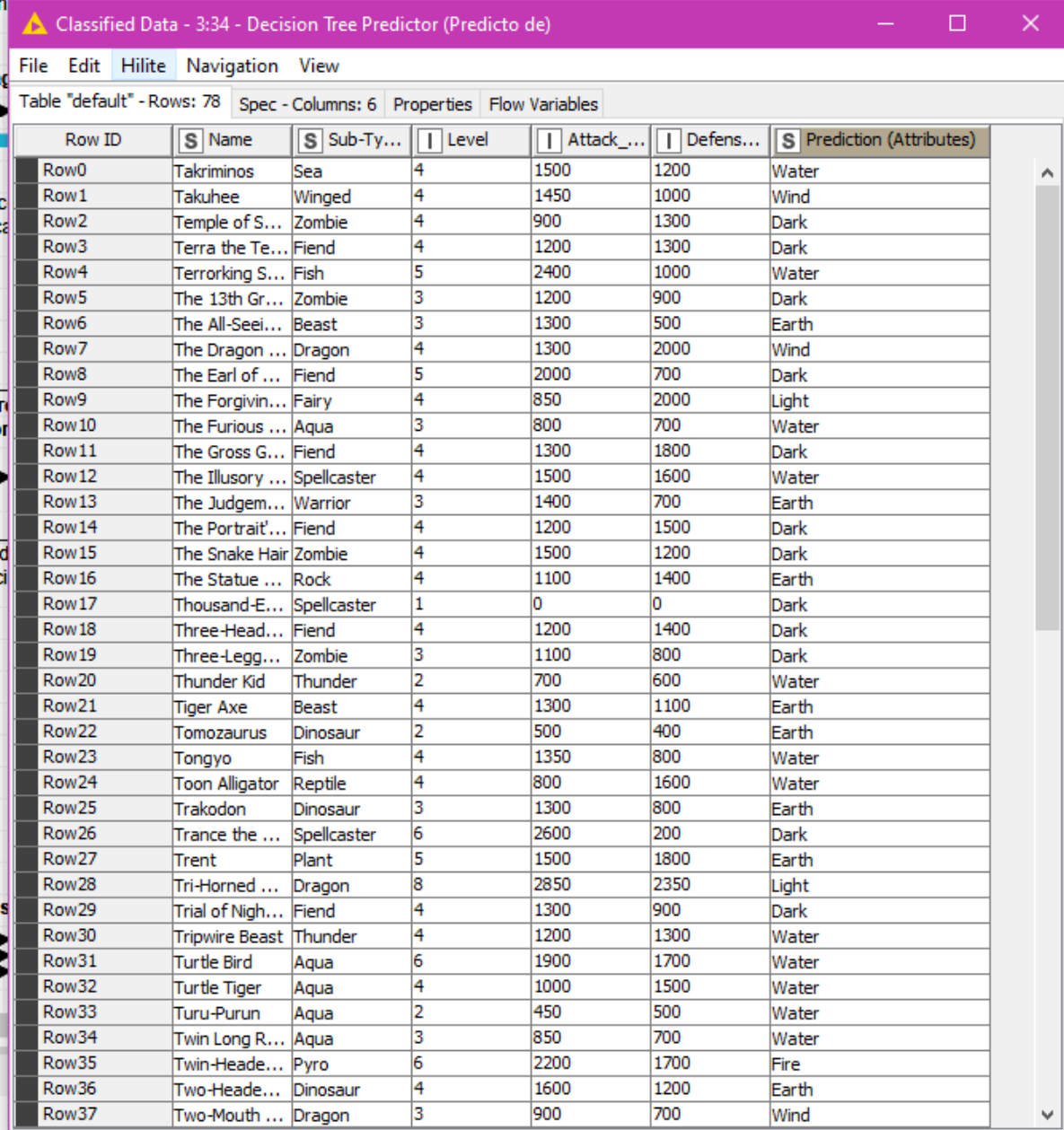
```
accuracy_attr: 0.5515297906602254
accuracy_species: 0.4669887278582931
accuracy_level: 0.3413848631239936
r2score_attack: 0.0804399379391485
r2score_defence: 0.04577024081390113
```

Como comentario adicional, encontré que el porcentaje de resultados correctos de predicciones cambia dependiendo de la distribución de la partición de datos con los que entrenamos el árbol.

Realicé varias pruebas y obtuve un promedio de 60% de exactitud, variando en un rango de 50%-70%.

Predicción

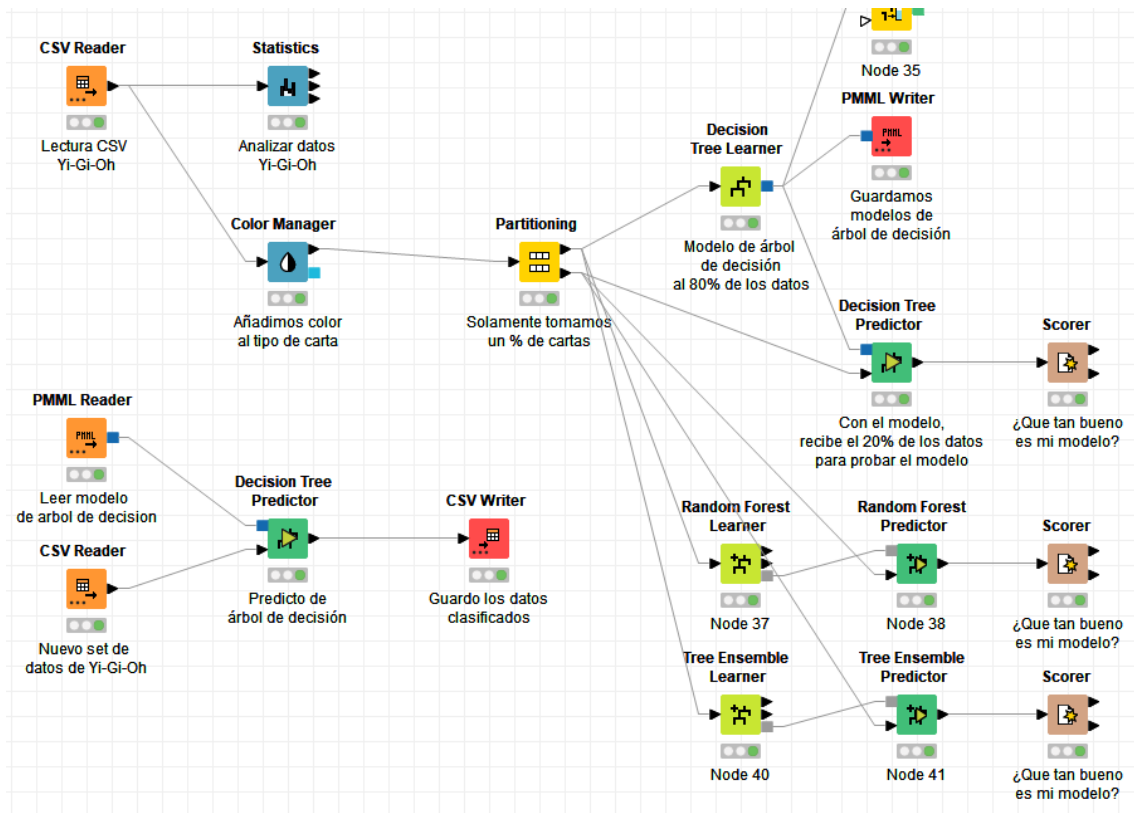
Por último, tome una sección de datos sin atributo (diferente a la usada para entrenar o validar el modelo) y usando el modelo de árbol de decisión, predije el valor del atributo.



