Group Mini Project | Ensemble Techniques

Please find below the Mini Project for Ensemble Techniques. Please note this is a group assessment. Kindly submit your post before the deadline.

Parkinson’s Disease (PD) is a degenerative neurological disorder marked by decreased dopamine levels in the brain. It manifests itself through a deterioration of movement, including the presence of tremors and stiffness. There is commonly a marked effect on speech, including dysarthria (difficulty articulating sounds), hypophonia (lowered volume), and monotone (reduced pitch range). Additionally, cognitive impairments and changes in mood can occur, and risk of dementia is increased.

Traditional diagnosis of Parkinson’s Disease involves a clinician taking a neurological history of the patient and observing motor skills in various situations. Since there is no definitive laboratory test to diagnose PD, diagnosis is often difficult, particularly in the early stages when motor effects are not yet severe. Monitoring progression of the disease over time requires repeated clinic visits by the patient. An effective screening process, particularly one that doesn’t require a clinic visit, would be beneficial. Since PD patients exhibit characteristic vocal features, voice recordings are a useful and noninvasive tool for diagnosis. If machine learning algorithms could be applied to a voice recording dataset to accurately diagnosis PD, this would be an effective screening step prior to an appointment with a clinician.

The data for this project is available at [https://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/parkinsons.dataLinks to an external site.](https://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/parkinsons.data). The data consists of those diagnosed with Parkinson Disease and those who do not.

Study the data

1. Read the column description and ensure you understand each attribute well
2. Split the data into training and test set in the ratio of 70:30 respectively
3. Use the classification algorithms and compare the models to find the best model

# Lab | Ensemble Techniques

Please find below the Lab for Ensemble Techniques course. Kindly make your submission before the deadline.

**Data Set -**

[**pima-indians-diabetes.data**](https://olympus.greatlearning.in/courses/2181/files/175763/download?wrap=1)**[View in a new window](https://olympus.greatlearning.in/courses/2181/files/175763/download?wrap=1)**

**ColumnNames Information -**[**https://www.kaggle.com/uciml/pima-indians-diabetes-database/dataLinks to an external site.**](https://www.kaggle.com/uciml/pima-indians-diabetes-database/data)

**Problem Set**

1. Load the PIMA Indian Diabetes file into Python DataFrame. You may first download it to a local folder and then load it into Python dataframe. Let us assume the data frame is named  “pima\_df.
2. It is always a good practice to eye-ball raw data to get a feel of the data in terms of number of structure of the file, number of attributes, types of attributes and a general idea of likely challenges in the dataset. You would notice that it is a comma separated file. There are no column names!. Check the associated folders and find out about each attribute the name. What information is available about the data.
3. Using univariate analysis check the individual attributes for their basic statistic such as central values, spread, tails etc. What are your observations (any two attributes)?
4. Using bivariate analysis check for useful relationships between attributes and attributes and the target class.
5. Split the pima\_df into training and test set in the ratio of 70:30 (Training:Test).
6. Create the model using “entropy” method of reducing the entropy and fit it to training data.
7. Test the model on test data and what is the accuracy achieved. Capture the predicted values and do a crosstab.
8. Would you get the same result if you recreate the training and test data using random function?
9. Use regularization parameters of max\_depth, min\_sample\_leaf to recreate the model. What is the impact on the model accuracy. How does regularization help?
10. Next implement the decision tree using Random Forest. What is the optimal number of trees that gives the best result?https://ssl.gstatic.com/ui/v1/icons/mail/images/cleardot.gif

# Supervised Learning

Please find below the subjective quiz for Supervised Learning. Kindly submit it before the deadline.

1. Load the file “imports-85.data” into a dataframe from [https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.dataLinks to an external site.](https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.data) and column names of this data set can be found here --[https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.namesLinks to an external site.](https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.names)
2. Explain the problem statement. What are you predicting and what attributes you have to predict?
3. Browse a sample record from the dataframe. Are there any missing values?
4. How many records are available in the data set and how many attributes. Do you think the depth (number of records) is sufficient given the breadth? In other words, is the sample likely to be a good representative of the universe?
5. Analyse the data distribution for the various attributes and share your observations.
6. Are there any independent attributes which have |R| close to 1?
7. Which attributes seem to have stronger relation with the dependent variable (Price of the car)?
8. Given the above analysis, which algorithm is likely to give a better accuracy? Why?

**Additional Pointers:**

* Hint for Q6 and Q7 - "*use the corr method (correlation) defined for pandas dataframe to answer these 2 questions after dealing with missing values."*

# Lab | Supervised Learning

Please find below the Lab for Supervised Learning course. Kindly make your submission before the deadline.

**Data Set** - [CarInsurance-1.csvView in a new window](https://olympus.greatlearning.in/courses/2180/files/164240/download?wrap=1)

'CarInsurance.csv' data set includes 85 predictors that measure demographic characteristics for over 5800 individuals.

**Data description**- Column descriptions can be found here: [http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/data.html (Links to an external site.)Links to an external site.](http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/data.html)

The response variance is 'purchase'. This indicates that whether or not a given individual purchases a car insurance policy.

**The following tasks are expected to be done:**

1. Separate given ‘carInsurance.csv’ data set into train and test sets ( 70% train and 30% test) ( 1 mark)
2. Find whether a given induvial will buy a car insurance policy or not by using   ( find test data set accuracy)    
   (8 marks)

2.1. 'Logistic regression'

2.2  'Naive Bayes'

2.3. 'KNN'

2.4. Support vector machines

1. Compare performances of the above three algorithms  ( on test data set)

# Mini Project | Supervised Learning

Please find below the Mini Project for Supervised Learning course. Kindly make your submission before the deadline.

**Case Study-**

**Campaign for selling personal loans**

This case is about a bank (Thera Bank) which has a growing customer base. Majority of these customers are liability customers (depositors) with varying size of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget.

The department wants to build a model that will help them identify the potential customers who have higher probability of purchasing the loan. This will increase the success ratio while at the same time reduce the cost of the campaign.

The file given below contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

Data Set - [Bank\_Personal\_Loan\_Modelling.xlsxView in a new window](https://olympus.greatlearning.in/courses/2180/files/164241/download?wrap=1)

**Considering the information provided above, follow the steps given below --**

1. Read the column description and ensure you understand each attribute well.
2. Study the data distribution in each attribute, share your findings.
3. Get the target column distribution. Your comments.
4. Split the data into training and test set in the ratio of 70:30 respectively.
5. Use a classification model to predict the likelihood of a liability customer buying personal loans.
6. Explain why you chose one model over the other.