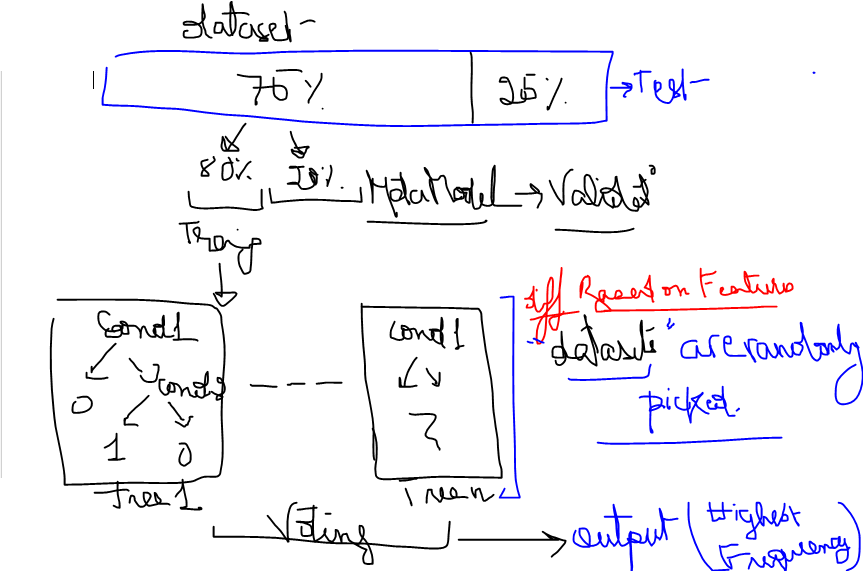
# URL:

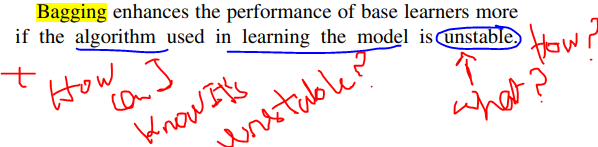
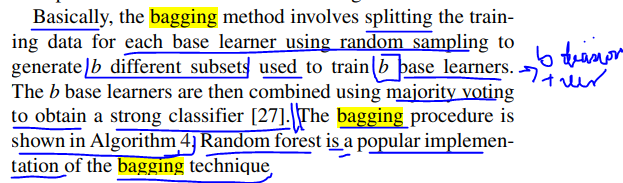
* Technical Doc: [RandomForestClassifier — scikit-learn 1.5.1 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)
* Youtube Video: <https://www.bing.com/videos/riverview/relatedvideo?&q=What+is+Random+Forest&&mid=9D2F45B57715544C41599D2F45B57715544C4159&&FORM=VRDGAR>
* Paper 1 : [IEEE Xplore Full-Text PDF:](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9893798)
* Paper 2: [A survey on ensemble learning | Frontiers of Computer Science (springer.com)](https://link.springer.com/article/10.1007/s11704-019-8208-z)
* Technical Doc: [Random Forest Classification with Scikit-Learn | DataCamp](https://www.datacamp.com/tutorial/random-forests-classifier-python)
* MultiLabel Classification: [pandas - How to use MultiLabelBinarizer for Multilabel classification? - Stack Overflow](https://stackoverflow.com/questions/65600815/how-to-use-multilabelbinarizer-for-multilabel-classification)
* MultiLabel Classification : [Scikit-Learn's preprocessing.MultiLabelBinarizer in Python (with Examples) | PythonProg](https://www.pythonprog.com/sklearn-preprocessing-multilabelbinarizer/)
* [python - How to choose n\_estimators in RandomForestClassifier? - Stack Overflow](https://stackoverflow.com/questions/60768008/how-to-choose-n-estimators-in-randomforestclassifier)

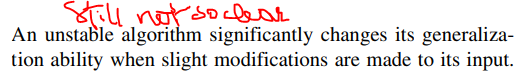
# How to start

Faezan, Random Forest, Is really close to Decision Trees, The diffference, Is that Random Forest, is much less sensitive to Training Data, Why?

