Data Underestanding, Data fetching techniques, EDA Techniques are mated in Jupyter Notebook. I was a live works also all salet -Feature Engineering & bib and oaking doors sandshipment on bandships → Data Estruct Algorithm aborona II A Dala engineer softner years works that you one Data Proparation herton silt office agree Common dol Reiterale till statisfactory model performance Feature Engineering edune! Feature Feature Feature Tranfo unation Extraction Selection Welliam Construction - Missing value Impulation Scheet Necessary Programmtically - Handling categorical features for Extract a Features Outlier Detection roja Johan new feature Add new from Features STOM YMAN -Feature Scaling existing one mom existing

Feature Scaling (Standardization): Almony Noted Before.

Lab instructions:

- J Import librories
- Import social Network detaset and trake three features from it (Age, Estimated Sel, Punchased)
- of Train, test, spiet
- Standard Scalere
 - Convert scaled data to DF again
 - Tcheck of describe 1). on both X-train and scaled deleset
 - Plot 2 subplats of before and after scaling data to check what charges happen
- I Plot KDE Polat in the same way
 - Check individual Distributions of Age and Sulary vering KDE plot (Before and afters)
 - Apply logistic Regression on both gealed and sunscaled dates to check who performs better. If scaling actually matters on not?
 - Apply Decision Trees in the same way to understand that not also all algorithm will work better on sealing deta.

Encodercs: Most popular encoders are

- 1) Ordinal Encodere

2) Label Encodera [Ata Basic workings

3) One Hot encoderc already noted)

When to use what?

Oredinal encodere -> Use it in oredinal contegory techne where you need to provide orders. But use it only for when the pure Independent Features

Label encoder -> It has been used for to encode Tranget features which are eategorical

One hot encoder -> It is used in Nominal category features who are not on ordered feature (means they an't be ordered)

pilos no nost a des. Un malingle la site due

Durmmy Variable Treap:

After doing one hot encoding, for a categorical feature which has no unique altegories, for that n new columns get created. But fore, no unique categories, from the n created new columns, we have to delete a column To solve the problem of multi collinearity. (Generally it has to be the Int column which needs to be deleted). So for no categories (n-1) columns will be kept after One hot encoding.

In a machine learning model, the input (independent) teatures should be not independent to each other means they should not have any mathematical relationship with each other. So, after one hot encoding as n new columns (feedures) get created, they show a mathematical relationship with each other for which it becomes a problem for some linear models like (linear Regnession, Logistic Regnession). So to solve the problem were need to remove I column. After ramoving I column with (an-i) columns, n categories still can be represent so that won't be a issue.

One hat encoding using most frequent Categories ignore shows promote

Sometimes a categorical column can have so much columns that it would not be voice to excete no new columns with that after one hat encoding. In those cases, we do a little bit modified one hat encoding. Suppose, we create new columns for those categories which are more frequent in the data column, and for remaining categories, we make a single feature morned 'Others' and encode all the lower frequent categories with that column.

The threeshold of shifting categories to "others" can be decided by person.

mediternedical recludionship with each other. It after one had one of a medienalist

relationalis with each other the alich it teames a problem

gone linear modals like (Luar Repression, Logistic Repression) - So-

solve the problem upo need to minore I column. All or minor

the school and the colonies or solution of the manifes I

ensel o sed - (mon)