**Guión:**

**(diapositiva 1)**

Buenas tardes, mi nombre es Alejandro y os voy a hacer una presentación donde intentaré dejar claro cuál ha sido el proceso que hemos seguido para aplicar modelos de regresión a una base de datos que nos han aportado en el material de prácticas.

**(diapositiva 2)**

Estos son los objetivos que hemos querido llevar a cabo y que voy a presentar:

1. Hacer un estudio de la base de datos aportada para entender estos datos y seleccionar los más representativos.
2. Aplicar modelos de regresión y estudiar los resultados.
3. Analizar los errores e intentar mejorarlos con distintas técnicas.
4. Como hemos resuelto el problema.
5. Y finalmente una conclusiones.

**(diapositiva 3)**

Nuestra base de datos contenía datos reales sobre las construcciones de apartamentos unifamiliares en la provincia de Teherán, Irán. Los datos estaban divididos en tres grandes bloques:

-En el primer bloque teníamos las fechas de los proyectos, donde en un inicio teníamos dudas de como estaban representadas estas fechas ya que en Irán actualmente es el año 1397.

-Un segundo bloque donde tenemos todas las variables físicas y financieras, donde tenemos el costo de construcción y el precio de venta de estos apartamentos.

-Y por último un tercer bloque con las variables económicas.

De todos estos datos, teniamos 2 que creíamos que eran los de principal interés para asignarlo como variable objetivo. Uno era el precio de los apartamentos y el otro el coste actual de la construcción. Finalmente decidimos utilizar el coste de la construcción, porque creemos que puede servir a una constructora para predecir cuáles van a ser sus costes y decidir si llevar a cabo el proyecto o no.

También comentar antes de todo que cuando hablemos de costos, estaremos analizandolos en IRR, que es la moneda de Irán, donde 1 euro son aproximadamente 48.600 IRR.

**(diapositiva 4)**

Para poder ver mejor las relaciones entre las distintas variables, hemos decidido hacer una

matriz de correlación, de esta manera podremos centrar mejor nuestra atención en cuáles

son las variables de interés y cuáles son aquellas que podemos descartar.

Algo que nos ha llamado la atención al ver estas correlaciones, es que la V2 y V3 (que son

variables relacionadas con el área de construcción) no tienen casi influencia en el coste y el

precio del apartamento. Por lo tanto, tanto estas variables como todas las que no tengan

influencia en el coste de la construcción, no las tendremos en cuenta a la hora de predecir.

Si vemos la matriz, podemos ver que la variable target (V10, que es el coste actual) tiene

una correlación alta con V5 ( que es la previsión de costes basada en los precios). Esto tiene

sentido, y significa que si esa variable sirve para hacer una predicción los estudios de

previsión que se han ido haciendo hasta el momento eran acertados.

**(diapositiva 5)**

Para poder tener claro también si había alguna variable importante o no, lo que hicimos fue calcular el MSE y el R2 para cada una de las variables individualmente, y comparar los resultados en un histograma. Como podemos observar, la variable V5 con diferencia es la que nos da un MSE más pequeño, y desde este histograma también podíamos observar con qué otras variables podríamos ir combinando.

**(diapositiva 6)**

Y otro valor que hemos analizado y que creemos que es importante es el R2, ya que nos permite orientarnos sobre si nuestra predicción está siendo precisa. Viendo el histograma tenemos claro que de nuevo V5 es la que nos está dando un mejor resultado y es un resultado muy cercano a 1, y poder descartar aquellas que tengan un resultado muy bajo.

**(diapositiva 7)**

Para la parte de desarrollo hemos utilizado jupyter notebook y spyder. En un inicio comenzamos usando un proyecto de spyder, pero finalmente sentimos que era más cómodo

ir trabajando en celdas como en una jupyter notebook, separando el código en celdas distintas según queramos obtener unos resultados u otros.

**(diapositiva 8)**

Las librerías que hemos usado son las mostradas en esta dispositivas, pandas para el trato de los datos en formato .csv, numpy para poder tener una estructura de datos rápida y trabajar las matrices de manera más cómoda, sklearn para hacer las regresiones y predicciones y matplotlib para dibujar las gráficas.

**(diapositiva 9)**

Para el desarrollar, siempre hemos ido representando gráficamente todos los resultados, porque esto nos permitía ver si el modelo se podía mejorar o si el espacio muestral no era correcto para aplicar una regresión lineal. Por ejemplo, la primera gráfica en un inicio parece que el espacio muestral puede ser óptimo para una regresión lineal pero hacia el final los datos comienzan a subir, por lo que quizá sería más eficaz una regresión no lineal. Y el último ejemplo, hay muchos datos concentrados y no parece ser una muestra que nos vaya a ayudar demasiado.

**(dispositiva 10) (diapositiva 11)**

Otro de los procesos que hemos llevado a cabo en el desarrollo ha sido la eliminación de datos inusuales que podían hacer que nuestro error fuera peor de lo que realmente era. Eliminando estos datos inusuales del conjunto de datos de X e Y, obtenemos una mejora en el MSE y el R2.

**(diapositiva 12)**

Como vemos en esta tabla, en la gráfica del ejemplo anterior, al eliminar el dato inusual el MSE mejora de 2093 a 1279 y el R2 también sube de 0,92 a 0,94.

**(diapositiva 13)**

En cuanto al tema de estandarización, hemos probado de mejorar el mejor modelo que teníamos hasta el momento estandarizando los datos, pero en nuestro caso no ha tenido prácticamente efecto. Creemos que esto es debido a que tanto la X como la Y se están tratando en medidas similares, ya que las dos están expresadas en IRR. En esta gráfica la parte naranja que son los datos estandarizados está interpuesta a la azul que son los datos sin estandarizar, y prácticamente no hay diferencia.

**(diapositiva 14)**

Probamos de hacer combinaciones de la mejor variable individual con algunas de las variables que habían obtenido mejores resultados de MSE y R2, o bien con variables que entendiendo el significado de estas, podíamos interpretar que complementarían a la variable individual. Este es un ejemplo donde como variables del conjunto X tenemos V5 y V8, que hacen referencia a la estimación de coste teniendo en cuenta los costes iniciales del proyecto y la V8 que son los precios por unidad. Como vemos en los resultados no obtenemos una mejora, si no que más bien hace aumentar el error. Probamos con distintas combinaciones y no conseguimos ninguna que superara al coste de la V5 individual, como mucho que lo igualaran.

**(diapositiva 15)**

En esta grafica hemos puesto un ejemplo donde al normalizar los datos si que mejoran brevemente los resultados, pero aun asi solo consigue igualar el valor de V5.

**(diapositiva 16)**

En cuanto a resultados, estos análisis se han llevado a cabo creando dos grupos en los datos, los de entrenamiento y los de validación. En un inicio hemos hecho todas las comprobaciones con un 80% de los datos en el conjunto de entrenamiento y un 20% de datos en el de validación, pero más adelante hicimos una prueba con un 50% y 50%, y aunque hubieron cambios pensábamos que influiria más. Lo que sí notamos es que con un 50% en cada conjunto, los resultados dados eran más inestables por lo que suponemos que harian que el modelo fuera menos fiable.

**(diapositiva 17)**

Y bueno como resultado final, no hemos conseguido ninguna combinación que mejorara al modelo que utiliza solo V5. Vemos que por el espacio muestral de los datos, era un ejemplo perfecto para aplicar una regresión lineal, y que eliminando los datos inusuales, hemos conseguido un R2 del 94%, lo que en un principio parece ser un buen resultado.

// TODO: conclusiones y actividad