# Topic: Ensemble Techniques

**Instructions**

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: Nukala Ayyappa Bharthwaj**

**Batch Id: DSWDMCOS 21012022**

**Topic: Ensemble Techniques.**

1. **Business Problem**
   1. **Objective**
   2. **Constraints (if any)**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.**

**Using R and Python codes perform:**

1. **Data Pre-processing**

**3.1 Data Cleaning, Feature Engineering, etc.**

**3.2 Outlier Imputation**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary**
   2. **Univariate analysis**
   3. **Bivariate analysis**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options)**
   2. **Perform Bagging, Boosting, Voting, Stacking on given datasets**
   3. **Train and Test the data, use grid search cross validation, compare accuracies using confusion matrix**
   4. **Briefly explain the model output in the documentation**

****

1. **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided**

Problem Statement: -

**Instructions**

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

1. **Business Problem**
   1. **Objective**
   2. **Constraints (if any)**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.**

**Using R and Python codes perform:**

1. **Data Pre-processing**

**3.1 Data Cleaning, Feature Engineering, etc.**

**3.2 Outlier Imputation**

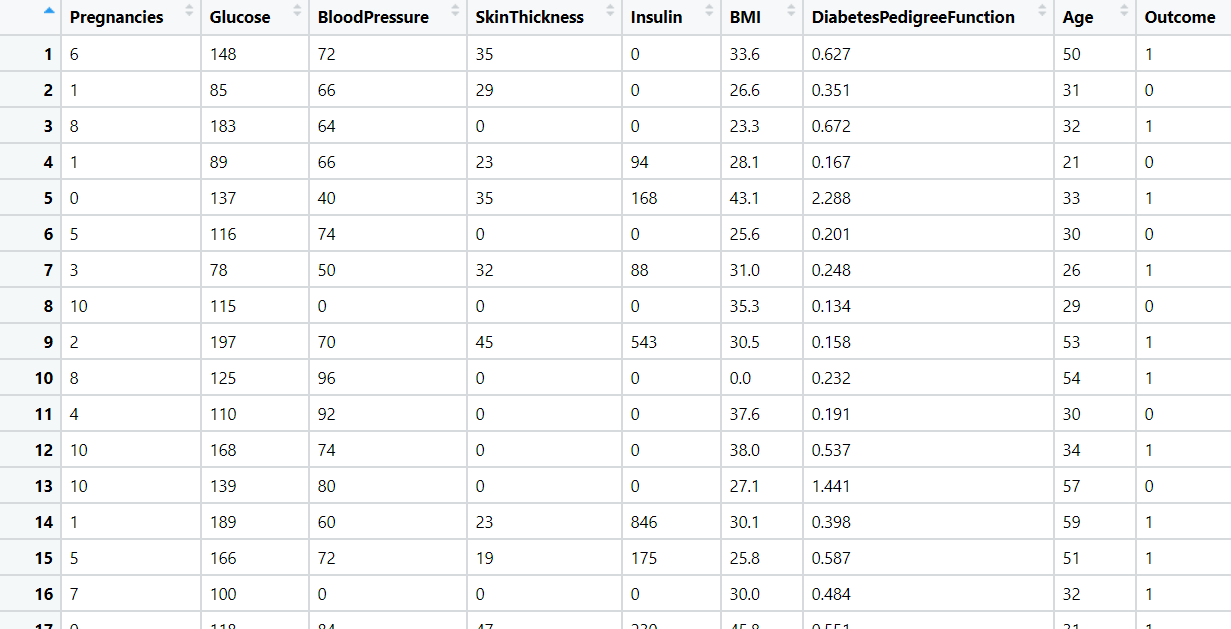
1. **Exploratory Data Analysis (EDA):**
   1. **Summary**
   2. **Univariate analysis**
   3. **Bivariate analysis**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options)**
   2. **Perform Bagging Boosting (adaboost, fastadaboost, Xgboost), Stacking, Voting on the given datasets in Hands on Material**
   3. **Train and Test the data and compare accuracies by Confusion Matrix and use different Hyper Parameters and also use GridSearchCV to improve your model performance**
   4. **Briefly explain the model output in the documentation**

