Group Details:

Group Name	SoleDS
Name	Anju Paul
Email	anjupaul14@yahoo.co.in
Country	UK
College	Anglia Ruskin University
Specialization	Data Science

Problem Description:

To understand the persistency of a drug as per the prescription given by the physician is an important question faced by the pharmaceutical companies. The problem here is to build a classification model to understand the persistency (persistent or not) of a drug for the given dataset.

EDA:

The 68 input variables needs to be reduced. The columns Risk_Type_1_Insulin_Dependent_Diabetes, Risk_Osteogenesis_Imperfecta, Risk_Rheumatoid_Arthritis, Risk_Untreated_Chronic_Hyperthyroidism, Risk_Untreated_Chronic_Hypogonadism, Risk_Untreated_Early_Menopause, Risk_Patient_Parent_Fractured_Their_Hip ,Risk_Smoking_Tobacco, Risk_Chronic_Malnutrition_Or_Malabsorption, Risk_Chronic_Liver_Disease, Risk_Family_History_Of_Osteoporosis ,Risk_Low_Calcium_Intake, Risk_Vitamin_D_Insufficiency, Risk_Poor_Health_Frailty, Risk_Excessive_Thinness, Risk_Hysterectomy_Oophorectomy, Risk_Estrogen_Deficiency, Risk_Immobilization Risk_Recurring_Falls are summed to Count_Of_Risks column. Hence all the above column can be dropped. This reduced the column count to 50. The Disease/Treatment Factor -the column name starts with 'Concom' or 'Comorb', this is grouped together by summing the number of positive for that and dropping the unwanted columns. This will reduce the number of columns of the dataframe to 28.

Then convert to numerical variable by applying the factorize method. Separate the x and y variables and divide it to test and train. Then apply standard scaler to the train and test values.

Apply PCA on that and plot a scree plot for identifying the number of major components. The scree plot has a leg after the first 6 factors. Hence I set n_components to 6 and apply the PCA for dimensionality reduction. The sum of explained variance from these 6 factors is 0.4691014092434735.