Row ID	Name	Sub-Ty...	Level	Attack_...	Defens...	Prediction (Attributes)
Row0	Takriminos	Sea	4	1500	1200	Water
Row1	Takuhee	Winged	4	1450	1000	Wind
Row2	Temple of S...	Zombie	4	900	1300	Dark
Row3	Terra the Te...	Fiend	4	1200	1300	Dark
Row4	Terroring S...	Fish	5	2400	1000	Water
Row5	The 13th Gr...	Zombie	3	1200	900	Dark
Row6	The All-Seei...	Beast	3	1300	500	Earth
Row7	The Dragon ...	Dragon	4	1300	2000	Wind
Row8	The Earl of ...	Fiend	5	2000	700	Dark
Row9	The Forgivin...	Fairy	4	850	2000	Light
Row10	The Furious ...	Aqua	3	800	700	Water
Row11	The Gross G...	Fiend	4	1300	1800	Dark
Row12	The Illusory ...	Spellcaster	4	1500	1600	Water
Row13	The Judgem...	Warrior	3	1400	700	Earth
Row14	The Portrait'...	Fiend	4	1200	1500	Dark
Row15	The Snake Hair	Zombie	4	1500	1200	Dark
Row16	The Statue ...	Rock	4	1100	1400	Earth
Row17	Thousand-E...	Spellcaster	1	0	0	Dark
Row18	Three-Head...	Fiend	4	1200	1400	Dark
Row19	Three-Legg...	Zombie	3	1100	800	Dark
Row20	Thunder Kid	Thunder	2	700	600	Water
Row21	Tiger Axe	Beast	4	1300	1100	Earth
Row22	Tomozaurs	Dinosaur	2	500	400	Earth
Row23	Tongyo	Fish	4	1350	800	Water
Row24	Toon Alligator	Reptile	4	800	1600	Water
Row25	Trakodon	Dinosaur	3	1300	800	Earth
Row26	Trance the ...	Spellcaster	6	2600	200	Dark
Row27	Trent	Plant	5	1500	1800	Earth
Row28	Tri-Horned ...	Dragon	8	2850	2350	Light
Row29	Trial of Nigh...	Fiend	4	1300	900	Dark
Row30	Tripwire Beast	Thunder	4	1200	1300	Water
Row31	Turtle Bird	Aqua	6	1900	1700	Water
Row32	Turtle Tiger	Aqua	4	1000	1500	Water
Row33	Turu-Purun	Aqua	2	450	500	Water
Row34	Twin Long R...	Aqua	3	850	700	Water
Row35	Twin-Heade...	Pyro	6	2200	1700	Fire
Row36	Two-Heade...	Dinosaur	4	1600	1200	Earth
Row37	Two-Mouth ...	Dragon	3	900	700	Wind

Implementación

Leemos el set de datos, lo particionamos y usamos el 80% de los datos para generar un árbol de decisión. Este árbol lo usamos para predecir el resultado tomando el 20% restante de la partición y comparamos los datos predecidos contra los originales para obtener la eficiencia del modelo.

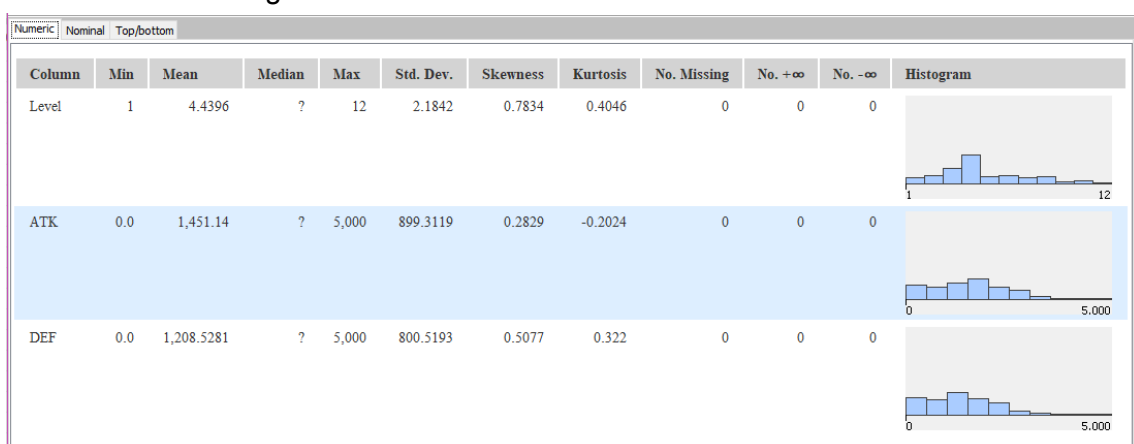


La razón por la que estoy analizando un conjunto de datos de mayor tamaño es para probar la hipótesis de “a mayor cantidad de datos, mayor calidad de análisis”.

Análisis

Puedo ver aún la distribución bayesiana pero ahora tiene una tendencia a priori.

En el juego existen cartas prohibidas[11] que contienen características no balanceadas (nivel y puntaje sumamente elevados), mi conclusión es que son estas cartas las que hacen que mi distribución tenga esta forma.

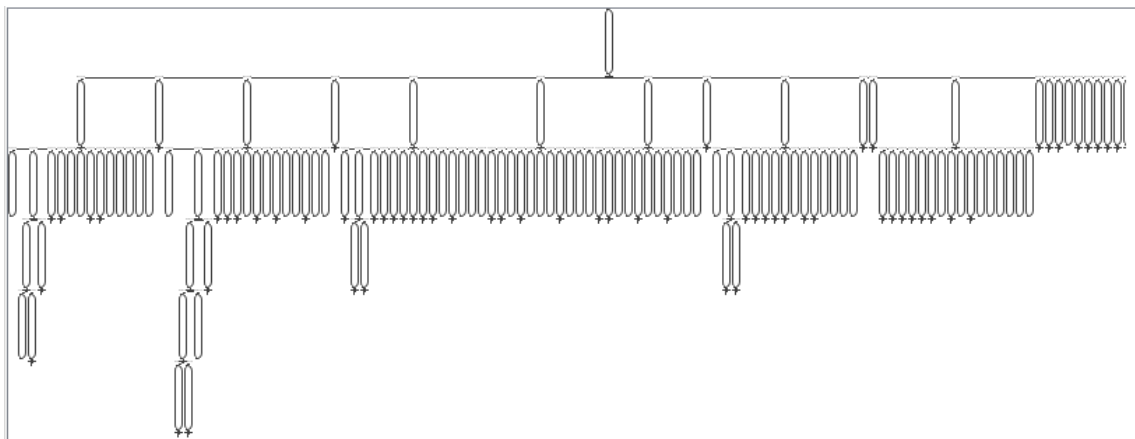


Cuando analizamos los atributos nominales nos concentramos en el de atributos y podemos comprobar que todas nuestras tarjetas caen en alguno de los 7 tipos de categorías.



