**Week 02 Assignment**

Supervised Learning Methods

Submitted by,

Group 06: Shruti Kamble, Shivani Jagtap,

Sai Prashanth Reddy, Pratik Mantri

College of Prof. Studies, Analytics

**A picture containing indoor

Description automatically generated**

EAI 6000: Fundamentals of Artificial Intelligence

Prof. Kasun Samarasinghe

Assignment Due Date: November 14, 2021

**Introduction:**

This dataset contains information about drug classification based on patient general information and its diagnosis. Machine learning model is needed to predict the outcome of the drug-type that will be suitable for the patient. There are 6 variables in this data set 4 categorical variables, and 2 continuous variables and 200 records. The aim is to perform data exploration, visualization and build Machine Learning models that will help in predicting drug type. The data has variables like Age, Sex, Blood Pressure Level, Cholesterol Level, Sodium to potassium Ratio in Blood and Drug Type.

Table

Description automatically generated

We have imported all the necessary libraries to perform data analysis. After importing the data, we verified the data using head() operator. The next step is to perform exploratory data analysis, hence, we first checked for null values in the data set. From the result, we observed there are no null/missing values in the dataset. The data is fully prepared exploratory data analysis.

**Data Exploration:**

This section will explore raw dataset that has been imported. We tried to analyze the data using one column at a time. We first considered ‘Drug’ column, it can be seen that, DrugY has more amount than other types of drugs. The stat shows 45.5% population uses DrugY. Next, we checked for the rest of the columns, distribution is normal. Further, we moved to numerical columns which is Age and Sodium to potassium Ratio in Blood.

We used describe () operator to get mean, count, std, min, max and others using describe function. The skewness value for each of the numerical variable is also shown in this section. Of the 200 records available in the data, the mean and median for their ages are 44.315 and 45 respectively with a standard deviation of 16.544315.

Table

Description automatically generated

While the lowest and highest ages are 15 and 74 respectively, the 25th percentile and 75th percentile ages are 31 and 58 respectively. Whereas in regard to Na\_to\_K values among these records, mean and median are 16.084 and 13.9365 respectively with a standard deviation of 7.223956. While the lowest and highest values are 6.269 and 38.247 respectively, the 25th percentile and 75th percentile ages are 10.44 and 19.38 respectively.

We calculated Age skewness and Na to K skewness, the result shows below positive skew which means the tail on the right side is longer than on the left. The distribution of 'Age' column is symetric, since the skewness value between -0.5 and 0.5. The distribution of 'Na\_to\_K' column is moderately skewed, since the skewness value is between 0.5 and 1. It can also be seen from the histogram for 'Na\_to\_K' column

Chart, histogram

Description automatically generated

Also, we plotted the graph for analyzing the other parameters of the dataset. From the distribution data in terms of gender, Blood Pressure, Cholesterol is shown as below:

Chart, bar chart

Description automatically generated

Considering the visuals above, and also the stats shows that cholesterol level goes above the rating of 100 and same is the case with Blood Pressure. The stat explains 51.5% of population from the dataset has high cholesterol level. Also, the blood pressure level is high among 38.5% of population.

Chart, bar chart

Description automatically generated

From the above visual, the consumption of DrugY is high among both the gender. And the second highest consumption by both the gender is DrugX.

Chart, bar chart

Description automatically generated

The above graph is about the parameter blood Pressure based on Cholesterol level.

Chart, scatter chart

Description automatically generated

The above visual is about

**2. Pre-processing of the data - standardize, normalize, one-hot-encoding, binning etc.**

**Pre-proceing of the data:** From below e can say that there Is one int one float and 4 object and null value can say that we are ready to implement machine learning algorithm.

**Graphical user interface, text

Description automatically generated**

Graphical user interface, text, application, email

Description automatically generated

E can ay that:

* DrugY has more amount than other types of drugs
* The distribution of patient gender is balanced.
* The distribution of blood pressure level is balanced.
* The distribution of cholesterol level is balanced.

