Group Name: PharmaPersist
Team Member Details:

- Name : Yunxin Gan
- Email: xin03040905@gmai.com
- Country: United Kingdom
- College: University of Exeter
- Specialization: Data Science

Problem description: To gather insights on the factors impacting the persistency, build a classification model for the given dataset.
Data that will be used in this project involves:

- Target variable (the variable that is used to indicate if the patient was persistent or not, binary variable)
- Demographics (categorical variables)
- Provider attributes (categorical variables)
- Clinical, factors (categorical variables)
- Disease/Treatment factors (categorical variables)

Github Repo link: https://github.com/YunxinG107112/VC/tree/main/week9

Data cleansing and transformation done on the data

- Use two different approaches to deal with unknown values in categorical columns, for example, for the 'Unkown' values in the Ntm_Speciality and Risk_Segment_During_Rx columns, we applied imputation based on the most frequent values of them.
- All categorical columns were encoded using Label Encoder, for example the column Gender was encoded so that it has only two values 0 and 1 to represent male and female.