Árbol de decisión

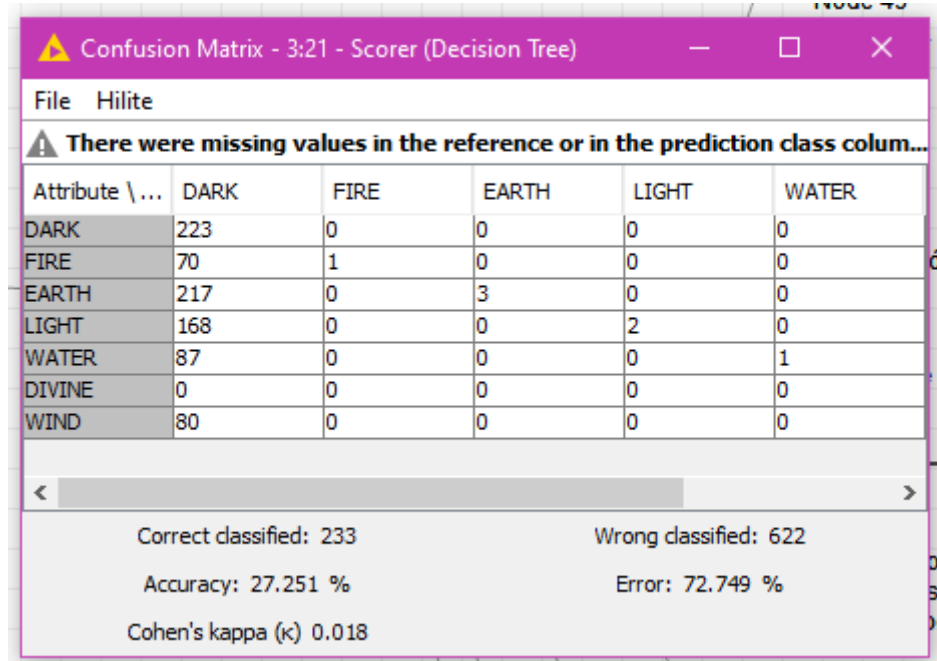
Tenemos un árbol de decisión con mayor complejidad comparado con el set de datos pequeño.



Eficiencia

Tras utilizar el árbol de decisión con el 20% de los datos restantes, podemos calcular una matriz de confusión.

Nuestro modelo predice una respuesta correcta el 27.51% de las veces que lo intenta.



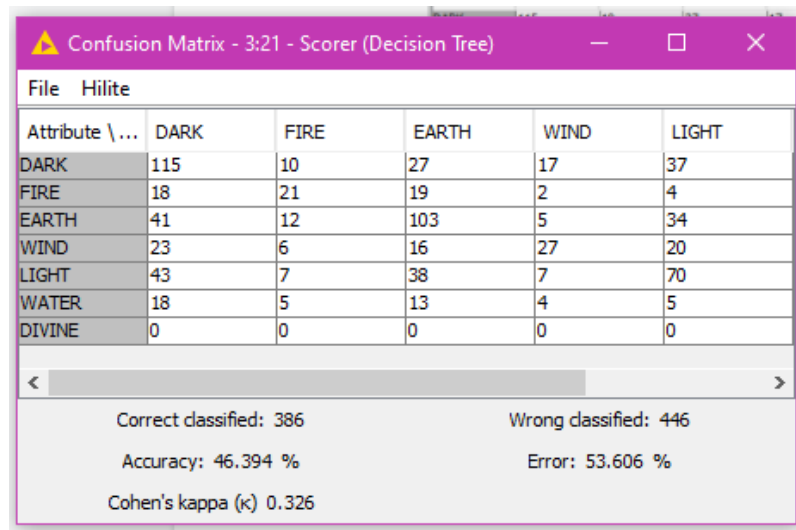
Attribute \ ...	DARK	FIRE	EARTH	LIGHT	WATER
DARK	223	0	0	0	0
FIRE	70	1	0	0	0
EARTH	217	0	3	0	0
LIGHT	168	0	0	2	0
WATER	87	0	0	0	1
DIVINE	0	0	0	0	0
WIND	80	0	0	0	0

Correct classified: 233	Wrong classified: 622
Accuracy: 27.251 %	Error: 72.749 %
Cohen's kappa (κ) 0.018	

Este bajo porcentaje se debe a que los datos de las cartas totales contienen cartas que no tienen información en bastantes columnas:

Row ID	S Name	S Type	I Level	S Race	S Attribute	I ATK	I DEF
Row0	Limit Reverse	Trap Card	?	Continuous	?	?	?
Row1	The 13th Gr...	Normal Mon...	3	Zombie	DARK	1200	900
Row2	Gem-Enhanc...	Trap Card	?	Normal	?	?	?
Row3	Magician's Ci...	Trap Card	?	Normal	?	?	?
Row4	Castle of Da...	Flip Effect M...	4	Fiend	DARK	920	1930
Row5	Cipher Spec...	Trap Card	?	Normal	?	?	?
Row6	Lava Golem	Effect Monster	8	Fiend	FIRE	3000	2500
Row7	Dark Magic ...	Spell Card	?	Quick-Play	?	?	?
Row8	Seismic Cras...	Effect Monster	3	Rock	EARTH	1400	300
Row9	Laval Lancel...	Effect Monster	6	Warrior	FIRE	2100	200
Row10	Skull Dice	Trap Card	?	Normal	?	?	?
Row11	Performapal...	Effect Monster	5	Winged Beast	WIND	1100	2400
Row12	Miracle Flipper	Effect Monster	2	Spellcaster	LIGHT	300	500
Row13	Key Mouse	Tuner Monster	1	Beast	EARTH	100	100
Row14	Mischief of t...	Trap Card	?	Normal	?	?	?
Row15	Vylon Hept	Effect Monster	4	Fairy	LIGHT	1800	800
Row16	Koa'ki Meiru ...	Effect Monster	4	Rock	EARTH	1800	1800
Row17	Koa'ki Meiru ...	Token	?	Rock	?	?	?
Row18	Hero Flash!!	Spell Card	?	Normal	?	?	?
Row19	E - Emergen...	Spell Card	?	Normal	?	?	?
Row20	Fenrir	Effect Monster	4	Beast	WATER	1400	1200
Row21	Ballista of P...	Spell Card	?	Equip	?	?	?

Tras limpiar el set de datos y solamente dejar las tarjetas que contienen información en todas sus columnas, el modelo predice una respuesta correcta el 46.394% de las veces que lo intenta.



Attribute \ ...	DARK	FIRE	EARTH	WIND	LIGHT
DARK	115	10	27	17	37
FIRE	18	21	19	2	4
EARTH	41	12	103	5	34
WIND	23	6	16	27	20
LIGHT	43	7	38	7	70
WATER	18	5	13	4	5
DIVINE	0	0	0	0	0

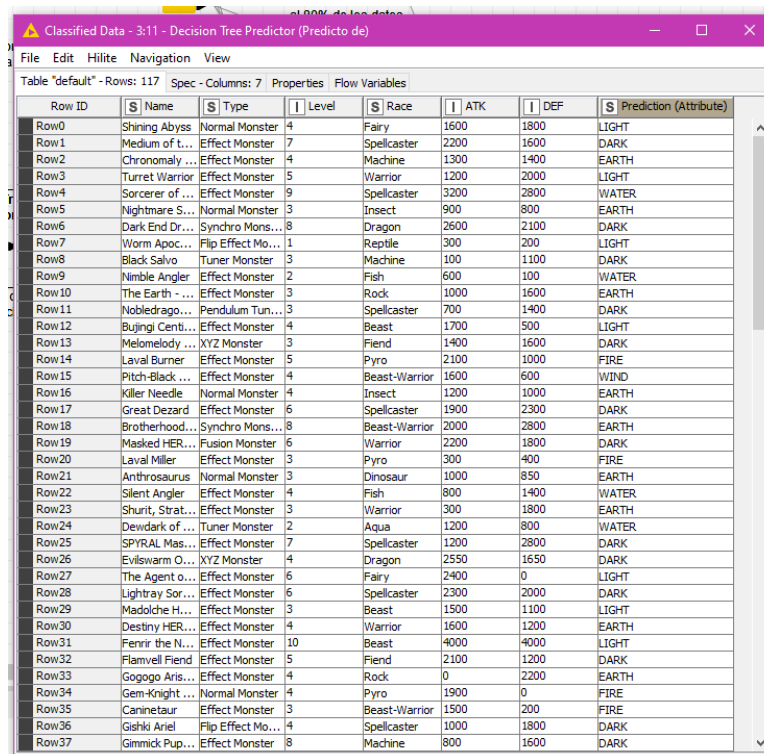
Correct classified: 386 Wrong classified: 446
 Accuracy: 46.394 % Error: 53.606 %
 Cohen's kappa (κ) 0.326

Recordando, el modelo de datos pequeño tuvo un porcentaje de exactitud promediada a 60%. A pesar de tener un set de datos con mayor número de registros y contar con columnas adicionales, nuestro modelo actual es peor.