-> Because when we are training / Creating the trees, we select the features on which we want to base our tree – This is called Bootstrappin  


Objective: Enhance the classification per616 formance of ML models by combining the predictions from 617 randomly generated training sets.





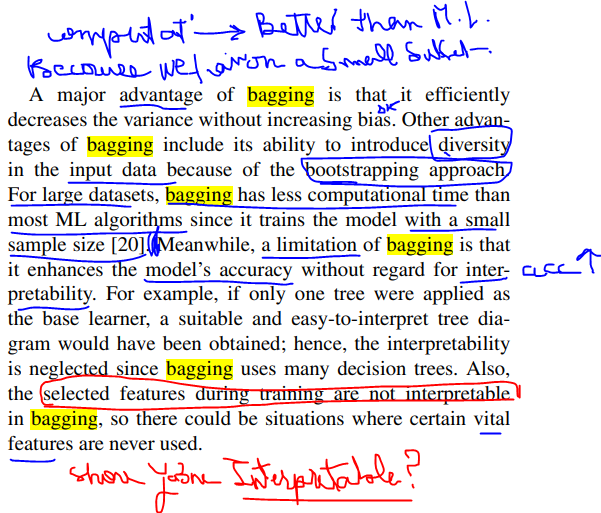
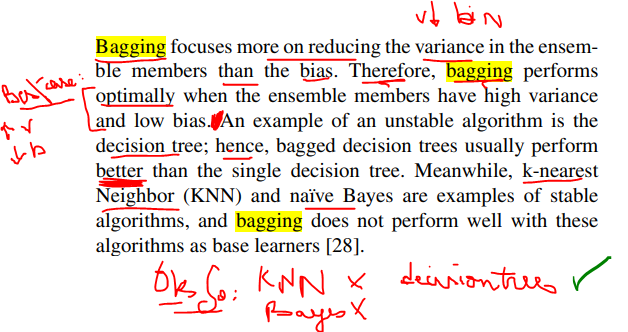
# Process:

Dataset used to Train, we will take the datasets we are sure of, What is the dataset?  
Nehna lyom mana edrin naamol 75/25 splitting as we please.

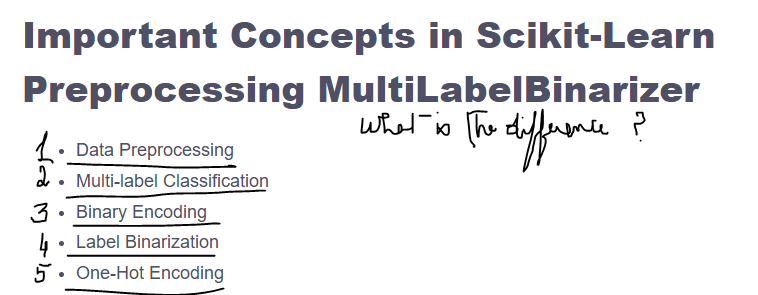
We are limited. Kif?

To train -> Dataset where 1&2&3&4 =0 and 1&2&3&4 =1 - Hol columns  
Yaane on ml-small-latest dataset:

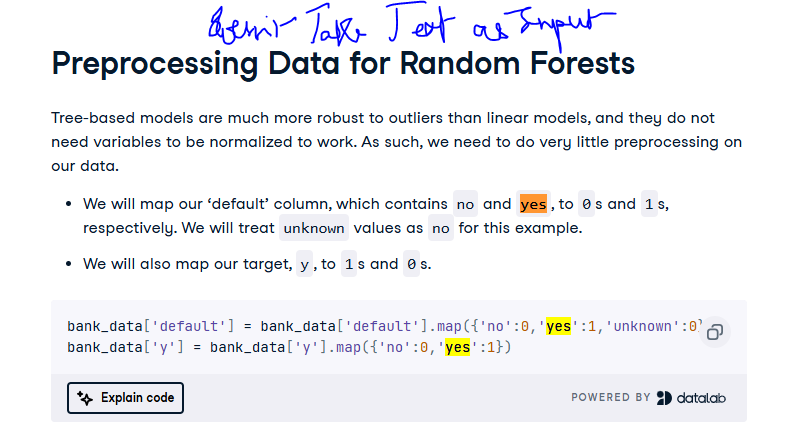
|  |  |  |
| --- | --- | --- |
| **1&2&3&4 = 0** | **6634** | **32.8936%** |
| **1&2&3&4 = 1** | **1** | **0.0049%** |
| **The rest (Testing)** | **20168 – 6635 = 13533** |  |
| **Not enough in my opinion** |  |  |
| **Total** | **20168** | **32.89855%** |

