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.

Business Understanding:

The act of continuing the treatment for the prescribed duration is called persistence. The success of any treatment is defined by taking proper medicines on time for prescribed time. The failure of this would lead to the adverse health effects and increased healthcare expenditure. In order to understand the persistency, the pharmaceutical companies would need help with automating the identification process of the persistent patients.

Data Understanding:

The data available in excel sheet with 69 columns and 3424 rows with a data type of each column as object. There are only 2 columns with numerical data, but the rest are available as categorical data. This was converted back to numerical by using the label encoding technique.

Null Values: There are no null values present in the data, hence it does not require any special treatment for that.

Skewness: The gender data is highly skewed towards the Male. The race data is majorly for cacussians. Ethnicity towards non-Hispanic. The age group >75 data is available more when compared to other groups. Likewise, almost all the groups have the data skewed.

Outliers: The data in the range of 25 to 75 are considered as the normal data, and the rest of the others is considered as outliers and is removed. Ie, Datapoints outside 25% and 75% Quarters are outliers and will be removed.

Data Cleansing and transformation:

The data available in excel sheet with 69 columns and 3424 rows with a data type of each column as object. There are only 2 columns with numerical data, the rest are available as categorical data. This was converted back to numerical by using the label encoding technique.

There are no null values present in the data, hence it does not require any special treatment for that. The gender data is highly skewed towards the Male. The race data is majorly for cacussians. Ethnicity towards non-Hispanic. The age group >75 data is available more when compared to other groups. Likewise, almost all the groups have the data skewed. The data in the range of 25 to 75 are considered as the normal data, and the rest of the others is considered as outliers and is removed. Ie, Datapoints outside 25% and 75% Quarters are outliers and will be removed. After that the data has been reduced to the shape 2956 rows × 69 columns.

The different methods for removing the outliers are explained in the notebook and the selected method is applied on the numerical data. Also, 2 different methods to convert the data from categorical to numerical is also explained (factorize and encoding).

EDA:

The 68 input variables need 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 data frame to 28.

Then the categorical variables are converted to numerical variable by applying the factorize method. Separated the x and y variables and divided it to test and train. Then applied 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.

EDA Presentation & Recommendation:

Summary

• From the Exploratory Data Analysis done, we are able to find how the different features/variables affects drug persistency.

- Dropped the unwanted columns

- The categorical variables are converted to numerical variables

- A total of 6 principal components can contribute a 46.9% of variance.

Recommendation

For the purpose of automating the process of drug persistency identification, the following machine learning models can be used:

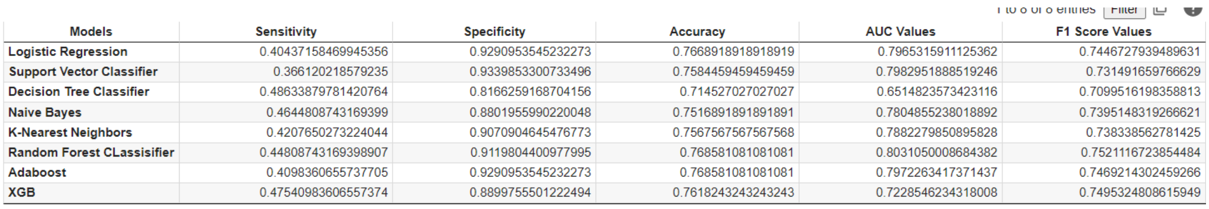
- **Logistic regression** – It is a type of linear model that is used for binary classification. It predicts output which is a categorical dependent variable. Such predictions are like yes or no, A or B, etc.

- **Random Forest** – It is a type of Bagging Ensemble Learning Classification. It predicts by the averaging of a number of decision tree classifiers.

- **Adaboost and XGBoost** – It is a type of Boosting model. It converts weak learners to strong learners.

Model:

Various machine learning models has been built and compared and a best model has been selected. The models used are Logistic Regression (Linear Model), Support Vector Classifier, Naïve Bayes, K-Nearest Neighbors, Decision Tree, Random Forest(Ensemble Learning- Bagging) , Adaboost(Boosting) and XGB(Boosting). The comparison matrix for various models is as shown below.


1 to 8 of 8 entries
Filter

Models	Sensitivity	Specificity	Accuracy	AUC Values	F1 Score Values
Logistic Regression	0.40437158469945356	0.9290953545232273	0.7668918918918919	0.7965315911125362	0.7446727939489631
Support Vector Classifier	0.366120218579235	0.9339853300733496	0.7584459459459459	0.7982951888519246	0.731491659766629
Decision Tree Classifier	0.48633879781420764	0.8166259168704156	0.714527027027027	0.6514823573423116	0.7099516198358813
Naive Bayes	0.4644808743169399	0.8801955990220048	0.7516891891891891	0.7804855238018892	0.7395148319266621
K-Nearest Neighbors	0.4207650273224044	0.9070904645476773	0.7567567567567568	0.7882279850895828	0.738338562781425
Random Forest CLassisifier	0.44808743169398907	0.9119804400977995	0.768581081081081	0.8031050008684382	0.7521116723854484
Adaboost	0.4098360655737705	0.9290953545232273	0.768581081081081	0.7972263417371437	0.7469214302459266
XGB	0.47540983606557374	0.8899755501222494	0.7618243243243243	0.7228546234318008	0.7495324808615949

From these values, Random Forest Classifier has a better matrix. Hence the final selected model is Random Forest Classifier.A prediction has also been done on the final selected model using the prediction system.

Conclusion:

From various Classification models, Random Forest Classifier was chosen based on the various parameter values. It has the highest accuracy of 76.9%. This could be the efficient model for the automation of the prediction of persistent or non-persistent drugs.