Con esto puedo concluir que el tener más datos no lleva a tener un mayor rango de predicciones correctas. Es mejor tener calidad que cantidad en los datos, para eso existen técnicas de limpieza, integración y reducción.

Predicción

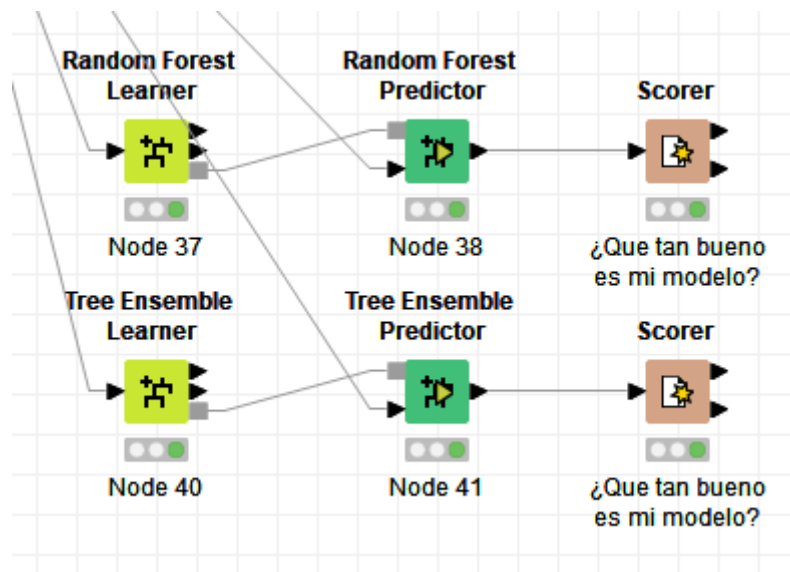
Por último, tome una sección de datos sin atributo (diferente a la usada para entrenar o validar el modelo) y usando el modelo de árbol de decisión, predije el valor del atributo.



Row ID	S Name	S Type	I Level	S Race	I ATK	I DEF	S Prediction (Attribute)
Row0	Shining Abyss	Normal Monster	4	Fairy	1600	1800	LIGHT
Row1	Medium of t...	Effect Monster	7	Spellcaster	2200	1600	DARK
Row2	Chronomaly ...	Effect Monster	4	Machine	1300	1400	EARTH
Row3	Turret Warrior	Effect Monster	5	Warrior	1200	2000	LIGHT
Row4	Sorcerer of ...	Effect Monster	9	Spellcaster	3200	2800	WATER
Row5	Nightmare S...	Normal Monster	3	Insect	900	800	EARTH
Row6	Dark End Dr...	Synchro Mons...	8	Dragon	2600	2100	DARK
Row7	Worm Apoc...	Flip Effect Mo...	1	Reptile	300	200	LIGHT
Row8	Black Salvo	Tuner Monster	3	Machine	100	1100	DARK
Row9	Nimble Angler	Effect Monster	2	Fish	600	100	WATER
Row10	The Earth - ...	Effect Monster	3	Rock	1000	1600	EARTH
Row11	Nobledrago...	Pendulum Tun...	3	Spellcaster	700	1400	DARK
Row12	Bujingi Centi...	Effect Monster	4	Beast	1700	500	LIGHT
Row13	Melomelody ...	XYZ Monster	3	Fiend	1400	1600	DARK
Row14	Laval Burner	Effect Monster	5	Pyro	2100	1000	FIRE
Row15	Pitch-Black ...	Effect Monster	4	Beast-Warrior	1600	600	WIND
Row16	Killer Needle	Normal Monster	4	Insect	1200	1000	EARTH
Row17	Great Dezaud	Effect Monster	6	Spellcaster	1900	2300	DARK
Row18	Brotherhood...	Synchro Mons...	8	Beast-Warrior	2000	2800	EARTH
Row19	Masked HER...	Fusion Monster	6	Warrior	2200	1800	DARK
Row20	Laval Miller	Effect Monster	3	Pyro	300	400	FIRE
Row21	Anthrosaurus	Normal Monster	3	Dinosaur	1000	850	EARTH
Row22	Silent Angler	Effect Monster	4	Fish	800	1400	WATER
Row23	Shurt, Strat...	Effect Monster	3	Warrior	300	1800	EARTH
Row24	Devdark of ...	Tuner Monster	2	Aqua	1200	800	WATER
Row25	SPYRAL Mas...	Effect Monster	7	Spellcaster	1200	2800	DARK
Row26	Evilswarm O...	XYZ Monster	4	Dragon	2550	1650	DARK
Row27	The Agent o...	Effect Monster	6	Fairy	2400	0	LIGHT
Row28	Lightray Sor...	Effect Monster	6	Spellcaster	2300	2000	DARK
Row29	Madolche H...	Effect Monster	3	Beast	1500	1100	LIGHT
Row30	Destiny HER...	Effect Monster	4	Warrior	1600	1200	EARTH
Row31	Fenrir the N...	Effect Monster	10	Beast	4000	4000	LIGHT
Row32	Flamvell Fiend	Effect Monster	5	Fiend	2100	1200	DARK
Row33	Gogogo Aris...	Effect Monster	4	Rock	0	2200	EARTH
Row34	Gem-Knight ...	Normal Monster	4	Pyro	1900	0	FIRE
Row35	Caninetaur	Effect Monster	3	Beast-Warrior	1500	200	FIRE
Row36	Gishki Ariel	Flip Effect Mo...	4	Spellcaster	1000	1800	DARK
Row37	Gimmick Pup...	Effect Monster	8	Machine	800	1600	DARK

Mejora en el modelo de predicción

En el material de apoyo encontré que existen varias técnicas para mejorar un modelo. Entre las que me llamaron la atención están los métodos de ensamblaje los cuales crean un modelo de modelos.



