Remove Duplicate Columns

Introduction

Removal of duplicate columns is an essential part of data preprocessing or cleaning. To understand this, think of a scenario where you are reading two books to prepare for an exam (which will be based on 5 different books) and both the books are identical. You are not going to learn anything new but you are complicating your exam preparation. As it will consume most of your time without providing any new concept. As well as you will be prepared very well on the concepts covered in those books but you won't have good knowledge of other subjects/concepts.

The simple reason why we need to remove duplicate columns is that, if more and more similar features will be provided to a machine learning algorithm then it will get biased towards these features, hence decreasing the variance of training data.

Objective

In this play you will learn about:

- How to identify duplicate columns?
- How to remove duplicate columns?

There are two different ways of removing duplicate columns based on the volume of data you have.

If a dataset has rows and columns less than 50000 and 100 respectively, then the steps will be:

1. Transpose the dataset i.e. rows will become columns and columns will become rows.
2. Pandas has a method to find duplicate rows. Use pandas "drop_duplicates" method to drop duplicate rows of transposed data frame and again transpose the data frame. The data frame will be again same as it was in step 1 before the transpose operation.
3. Dataframe returned in the step 2 won't have any duplicated columns.

If a dataset has rows and columns in access of 50000 and 100 respectively then a custom python has to be written to remove duplicate columns as transpose is an exhaustive operation in terms of memory and processing as well as drop_duplicates() and duplicate() methods also do not work when there are more numbers of rows to be compared on the basis of huge number of columns as it throws Recursion Error because stack gets full.