DATA BINNING:

Data binning is a type of data preprocessing, a mechanism which includes also dealing with missing values, formatting, normalization and standardization. Binning can be applied to convert numeric values to categorical or to sample numeric values.

Graphical user interface, text, application

Description automatically generated

One hot encoding:

**one-hot encoding**, which is **transforming categorical variables into a form that could be provided to ML algorithms to do a better prediction**.

Table

Description automatically generated

Here we have transformed the categorical variable which is ex BP\_high to a form that could be provided to ML algorithm to do better prediction.

**3. Apply at least 4 ML models (Eg: Decision Tree, Random Forest, XgBoost, Linear/ Logistic Regression, SVM etc.)**

1. **Logistic Regression:**

Linear Regression is implemented to find the relationship and predict data values of two variables. As we have substituted the data into Train and test we performed the Logistic Regression on those data, Where we can clearly see into the image bellow that we get accuracy of 82.86%, Which is pretty good accuracy rate.

Table

Description automatically generated

1. **K Neighbours:**

K nearest neighbors is a simple algorithm that stores all available cases and predict the numerical target based on a similarity measure. KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood. The size of the neighbourhood needs to be set by the analyst or can be chosen using cross-validation (we will see this later) to select the size that minimises the mean-squared error.

If we look at the below screen shot we can see that the accuracy rate of KNN is 71.**43%**

A picture containing text, receipt

Description automatically generated

With the help of the K Neighbors we decided to plot the graph and to get the KNN Account Maximum percentage which gets up to 75.71%.

Chart, line chart

Description automatically generated

1. **Support Vector Machine(SVM)**

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences.

The Accuracy for Support Vector Machine is 84.29%

Table

Description automatically generated

1. **Naïve Bayes**

* **Categorical NB:**

The categorical Naive Bayes classifier is suitable for classification with discrete features that are categorically distributed. The categories of each feature are drawn from a categorical distribution. Parameters alphafloat, default=1.0. Additive (Laplace/Lidstone) smoothing parameter (0 for no smoothing). The Accuracy for Categorical Naïve Bayes is 82.86%

Table

Description automatically generated

* **Gaussian NB:**

Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. The Accuracy for Gaussian Naïve Bayes is 82.86%

A picture containing text, receipt

Description automatically generated

1. **Decision Tree**

A tree diagram which is used for making decisions in business or computer programming and in which the branches represent choices with associated risks, costs, results, or probabilities.

Table

Description automatically generated

Chart, line chart

Description automatically generated

1. **Random Forest Model:**

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

Table

Description automatically generated

Chart, line chart

Description automatically generated

**4. Performance measurement**

Comparing all the seven models which we performed on the dataset we compare all of them with each other to get a correct outcome of the results,

Table

Description automatically generated

By looking at the above from the results, it can be seen that most of ML models can reach **up to 80% accuracy** in predicting classification of drug type.

Table

Description automatically generated

**5. Conclusion:**

To conclude this assignment, by examining the dataset using the exploratory data analysis using python. After that, we have checked the data for the cleaning process and prepared the data for the modeling. We need to focus on the drug classification based on patient general information and its diagnosis, and which factors are affected by that. So, we have performed the Logistic regression model , Random Forest, K Neighbors, SVM, Naïve Bayes, Decision Tree to get the clarity through our models.

**Reference:**

Duca, A. L. (2021, April 1). *Data Preprocessing with Python Pandas — Part 5 Binning*. Medium. <https://towardsdatascience.com/data-preprocessing-with-python-pandas-part-5-binning-c5bd5fd1b950>

*Numpy, Matplotlib & Scipy Tutorial: Binning Data with Python and with Pandas*. (n.d.). Python-Course.eu. Retrieved November 15, 2021, from <https://python-course.eu/pandas_python_binning.php>

‌

‌