Tras realizar un proceso similar con ambas bases, pude mejorar la base de datos pequeña a un 72.75% de exactitud y la base de datos grande a un 51.42%

Confusion Matrix - 3:13 - Scorer (Decision Tree) File: Hilito Attributes ... Water Earth Fire Light Dark Water 4 2 0 1 0 Earth 0 22 1 1 2 Fire 0 0 5 1 1 Light 0 0 0 7 0 Dark 1 3 2 3 15 Wind 1 2 0 0 1 Divine 0 0 0 0 0 Correct classified: 57 Wrong classified: 23 Accuracy: 71.25 % Error: 28.75 % Cohen's kappa (κ) 0.628	Confusion Matrix - 3:39 - Scorer (Random Forest) File: Hilito Attributes ... Water Earth Fire Light Dark Water 3 3 0 1 0 Earth 1 23 0 1 1 Fire 0 1 4 1 1 Light 0 0 0 7 0 Dark 1 4 0 1 18 Wind 1 1 0 0 2 Divine 0 0 0 0 0 Correct classified: 59 Wrong classified: 21 Accuracy: 73.75 % Error: 26.25 % Cohen's kappa (κ) 0.653	Confusion Matrix - 3:42 - Scorer (Tree Ensemble) File: Hilito Attributes ... Water Earth Fire Light Dark Water 3 3 0 1 0 Earth 1 22 0 1 1 Fire 0 1 4 1 1 Light 0 1 0 5 2 Dark 2 5 0 0 16 Wind 1 1 0 0 2 Divine 0 0 0 0 0 Correct classified: 54 Wrong classified: 26 Accuracy: 67.5 % Error: 32.5 % Cohen's kappa (κ) 0.569
Confusion Matrix - 3:21 - Scorer (Decision Tree) File: Hilito Attribute \ ... DARK FIRE EARTH WIND LIGHT DARK 115 10 27 17 37 FIRE 18 21 19 2 4 EARTH 41 12 103 5 34 WIND 23 6 16 27 20 LIGHT 43 7 38 7 70 WATER 18 5 13 4 5 DIVINE 0 0 0 0 0 Correct classified: 386 Wrong classified: 446 Accuracy: 46.394 % Error: 53.606 % Cohen's kappa (κ) 0.326	Confusion Matrix - 3:47 - Scorer (Random Forest) File: Hilito Attribute \ ... DARK FIRE EARTH WIND LIGHT DARK 127 7 38 12 25 FIRE 16 22 20 3 3 EARTH 30 9 130 5 24 WIND 22 5 24 26 15 LIGHT 37 6 47 3 73 WATER 13 4 16 3 9 DIVINE 0 0 0 0 0 Correct classified: 428 Wrong classified: 404 Accuracy: 51.442 % Error: 48.558 % Cohen's kappa (κ) 0.386	Confusion Matrix - 3:43 - Scorer (Tree Ensemble) File: Hilito Attribute \ ... DARK FIRE EARTH WIND LIGHT DARK 128 7 36 14 24 FIRE 17 22 21 1 3 EARTH 32 11 126 6 24 WIND 23 5 24 25 15 LIGHT 43 6 42 4 71 WATER 12 3 21 3 6 DIVINE 0 0 0 0 0 Correct classified: 422 Wrong classified: 410 Accuracy: 50.721 % Error: 49.279 % Cohen's kappa (κ) 0.376

Puedo ver una mejora en sus porcentajes pero de nuevo encontré que los resultados cambian dependiendo de los datos que la partición elija para entrenar los árboles.

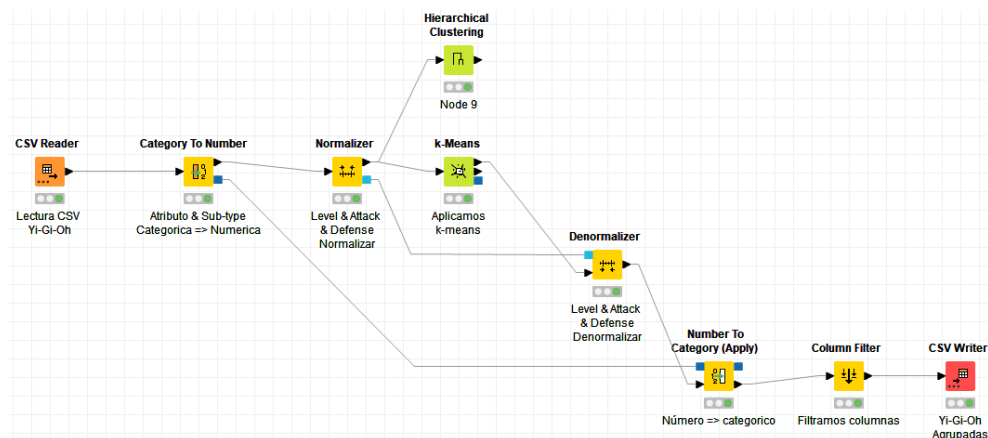
Agrupamiento

Para el agrupamiento usaremos la técnica de K-medias el cual tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.

Set yu-gi-oh_small

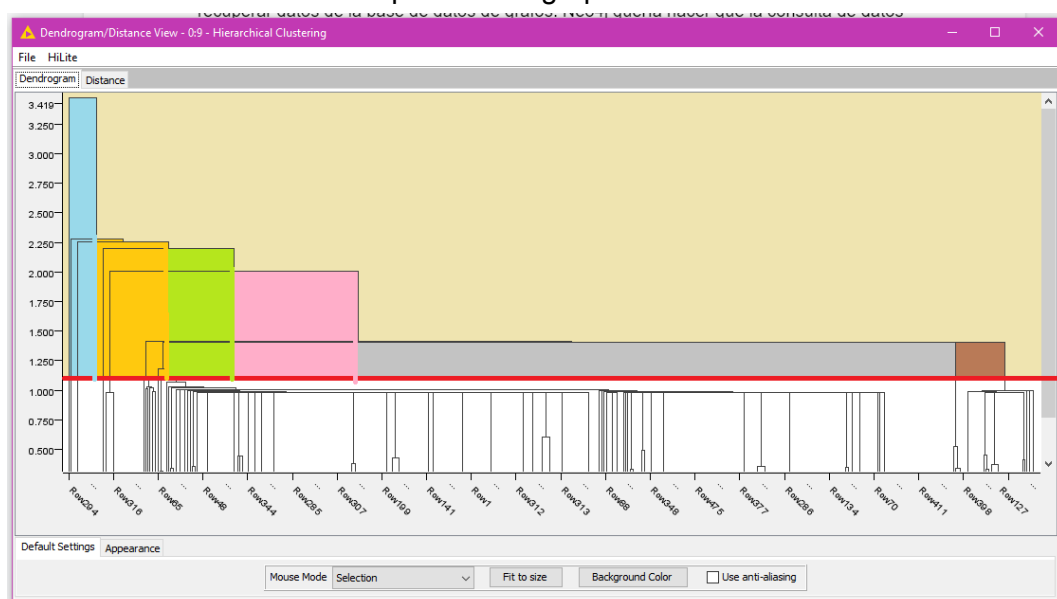
Implementación

Leemos el set de datos, pasamos los datos categóricos (cualquier dato tipo texto) a tipo numérico y normalizamos los datos numéricos (a un rango de 0-1 y solamente las columnas originales). Aplicamos el algoritmo de agrupación K-means. Lo restante es regresar los datos a su formato original denormalizando los valores numéricos y recuperando los valores categóricos.



Análisis de clusters

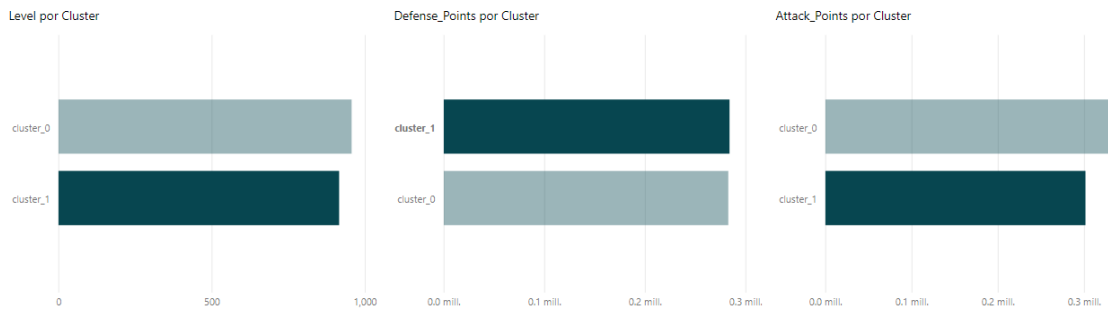
El parámetro principal de nuestro algoritmo de agrupación es el número de clusters. Para decidir cuántos clusters realizar, creamos un agrupamiento jerárquico el cual es un método de análisis de grupos puntuales, el cual busca construir una jerarquía de grupos y nos regresa la distribución de datos respecto a los grupos.



