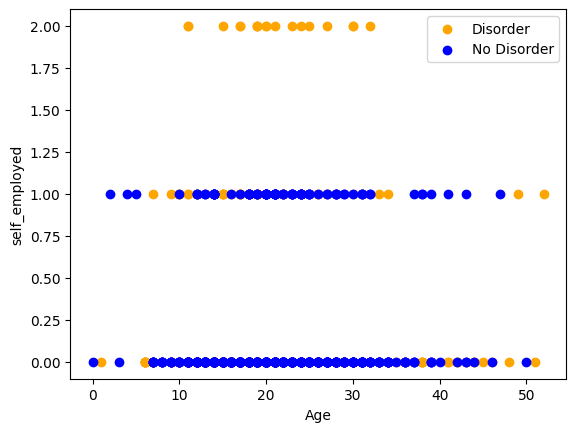
# Clustering on Mental Health survey data

In this part we take a mental health survey [data](https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey) and perform clustering in an unsupervised manner to extract semantic relations from the dataset. The dataset is based on a 2014 survey that measures the occurrence of mental health disorders and awareness in a workplace.



First we begin with the preprocessing of the dataset, since there are a lot of categorical variables we first convert the text classes into numerical classes which can be fed into the model using the labelencoder of sklearn. This is very important since most of the machine learning models cant comprehend the text labels and they have to be mandatorily converted into integer labels.

Next up we use the kmeans algorithm for our clustering. In crude sense, clustering means grouping the data based on the various indicators or features of the data without any supervision or labels. Kmeans is one of the most popular and widely used clustering algorithms. We use kmeans to sort the data into two clusters or we can say groups based on if there is a mental health disorder or not.



## CODE:

## Installation

Pip install scikit-learn matplotlib pandas

## Imports

import pandas as pd

from sklearn.preprocessing import LabelEncoder

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

## Data Loading

df = pd.read\_csv('survey.csv')

## Data Pre-processing

df = df.apply(LabelEncoder().fit\_transform)

print(sns.heatmap(df.corr()))

## Unsupervised Clustering

kmeans = KMeans(n\_clusters=2, random\_state=0, n\_init="auto").fit(df)

df['cluster']=kmeans.labels\_

## Plotting the results

plt.scatter(df['Age'],df['self\_employed'],c=df['cluster'])

plt.xlabel('Age')

plt.ylabel('self\_employed')

plt.show()