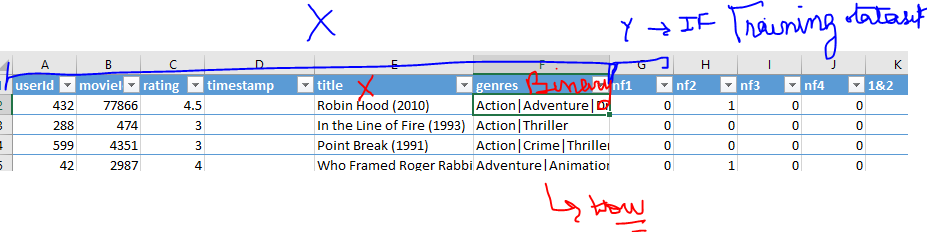


## How do we handle text input/output in this algo?

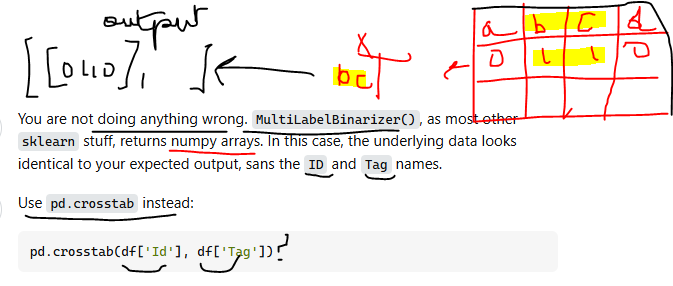


1/ Data Preprocessing -> Usual replace/ join ....  
2/ Multi class output -> For videos/pictures/ an object can belong to multiple classes, for example a dog can be young, animal, female...  
3/ Binary Encoding -> yes is 1 0 is no, done on single labels  
4/ Label Binarization -> on multi labels -> What we will use  
5/ One-Hot Encoding -> single-label encoding for simplicity yaane bikoun fi colonne la genre thriller lahal (1/0) romance (1/0) ... akid wont use

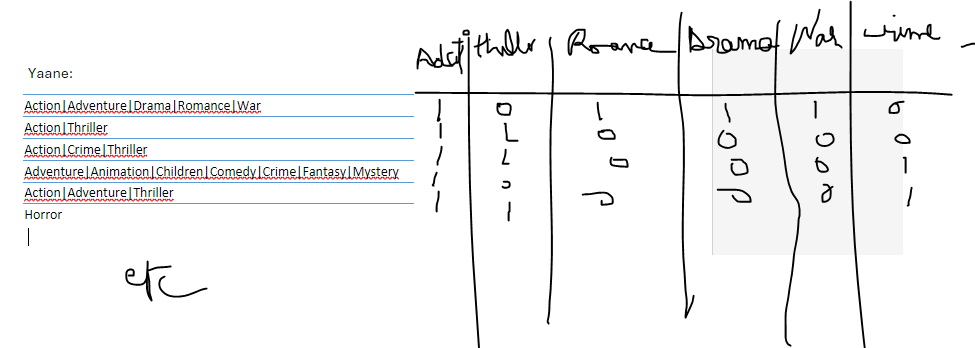
So, for genres, and title, we can’t leave them to be handled