Identificamos 7 grupos. Para tener un análisis completo, realizaremos el mismo proceso con 2 clusters y con la media entre el mínimo y máximo, 4 clusters.

Agrupamiento del modelo con 2 clusters

Con el agrupamiento con 2 clusters podemos obtener cuales son las tarjetas que mejor funcionan para defender y cuales funcionan para atacar.



Si quiero defenderme usando la mejor tarjeta disponible, usaría una con atributo fuego de subtipo dragon o máquina.

Cluster	Attributes	Recuento de Name	Attack_Points	Defense_Points	Cluster	Attributes	Sub-Types	Recuento de Name	Mediana de Level	Promedio de Attack_Points	Promedio de Defense_Points
cluster_1	Earth	98	124,250.00	126,350.00	cluster_1	Fire	Dragon	3	6.00	1,833.33	2,100.00
cluster_1	Dark	30	49,500.00	42,850.00	cluster_1	Fire	Machine	6	4.50	1,383.33	2,050.00
cluster_1	Light	33	44,400.00	42,800.00	cluster_1	Water	Machine	1	5.00	1,600.00	1,900.00
cluster_1	Water	21	32,250.00	25,400.00	cluster_1	Dark	Dragon	8	6.00	2,062.50	1,587.50
cluster_1	Wind	23	31,500.00	24,200.00	cluster_1	Light	Dragon	8	7.00	2,068.75	1,575.00
cluster_1	Fire	13	19,750.00	22,600.00	cluster_1	Earth	Rock	18	4.00	1,247.22	1,525.00
Total		218	301,650.00	284,200.00	cluster_1	Water	Reptile	5	4.00	1,680.00	1,440.00
					cluster_1	Light	Machine	4	4.00	1,562.50	1,425.00
					cluster_1	Dark	Machine	17	5.00	1,444.12	1,411.76
					cluster_1	Dark	Fairy	1	5.00	1,600.00	1,400.00
					cluster_1	Wind	Machine	2	4.50	1,950.00	1,400.00
					cluster_1	Dark	Beast	3	4.00	1,633.33	1,366.67
					cluster_1	Earth	Dragon	3	5.00	1,433.33	1,366.67
					cluster_1	Earth	Machine	14	4.00	1,128.57	1,332.14
					cluster_1	Fire	Beast	3	5.00	1,950.00	1,300.00
					Total			218	4.00	1,383.72	1,303.67

Si quiero atacar usando la mejor tarjeta permitida y disponible, usaría una con atributo agua y subtipo Wyrn o una tarjeta con atributo obscuro y subtipo Cyberse.

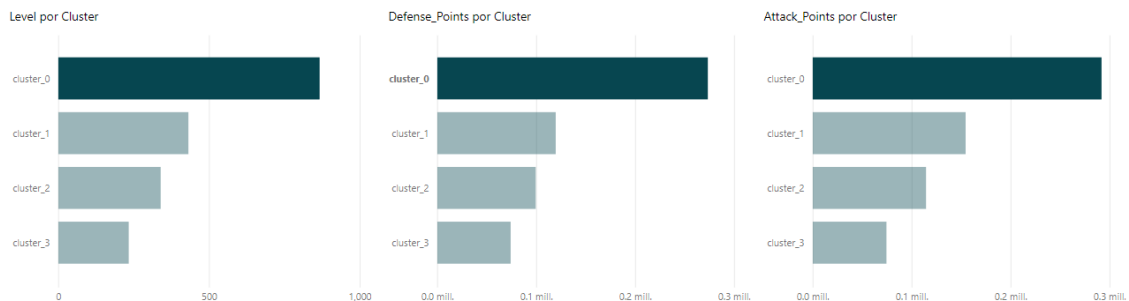
Cluster	Attributes	Recuento de Name	Attack_Points	Defense_Points	Cluster	Attributes	Sub-Types	Recuento de Name	Mediana de Level	Promedio de Attack_Points	Promedio de Defense_Points
cluster_0	Dark	79	100,550.00	82,330.00	cluster_0	Divine	Divine	1	10.00	4,000.00	4,000.00
cluster_0	Earth	73	89,600.00	78,580.00	cluster_0	Water	Wyrn	1	8.00	2,900.00	2,900.00
cluster_0	Water	45	58,550.00	56,650.00	cluster_0	Dark	Cyberse	1	8.00	2,800.00	2,600.00
cluster_0	Light	20	30,950.00	22,200.00	cluster_0	Light	Wyrn	1	7.00	2,800.00	1,000.00
cluster_0	Wind	20	25,350.00	19,200.00	cluster_0	Earth	Spellcaster	2	5.00	2,250.00	550.00
cluster_0	Fire	20	25,200.00	20,000.00	cluster_0	Light	Psychic	2	5.00	2,250.00	0.00
cluster_0	Divine	1	4,000.00	4,000.00	cluster_0	Water	Dinosaur	3	4.00	2,100.00	1,066.67
Total		258	334,200.00	282,960.00	cluster_0	Dark	Wyrn	1	4.00	2,000.00	0.00
					cluster_0	Water	Sea	6	4.50	1,833.33	1,400.00
					cluster_0	Light	Warrior	6	4.00	1,733.33	1,466.67
					cluster_0	Fire	Aqua	1	4.00	1,600.00	1,500.00
					cluster_0	Fire	Psychic	1	3.00	1,600.00	0.00
					cluster_0	Water	Warrior	1	4.00	1,600.00	900.00
					cluster_0	Earth	Zombie	3	4.00	1,550.00	333.33
					cluster_0	Earth	Thunder	2	4.00	1,400.00	1,550.00
					Total			258	4.00	1,295.35	1,096.74

Si quisiera crear un mazo que maximice mi estrategia, usaría una tarjeta con atributo tierra ya que es la que tiene mayor probabilidad de maximizar los puntos de defensa sin descuidar los puntos de ataques.

Agrupamiento del modelo con 4 clusters

Con el agrupamiento con 4 clusters podemos obtener cuales son las tarjetas recomendadas para el mazo principal (cluster 3 y cluster 2) y las tarjetas recomendadas para el mazo secundario.

Adicionalmente tenemos las tarjetas que no se recomienda usar en absoluto (cluster 0).



Parece contraintuitivo llegar a esta conclusión solamente viendo la distribución de cartas y la suma de los puntos pero si analizamos los clusters, ya no tienen una distribución de cartas uniforme por lo que tenemos que comparar mediana de puntos y no suma de puntos.

Cluster	Recuento de Name	Mediana de Attack_Points	Mediana de Defense_Points
cluster_1	119	1,300.00	1,000.00
cluster_0	226	1,300.00	1,200.00
cluster_2	76	1,500.00	1,250.00
cluster_3	55	1,400.00	1,400.00
Total	476	1,400.00	1,200.00

Si busco las mejores tarjetas para una estrategia enfocada en ataque sin descuidar la defensa, usaría Blue-Eyes White Dragon, Rabidragon, Gogiga Gagagigo, Tri-Horned Dragon y Wingweaver.

