Logo, company name

Description automatically generated

**Introducción a la Ciencia de Datos**

Parcial 09 de Octubre 2023

1. **Suponga que usted trabaja como científico de datos para una empresa de e-commerce que quiere mejorar su sistema de recomendación. El conjunto de datos de la empresa incluye información sobre las interacciones de los clientes con los productos, como compras, clicks y búsquedas. Qué tipo de features pueden crearse a partir de este tipo de datos?**

Pueden crearse multiples tipos de features tales como:

* Engagement features: numero total de clicks, numero total de compras, numero total de búsquedas.
* Features de comportamiento: tiempo total en la plataforma, si el usuario es mayoritarimente un comprador diurno o nocturno, si el usuario entra en días particulares, si el usuario compro en días de descuentos, si el usuario es nuevo en la plataforma o no.
* Features demográficas: si el usuario pertenece a un cluster especifico demográfico, si es un comprador avido o no, total de dinero en sus compras.
* Features de los productos: si el producto es popular, si esta en descuento, cluster de productos similares, rating del producto, numero total de clicks del producto, etc.
* Interacciones binarias producto de un pivot.
* Time related features: por ejemplo dia de la semana, mes, hora, si el dia es una fecha especial (navidad, año nuevo, Halloween, pascua, black Friday).
* Otras.

1. **Su cliente le proporciona un dataset para que usted entrene un algoritmo de clasificación binaria. El cliente le sugiere que quiere evaluar la performance predictiva usando *accuracy*. Usted, tras analizar la distribución del *target* o columna objetivo, le aconseja no usar esta métrica. Qué es lo que usted notó particular en la distribución y qué métrica aconsejaría utilizar?**

Como hemos visto en clase, la metrica accuracy para la evaluación de modelos frecuentemente puede resultar engañosa. Hemos visto en clase que un gran problema de usar accuracy es cuando tenemos un dataset desbalanceado. Es decir, que en nuestra columna objetivo tenemos una categoría mayoritaria y una minoritaria en razones desproporcionadas. Hemos visto que un modelo “ciego” (modelo que siempre predice la categoría mayoritaria sin importar las input features X) tendrá una accuracy muy alta, solamente por el simple hecho de la proporción inherente en nuestros datos.

1. **Un modelo fue entrenado para predecir el valor de propiedades en el rubro de bienes raíces con datos del 1990. Usted utiliza este modelo para predecir valores de propiedades actuales y realiza a posteriori un análisis de los residuos obteniendo el diagrama que se ve en el anexo. Interprete el mismo y justifique que acción se debe tomar.**

El problema en cuestión es que el modelo fue entrenado con datos muy viejos. Por la naturaleza misma del rubro de bienes raíces, la columna objetivo (precio de las propiedades) fluctúa acorde a variables que tienen una dependencia temporal. Por ejemplo, inflación, apreciación/depreciación de una zona, deterioro, tipo de cambio,etc. Por tanto, un modelo entrenado en datos antiguos no tendrá una buena performance con los datos actuales. En concreto, en el diagrama de residuos podemos ver claramente que el modelo tiene un sesgo, prediciendo valores inferiores a los que en realidad son. Esto es un fenómeno frecuente, ya que las propiedades muchas veces con el tiempo tienden a valorizarse. El fenómeno opuesto podría suceder también en otro escenario o rubro. Lo importante es identificar un sesgo en las predicciones de nuestro modelo. La acción a tomar es reentrenar el modelo con datos actuales o (posiblemente con efectos peores)hacer un ajuste de nuestra variable objetivo al tiempo actual.

1. **Por qué cree que es importante lidiar con outliers? Justifique su respuesta dando al menos tres razones.**

Existen multiples razones, entre ellas:

* Los outliers pueden tener efectos adversos como sesgos a la hora de entrenar modelos pues introducen sesgos y deterioran la performance.
* En unsupervised learning puede distorsionar la forma de los clusters dependiendo el algoritmo que usemos.
* Mejora en visualizaciones: remover outliers puede mejorar distintas graficas al removerlos.
* Lidiar con outliers puede mejorar la calidad de nuestro ingreso de datos pues pueden revelar anomalías, errores de entrada, malfuncionamiento de medidores.
* Lidiar con outliers puede revelarnos comportamientos inusuales.
* Mayor interpretabilidad al hacer un estudio analítico.

1. **Por qué razones cree que las gráficas presentes en el anexo no son apropiadas? Justifique su respuesta.**

La grafica del mapa de USA fue impresa en blanco y negro por lo cual no se pueden apreciar las distinciones. Asimismo la grafica carece de títulos, que se esta graficando? Tampoco parece haber sido realizada con colores en una escala continua, por lo que al ser discreta no nos da una estimación de que tan fuerte es el valor de la variable.

La grafica parece lineal pero en realidad no lo es. Los valores en el eje x no están equidistantes en el tiempo. Tampoco es razonable hacer una comparación entre estos periodos.

En la ultima grafica por que los valores del eje Y están tomados a esos intervalos? Por que utiliza una grilla? Hay razón de hacer un grafico solo para informar que solo 3 provincias tienen edades de 18 y el resto 19? Es la mejor grafica un grafico de barras? Quizas podría ser mejor un mapa geografico binario.