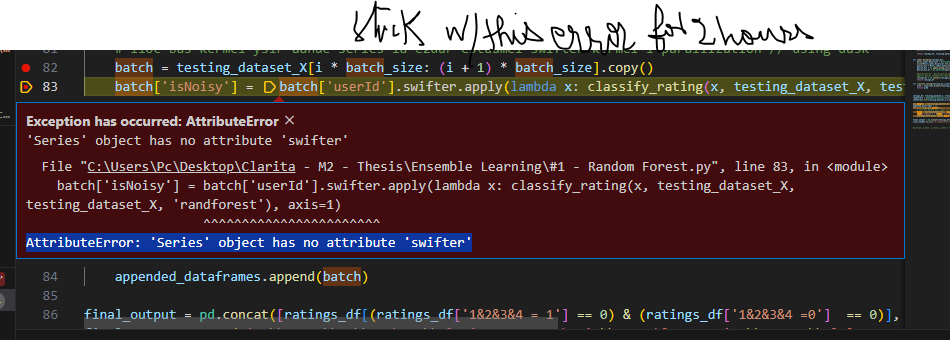


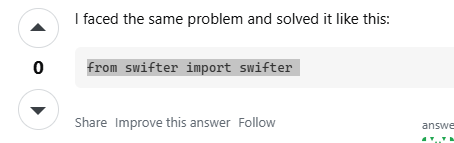
We will use the multilabel binarizer used in NLP, to be able to transform the genres to numbers, As for the name, it is not needed, as we have the id to identify it, we didn’t take the year into consideration – Can be added

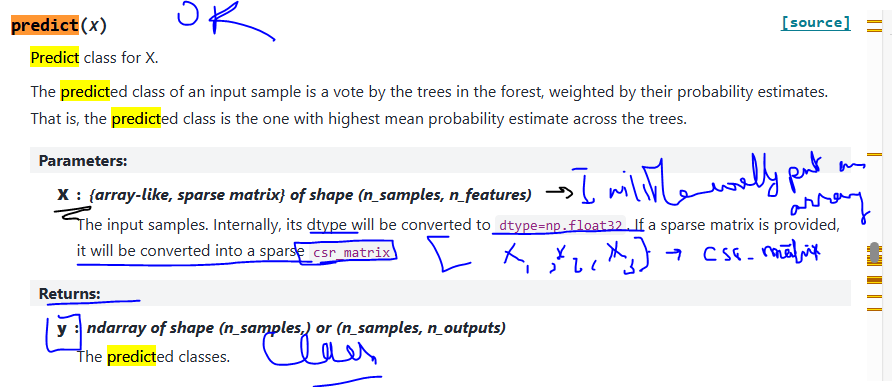
example:  
Yaane:

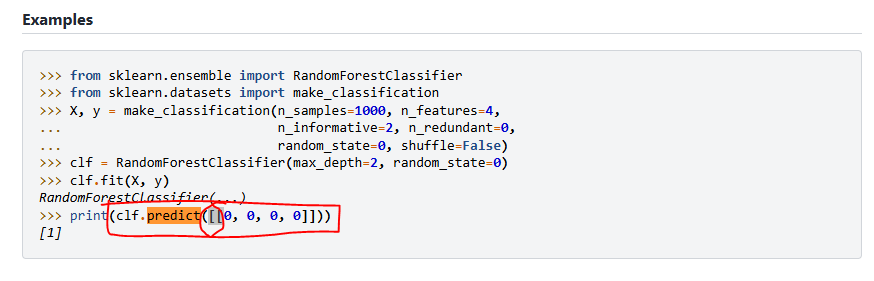
|  |
| --- |
| Action|Adventure|Drama|Romance|War |
| Action|Thriller |
| Action|Crime|Thriller |
| Adventure|Animation|Children|Comedy|Crime|Fantasy|Mystery |
| Action|Adventure|Thriller |
| Horror |

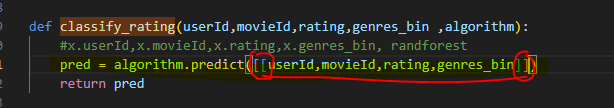
Swifter for parallelism

[python - How can I fix the error AttributeError: 'Series' object has no attribute 'swifter'? - Stack Overflow](https://stackoverflow.com/questions/62843445/how-can-i-fix-the-error-attributeerror-series-object-has-no-attribute-swifte)