Cluster	Attributes	Sub-Types	Name	Mediana de Level	Promedio de Attack_Points	Promedio de Defense_Points
cluster_2	Light	Dragon	Blue-Eyes White Dragon	8.00	3,000.00	2,500.00
cluster_2	Light	Dragon	Rabidragon	8.00	2,950.00	2,900.00
cluster_3	Water	Reptile	Gogiga Gagagigo	8.00	2,950.00	2,800.00
cluster_2	Dark	Dragon	Tri-Horned Dragon	8.00	2,850.00	2,350.00
cluster_2	Light	Fairy	Wingweaver	7.00	2,750.00	2,400.00
cluster_2	Wind	Machine	Cyber-Tech Alligator	5.00	2,500.00	1,600.00
cluster_2	Light	Dragon	Seiyaryu	7.00	2,500.00	2,300.00
cluster_2	Light	Dragon	Wattaildragon	6.00	2,500.00	1,000.00
cluster_3	Water	Reptile	Giga Gagagigo	5.00	2,450.00	1,500.00
cluster_2	Wind	Dragon	Luster Dragon #2	6.00	2,400.00	1,400.00
cluster_2	Dark	Dragon	Red-Eyes B. Dragon	7.00	2,400.00	2,000.00
cluster_3	Water	Fish	Amphibian Beast	6.00	2,400.00	2,000.00
cluster_3	Water	Fish	Terroring Salmon	5.00	2,400.00	1,000.00
cluster_2	Dark	Dragon	Serpent Night Dragon	7.00	2,350.00	2,400.00
Total				4.00	1,445.04	1,327.10

Si busco las mejores tarjetas para una estrategia enfocada en defensa sin descuidar el ataque, usaría Radidragon, Hyozanryu, Gogiga Gagagigo, Ryu-Ran y Blue-Eyes White Dragon.

Cluster	Attributes	Sub-Types	Name	Mediana de Level	Promedio de Attack_Points	Promedio de Defense_Points
cluster_2	Light	Dragon	Rabidragon	8.00	2,950.00	2,900.00
cluster_2	Light	Dragon	Hyozanryu	7.00	2,100.00	2,800.00
cluster_3	Water	Reptile	Gogiga Gagagigo	8.00	2,950.00	2,800.00
cluster_3	Fire	Dragon	Ryu-Ran	7.00	2,200.00	2,600.00
cluster_2	Light	Dragon	Blue-Eyes White Dragon	8.00	3,000.00	2,500.00
cluster_3	Fire	Machine	Launcher Spider	7.00	2,200.00	2,500.00
cluster_3	Fire	Machine	Woodborg Inpachi	5.00	500.00	2,500.00
cluster_2	Dark	Dragon	Serpent Night Dragon	7.00	2,350.00	2,400.00
cluster_2	Light	Fairy	Wingweaver	7.00	2,750.00	2,400.00
cluster_2	Dark	Dragon	Mikazukinoyaiba	7.00	2,200.00	2,350.00
cluster_2	Dark	Dragon	Tri-Horned Dragon	8.00	2,850.00	2,350.00
cluster_2	Light	Dragon	Seiyaryu	7.00	2,500.00	2,300.00
cluster_2	Dark	Machine	Slot Machine	7.00	2,000.00	2,300.00
cluster_3	Earth	Machine	Steel Ogre Grotto #2	6.00	1,900.00	2,200.00
Total				4.00	1,445.04	1,327.10

Podemos formar un mazo maximizado juntando ambos análisis, eliminando repetidos y llegando a 40 tarjetas en total.

Repetimos este análisis en el cluster 0 y es evidente que las cartas no están balanceadas ya que tienen un valor “alto” en defensa pero 0 o extremadamente bajo en ataque o viceversa.

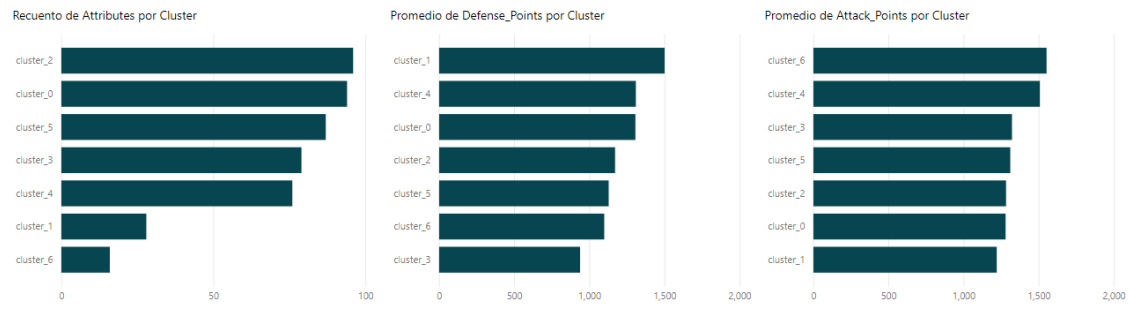
Cluster	Attributes	Sub-Types	Name	Mediana de Level	Promedio de Attack_Points	Promedio de Defense_Points
cluster_0	Earth	Warrior	Beckoned by the World Chalice	4.00	1,800.00	0.00
cluster_0	Earth	Beast	Chu-Ske the Mouse Fighter	3.00	1,200.00	0.00
cluster_0	Water	Dinosaur	Megalosmasher X	4.00	2,000.00	0.00
cluster_0	Earth	Insect	Shiny Black C" Squadder"	4.00	2,000.00	0.00
cluster_0	Dark	Spellcaster	Thousand-Eyes Idol	1.00	0.00	0.00
cluster_0	Earth	Rock	Gem-Knight Lapis	3.00	1,200.00	100.00
cluster_0	Earth	Beast	Gene-Warped Warwolf	4.00	2,000.00	100.00
cluster_0	Earth	Insect	Fiend Scorpion	2.00	900.00	200.00
cluster_0	Earth	Insect	Petit Moth	1.00	300.00	200.00
cluster_0	Dark	Fiend	Souls of the Forgotten	2.00	900.00	200.00
cluster_0	Earth	Spellcaster	Trance the Magic Swordsman	6.00	2,600.00	200.00
cluster_0	Dark	Spellcaster	Left Arm of the Forbidden One	1.00	200.00	300.00
cluster_0	Dark	Spellcaster	Left Leg of the Forbidden One	1.00	200.00	300.00
cluster_0	Dark	Spellcaster	Right Arm of the Forbidden One	1.00	200.00	300.00
Total				4.00	1,291.59	1,210.22

Cluster	Attributes	Sub-Types	Name	Mediana de Level	Promedio de Attack_Points	Promedio de Defense_Points
cluster_0	Water	Spellcaster	Crowned by the World Chalice	2.00	0.00	2,100.00
cluster_0	Earth	Rock	Labyrinth Wall	5.00	0.00	3,000.00
cluster_0	Earth	Warrior	Millennium Shield	5.00	0.00	3,000.00
cluster_0	Light	Beast	Ojama Black	2.00	0.00	1,000.00
cluster_0	Light	Beast	Ojama Green	2.00	0.00	1,000.00
cluster_0	Light	Beast	Ojama Yellow	2.00	0.00	1,000.00
cluster_0	Earth	Beast	Soul Tiger	4.00	0.00	2,100.00
cluster_0	Dark	Spellcaster	Thousand-Eyes Idol	1.00	0.00	0.00
cluster_0	Light	Fiend	White Duston	1.00	0.00	1,000.00
cluster_0	Dark	Fiend	D.D. Trainer	1.00	100.00	2,000.00
cluster_0	Dark	Fiend	Renge, Gatekeeper of Dark World	4.00	100.00	2,100.00
cluster_0	Earth	Beast	Bunilla	1.00	150.00	2,050.00
cluster_0	Earth	Warrior	Chamberlain of the Six Samurai	3.00	200.00	2,000.00
cluster_0	Dark	Spellcaster	Left Arm of the Forbidden One	1.00	200.00	300.00
Total				4.00	1,291.59	1,210.22

Este cluster nos puede dar la información de cuales cartas son la mejor opción para sacrificar en una partida.

