Data Challenge

Monster Hunt en el ITAM

En Octubre de 2023 el ITAM fue embrujado y rectoría necesita de tu ayuda. Los expertos del Departamento de Estadística estiman que existen alrededor de 1000 criaturas rondando por los pasillos del ITAM aterrorizando a l@s alumn@s. El Departamento pidió ayuda al equipo de *Control de Criaturas Fantasmagóricas* de Kaggle, ya que este sitio sufrió un embrujo similar en 2016, donde reclutó colaboradores a través de una competencia. En dicha ocasión, los colaboradores en Kaggle pudieron identificar correctamente a 300 criaturas. Para el ITAM es crítico poder clasificar de manera correcta a las criaturas, debido a que el procedimiento para deshacerte de ellas es distinto dependiendo de la especie de monstruo. Se sabe que existen 3 tipos diferentes de criaturas: $\{ghost, goblin, ghouls\}$. El Departamento de Estadística recopiló los datos que consideraba confiables de Kaggle que continenen información de las criaturas capturadas en dicha ocasión, tales como su clasificación y características. Dichos datos los puso a disposición de ustedes vía Canvas (kaggle_train.csv). La *Tabla 1* incluye un diccionario de las variables incluidas.

Reto A: Código para clasificar monstruos.

Entrega de esta parte: **Miércoles 8 de noviembre**.

Utilizando la base de datos (kaggle_train.csv) tendrán que entrenar un código para clasificar a los monstruos encontrados en el ITAM. Es importante notar que la base de datos que se les entrega incluye la clasificación correcta de los monstruos capturados en el embrujo de Kaggle. En el caso de los monstruos que han aparecido en el ITAM, el Departamento de RI esta recopilando una base de datos con las características que puede observar de los monstruos, pero dicha base no tendrá su clasificación. Precisamente rectoría esta esperando que, usando estas características y su código, ustedes puedan decir qué tipo de criatura se trata en cada caso, para poder eliminarla.

Para inspirarse en la creación de su código, pueden consultar los distintos códigos que se emplearon en la competencia de Kaggle. En específico, se les recomienda consultar códigos en ***R***, basarse en alguno de ellos y modificarlo para adaptarlo al embrujo del ITAM, aportando algo a dicho código. Su código deberá tener un marcador o señalamiento en las secciones del código que modificaron y además incluir el link al código original. **Ojo:** los miembros del Departamento de Estadística revisarán y compararán el código original con la modificación que ustedes hayan hecho. Para cumplir con la *Misión 1*, se requiere presentar una nota ejecutiva que describa las modificaciones realizadas en el código y justifique las razones que llevaron a tomar dichas decisiones. Asimismo, en esta nota deberán describir

con alguna gráfica o tabla cómo le hicieron, con la base de datos que se les entregó, para juzgar si su código era bueno.

Por último, en esta nota deberán describir brevemente cuáles fueron las contribuciones de los distintos miembros de su equipo en este trabajo. El código y la nota representan el 70 % de la calificación de su *Data Challenge*.

Reto B: Diseñando un monstruo

Entrega de esta parte: **Martes 7 de noviembre**.

Al rondar por los pasillos embrujados del ITAM, te encuentras con un monstruo que parece ser amigable. Usando gestos y algo cercano a un lenguaje lugubre, le pide ayuda a tu equipo para que el Departamento de Estadística no lo vaya a capturar. Usando tus habilidades de hackeo que te enseño un compañero de Ciencia de Datos, entras a los registros de monstruos recopilados por RI y encuentras la línea con las características precisas del monstruo que te pidió ayuda. Modifica las características de dicho monstruo de manera tal que la mayoría de los otros equipos no lo puedan clasificar correctamente, pero algunos sí lo logren para que no sea demasiado obvio que lo ayudaste en caso de que te descubran. **Nota:** Tu equipo debe de ser capaz de clasificar correctamente su monstruo modificado utilizando su código.

Una vez que tengas listas las características que pondrás, deberás subir en *Canvas* una línea en la base de datos que corresponda a tu monstruo. En esta base debes, además de inlcuir las característcas, indicar qué tipo de monstruo es, asi como darle un nombre a tu monstruo que usarás para nombrar la base de datos.

Más adelante se describe cómo tu monstruo te puede ayudar a ganar puntos adicionales para el examen final.

Reto C: Cacería de monstruos

Día del *Monster Hunt: Miércoles 8 de noviembre.*

Tu profesor agregará los monstruos enviados por los distintos equipos a una base de datos (a.k.a. caldero mágico) con la información recopilada por el Departamento de RI de los monstruos que fueron identificados rondando por el ITAM. Haciendo uso de tu código en tiempo real, es decir, durante el tiempo de clase, intentarás clasificar correctamente a todos los monstruos. Tras ejecutar tu código enviarás al profesor la clasificación predicha para los

monstruos. Deberán entregar archivo *.csv* donde la primer columna sea el *id* de la criatura y al segunda sea la categoría en la que clasificaron a la criatura *type*. Ejemplo de Archivo

En esta actividad se revelará la identidad de cada monstruo con lo cual podremos evaluar el desempeño de los distintos códigos. Para calificar esta parte, cada equipo tendrá chance de clasificar incorrectamente 20 % de los monstruos sin ser penalizados. A partir de entonces, cada monstruo mal clasificado representará una penalización de $100 * \frac{1}{N_p}$ puntos, donde $N_p = 0.8N$ y N es el total de monstruos que se les entreguen para clasificar el día del Monster Hunt.

Ejemplo: si se les entregan 200 monstruos para clasificar, pueden equivocarse en 40 monstruos sin perder puntos. Si algun equipo se equivocara clasificando 48 monstruos en total (teniendo 152 correctamente clasificados), su calificación de esta parte sería de 95.

Puntos extra basado en el Reto B

Los puntos extras se aplicarán sobre el examen final a todos los miembros del equipo por igual.

Si tu equipo creo un monstruo que ... % de equipos que lo descubrieron	Puntos
(45 %, 55 %)	0.6 puntos
(35 %, 45 %] o [55 %, 65 %)	0.4 puntos
(25 %, 35 %] o [65 %, 75 %)	0.2 puntos
e.o.c.	0 puntos

Tabla 1: Descripción de variables.

Variable	Descripción
id	identificador único de la criatura
bone_length	largo promedio de loss huesos en la criatura, normalizado entre 0 y 1
rotting_flesh	porcentaje de carne podrida en la criatura
hair_length	longitud promedio del cabello, normalizada entre 0 y 1
has_soul	porcentaje de alma en la criatura
color	color dominante de la criatura: blanco, negro, transparente, azul, verde, sangre
type	variable objetivo: Ghost, Goblin y Ghoul