1. **Suponga que usted posee un dataset de 3 columnas numéricas continuas, la cual una de ellas presenta missing values. Como imputaría los valores faltantes de la misma utilizando KNN? Describa detalladamente el proceso incluyendo las fórmulas que considere necesarias.**

Para imputar valores faltantes usando KNN usted debe definir el hiperparametro *cantidad de vecinos* a un valor razonable*.* Previamente, usted debe escalar las columnas a un rango común para que las distancias no se vean afectadas por una desproporcion en las escalas de las mismas y potencialmente remover outliers. Para cada punto el algoritmo debe calcular la distancia a cada uno de los otros puntos con valores no faltantes (pair-wise distance). La distancia usualmente es la euclideana (podría ser Manhattan o Minkowski pero es poco frecuente). Una vez que determina los puntos mas cercanos puede imputar el valor haciendo un promedio ponderado de los vecinos mas cercanos.

1. **Explique los distintos componentes de una serie de tiempo en base al ejemplo ilustrado en el anexo.**

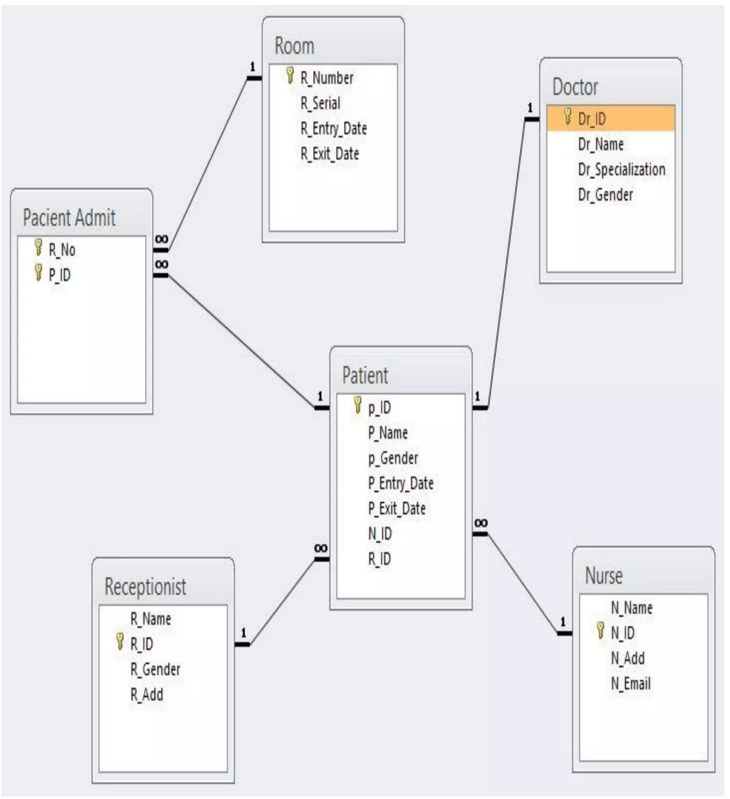
Hemos visto en clase que una serie de tiempo puede desglosarse en cuatro componentes:

1. Tendencia: en la grafica puede verse una tendencia clara entre 1960 y 1975.
2. Estacionalidad: Se puede ver claramente que hay una periodicidad anual por los frecuentes picos y valles de la serie de tiempo.
3. Cambios cíclicos: Los cambios cíclicos solo pueden apreciarse en un panorama mayor, en la grafica se pueden ver dos cambios cíclicos representados por dos valles abruptos en 1975 y 1983. Pueden deberse, por ejemplo, a crisis económicas
4. Error irreducible: la grafica no tiene la granularidad suficiente como para apreciar el error irreducible. Sin embargo, si hiciéramos “zoom” podríamos ver fluctuaciones aleatorias a un nivel mas bajo por ejemplo semanalmente o diariamente.
5. **Calcular la Information Gain para la columna Status presente en la tabla del anexo.**

Ver anexo con foto de solución.

1. **Usted trabaja en un nuevo hospital. El director del mismo le solicita que cree la base de datos del mismo. Realice el diagrama Entidad-Relación del mismo identificando al menos 5 entidades, especificando las relaciones entre las mismas, incluyendo atributos y keys.**

En esta pregunta cada entidad debe estar correctamente definida, debe presentar al menos 2 o 3 atributos y especificar cual es la primary key o foreign key para establecer las relaciones pertinentes entre las distintas entidades. A modo de ejemplo se adjunta un diagrama similar.



1. **Corredores de distintos países corren 10 carreras y obtienen los tiempos que se muestran en la tabla del anexo. La desviación estándar para los corredores del país 1 es de 103 segundos mientras que para los corredores del país 2 es de 101 segundos. Estableciendo un nivel de significancia alpha = 0.05, existe una diferencia significativa en la media de tiempos de ambos corredores?**

**La tabla del anexo posee valores negativos del z-score porque es de la cola izquierda, pero es igual para los valores positivos del z-score si fuera la cola derecha.**

Acá en este ejemplo se hace un two sample z-test para medias. Hay que plantear la hipótesis nula y la hipótesis alternativa, usar la fórmula vista en clase y obtener el z-score. El valor del z-score da 0.578 se usa la tabla y el p-value asociado es aprox 0.28. No se rechaza la hipótesis nula ya que para el nivel de significancia 0.05 el p-value es mayor.

A paper with writing on it

Description automatically generated