Agrupamiento del modelo con 7 clusters

Mientras más grupos hay, más trabajo cuesta identificar para qué me sirve la información. Después de moverle, analizar y tratar de encontrar uso a mis datos, llegué a la conclusión que generar 7 clusters me puede dar cuales son las mejores cartas y peores cartas dentro de cada atributo.



Puede ser útil conocer esta información ya que las tarjetas tipo hechizo y trampa se deben de utilizar en conjunto con una tarjeta de atributo específico para que tengan efecto.

Teóricamente, podríamos recomendar cual es la mejor tarjeta a usar dependiendo del atributo.

Por ejemplo, si tenemos una tarjeta de magia tipo “Espada de la Destrucción Oscura”[12]. Esta tarjeta solamente puede usarse sobre una carta de monstruo de oscuridad. Éste gana 400 ATK y pierde 200 DEF.

Para que sea conveniente entonces analizamos cuales clusters tienen tarjetas con atributo oscuro.

Cluster	Recuento de Name	Mediana de Attack_Points	Mediana de Defense_Points
cluster_4	26	1,650.00	1,500.00
cluster_5	66	1,300.00	1,150.00
cluster_3	17	1,200.00	700.00
Total	109	1,400.00	1,200.00

El cluster 3 no nos sirve ya que no queremos perder más puntos de defensa sobre una carta débil.

Analizamos los valores en los cluster restantes y obtenemos las cartas que nos conviene utilizar.

Cluster	Attributes	Sub-Types	Name	Mediana de Level	Attack_Points	Defense_Points
cluster_5	Dark	Spellcaster	Cosmo Queen	8.00	2,900.00	2,450.00
cluster_4	Dark	Dragon	Serpent Night Dragon	7.00	2,350.00	2,400.00
cluster_4	Dark	Dragon	Mikazukinoyaiba	7.00	2,200.00	2,350.00
cluster_4	Dark	Dragon	Tri-Horned Dragon	8.00	2,850.00	2,350.00
cluster_4	Dark	Machine	Slot Machine	7.00	2,000.00	2,300.00
cluster_5	Dark	Spellcaster	Illusionist Faceless Mage	5.00	1,200.00	2,200.00
cluster_5	Dark	Fiend	Beast of Talwar	6.00	2,400.00	2,150.00
cluster_5	Dark	Fiend	Metal Guardian	5.00	1,150.00	2,150.00
cluster_5	Dark	Spellcaster	Dark Magician	7.00	2,500.00	2,100.00
cluster_4	Dark	Fiend	Renge, Gatekeeper of Dark World	4.00	100.00	2,100.00
cluster_4	Dark	Machine	Pendulum Machine	6.00	1,750.00	2,000.00
cluster_4	Dark	Dragon	Red-Eyes B. Dragon	7.00	2,400.00	2,000.00
cluster_5	Dark	Fiend	D.D. Trainer	1.00	100.00	2,000.00
cluster_5	Dark	Fiend	Ushi Oni	6.00	2,150.00	1,950.00
cluster_5	Dark	Fiend	Zoa	7.00	2,600.00	1,900.00
Total				4.00		

Comparativa entre los árboles de decisión y el agrupamiento

Encontré que cada método te permite predecir/recomendar información diferente.

Me parece que los árboles de decisiones son útiles cuando tenemos información de entrada que nos puede llevar a concluir un dato. Le encuentro utilidad sobre todo en modelos de reconocimiento, es decir, que le pongan ciertos datos de una tarjeta y te diga cual es dentro de la colección, pero no para un modelo de predicción de movimientos, al menos no con el set de datos que elegí.

Al contrario, en el modelo de agrupamiento encontré que era más sencillo identificar como ciertos grupos de cartas son buenos para algunos movimientos específicos pero siento que no es suficiente, necesitamos tomar los cluster y usarlos como información preprocesada para otro algoritmo.

En resumen, me di cuenta que el árbol de decisión está más enfocado a los datos categóricos, mientras que el de agrupamiento por k-means funciona mejor con datos únicamente numéricos.

Conclusiones y Aprendizajes

Me gustó la práctica, requirió que entendiera un tema externo y realmente aplicara un análisis a los datos que estaba obteniendo.

Se siente refrescante trabajar con un set de datos diferente y ver que podemos tomar información de casi cualquier tema para analizar. Al inicio estaba buscando una base de datos con estadísticas de crímenes porque tengo bastante interés en el tema del ViCAP del FBI y como la interacción e interconexión de varias bases de datos ha llevado a la captura de criminales a un paso mucho más rápido. El problema es que las bases de datos públicas de ese tema contenían poca información o información que requería un proceso de limpieza exhaustivo y las bases de datos proveídas para investigación (Radford/FGCU Database) tienen un proceso de pre aprobación de aproximadamente 6 meses y está sujeta a constante monitoreo.

Me hubiera gustado ver las conclusiones a las que mis demás compañeros llegaron ya que los temas que alcancé a ver en el registro se veían interesantes.

Bibliografía

- [1] "Self-Paced Courses List." <https://www.knime.com/knime-self-paced-courses> (accessed Nov. 26, 2021).
- [2] "1.10. Decision Trees." <https://scikit-learn/stable/modules/tree.html> (accessed Nov. 26, 2021).
- [3] "kmeans." https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html (accessed Nov. 26, 2021).
- [4] "Yu-Gi-Oh! API Guide - Yu-Gi-Oh! Card Database - YGOPRODECK." <https://db.ygoprodeck.com/api-guide/> (accessed Nov. 25, 2021).
- [5] A. Lowhur, "I made an AI to recognize over 10,000 Yugioh cards," *Towards Data Science*, Dec. 02, 2020. <https://towardsdatascience.com/i-made-an-ai-to-recognize-over-10-000-yugioh-cards-26fc6aed1588> (accessed Nov. 25, 2021).
- [6] "Yu-Gi-Oh!" <https://en.wikipedia.org/wiki/Yu-Gi-Oh!> (accessed Nov. 25, 2021).
- [7] "Yu-Gi-Oh! Trading Card Game." https://en.wikipedia.org/wiki/Yu-Gi-Oh!_Trading_Card_Game (accessed Nov. 25, 2021).
- [8] "Atributo." <https://yugioh.fandom.com/es/wiki/Atributo> (accessed Nov. 25, 2021).
- [9] "Atributo." <https://yugioh.fandom.com/es/wiki/Atributo> (accessed Nov. 26, 2021).

- [10] "Predicting offensive and defensive attributes from the Yu-Gi-Oh! Card name --Yu-Gi-Oh! Data Science 3. Machine Learning," Nov. 16, 2020.
<https://linuxtut.com/en/51ff9585a79bac4fd8a2/> (accessed Nov. 26, 2021).
- [11] "Forbidden & Limited Card List." <https://www.yugioh-card.com/es/limited/> (accessed Nov. 26, 2021).
- [12] "Espada de la Destrucción Oscura."
https://yugioh.fandom.com/es/wiki/Espada_de_la_Destrucci%C3%B3n_Oscura
(accessed Nov. 26, 2021).