**19AI704 Applied Predictive Analytics**

**Evaluation Lab: Logistic Regression**

User Database (pn123.csv) : The dataset contains information of users **from** **a** company’s database. It contains information about

UserID, Gender, Age, EstimatedSalary, and Purchase. We are usingthis dataset **for** predicting that a user will purchase the company’s newly launched product **or** **not**.

Now, to predict whether a user will purchase the product or not,

one needs to find out the relationship between Age and Estimated

Salary. Here User ID and Gender are not important factors for

finding out this.

Split the dataset to train and test. 75% of data is used for train-ing the model and 25% of it is used to test the performance of

the model.

NB: Use sklearn.model\_selection and import train\_test\_split

Perform feature scaling here because Age and Estimated Salary

values lie in different ranges. If we don’t scale the features theEstimated Salary feature will dominate Age feature when the model finds the nearest neighbor to a data point in data space. Scale inthe range -1 to 1, such that, each feature will contribute equallyin decision making i.e. in finalizing the hypothesis.

Finally,train a Logistic Regression model, and analyse performancemeasures, accuracy and confusion matrix. Draw a graph to visualizeyour results–and show that your model performance is satisfactory.

Dr PN Kumar