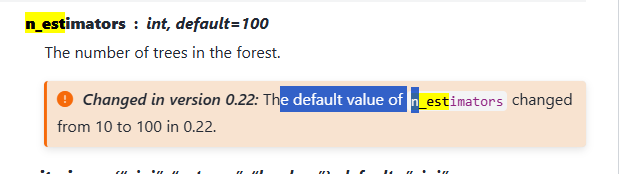




Timestamp is still empty bi aleb l dataset laean 25m/5m doenst open/wl 1m msayve bala timestamp

Laheik u should find a solution

N\_estimators by default is 100



Oke aam taamoul error when I added all the data

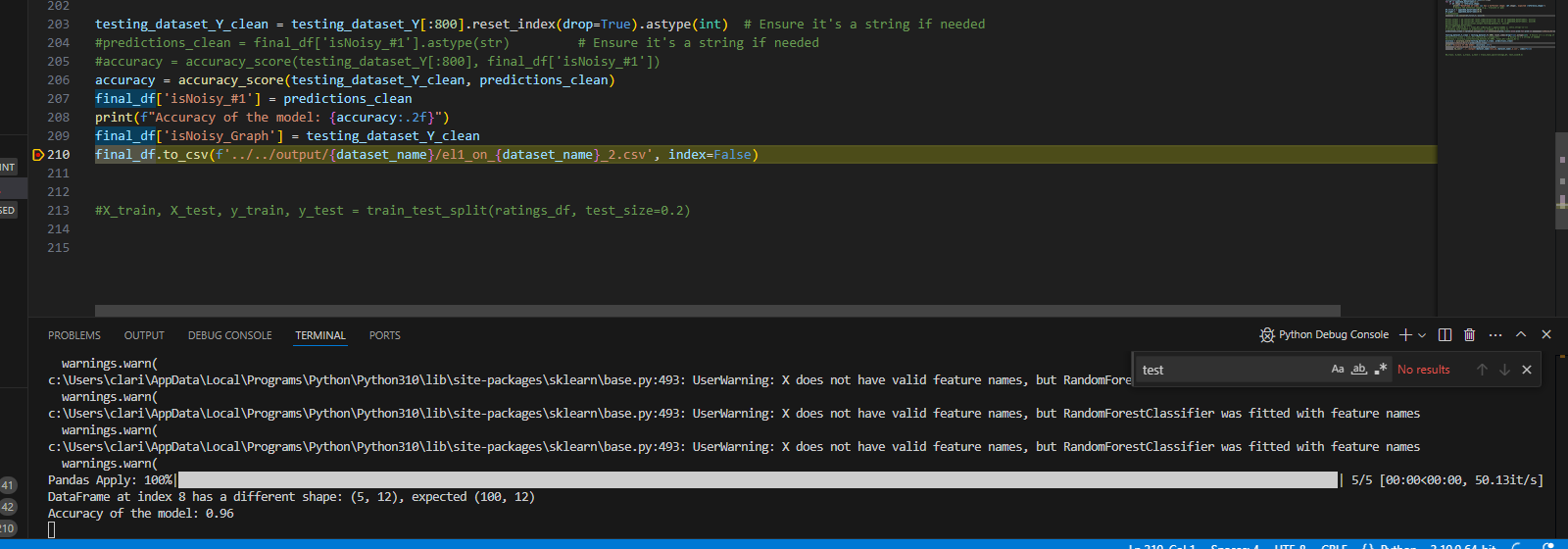
batch['isNoisy'] = batch.swifter.apply(lambda x: classify\_rating(x.userId,x.movieId,x.rating,x.timestamp,x.genres\_bin,x.nf1,x.nf2,x.nf3,x.nf4,x.cluster\_ab\_nDCG,x.equiv\_ab\_nDCG, x.ab\_ndcg,x.perc\_change,x.condition\_1, x.user\_serendipity, randforest), axis=1)

Hon aam bttoj aal prediction, All fi sequence,beete2ed l mashkal mn gher shi

I added a part to label some fields that exist already like

, ['user\_group\_x', 'user\_cat\_x', 'rating\_group\_x', 'item\_cat\_x'])

Test 2 takes into consideratoim labeled data from the algorithm we set



Everything was set to 0

So no need to redo the grp val and Ser

As the delta will be 0

If 0 -> nothing changed!

Bas hon it’s a sma;;lll dataset!