Machine learning notes

- Building a dataset
- Import the libraries that we will working with
- Import the dataset
- Data Cleaning
- Analyzing the dataset features
- Splitting the dataset
- Encoding the categorical data

- ✅ dataset

https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning/data

- ✅ libraries (**Pandas, scikit learn**)
- ✅ Importing the dataset

To import the dataset, panda's library offers a method **read_scv(),** it takes the file path of the CSV file as an argument. It then returns a DataFrame containing the data from the CSV file.
And to ensure that the data is successfully imported we check it out by using the **head()** Methode of the dataFrame object.

- ✅ Data Cleaning (**Removing duplicates, Handling missing values**)

Removing duplicates: Identifying and removing duplicate records from the dataset. Duplicate records can skew analysis and modeling results by inflating the importance of certain observations.
Handling missing (null) values: Identifying and dealing with missing data points in the dataset. This could involve imputation (replacing missing values with a statistical measure like mean, median, or mode), deletion of rows or columns with missing values, or using advanced techniques like predictive modeling to estimate missing values.
Pandas' library offers huge methods allow us to clean our data:
Unique() : to know the unique elements of a dataframe column (in our case to know how many labels or target we have to encode it).
Isnull() : Detect missing values in the DataFrame
Fillna() : Replace missing values with specified values
Dorpna() : Drop rows or columns containing missing values
L**oc[: , :] and iloc[: , :]** : Select subsets of rows and columns based on labels or integer indices.

- ?? Analyzing the dataset features

Creating new features or modifying existing ones to improve the performance of machine learning models. This could include extracting useful information from raw data, combining multiple features, or transforming features to better capture relationships or patterns in the data.

- ☑ Encoding the categorical data

Because Machine learning algorithms operate on numerical data, by encoding categorical variables into numerical format, we enable algorithms to process and learn from the data effectively.

There are a lot of encoding types in machine learning (ordinal, one-hot ...)

Ordinal Encoding data is a method used to convert categorical variables (string format) into numerical format. It assigns a unique integer value to each category based on its order; it means that each number represent a unique category.

To encode data, we generally use the LabelEncoder class of preprocessing module from scikit learn library (Sklearn.preprocessing import LabelEncoder)

- ☑ Splitting the dataset

Splitting the dataset is a very important step in machine learning for training and evaluating predictive models. The dataset is typically divided into two or three subsets: training set, validation set, and test set.

Training Set: The training set contains a large portion of the dataset; it is used to train the machine learning model.

Validation Set: The validation set used to evaluate the performance of the model during training. It helps in preventing overfitting by providing an independent dataset to discover model performance.

Test Set: The test set used to evaluate the final performance of the trained model based some metrics like the accuracy which is an estimate of the model's performance on new, unseen data.