****

1. **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided**

**Problem Statement: -**

Diabetes is disease caused by increase in Blood glucose levels in your body, Blood glucose is the main source of energy and it gets from the food you eat. Insulin a hormone, made by the pancreas helps to get glucose from your food reach every cell of your body for energy. Sometimes, your body

does not make enough or use Insulin at all, due to this the glucose levels in your body increases which leads to severe health problems and moreover diabetes has no cure, you can only avoid taking sugar filled foods and take precautions. In Pregnant women’s the rising of glucose levels is a danger for to be mother and the baby, if we can predict accurately whether a pregnant women will become diabetic or not can be help doctors to treat patients in a much efficient way and also for pregnant women can avoid becoming diabetic by taking necessary steps during pregnancy. Build an ensemble model to correctly classify the outcome variable and improve your model prediction by using GridSearchCV. You must apply Bagging, Boosting, Stacking and Voting on the dataset.

**Sol:**

**Business Objective:** To predict the diabetes disease for the new patient using the ensemble techniques.

**Constraints:** Lack of analysis of the patient’s data.

**Data Types:** given data and its data types are shown below

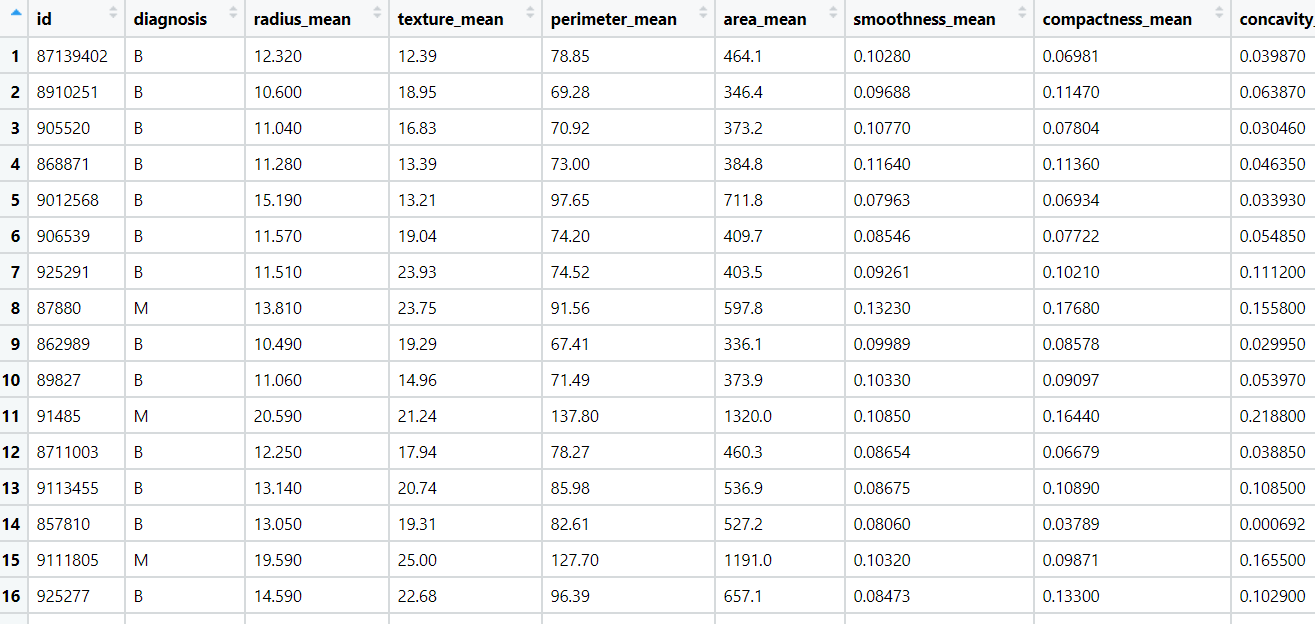
|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Data type | Relevance |
| Number of times pregnant | Number of time the woman got pregnancy | Internal | Relevant |
| Plasma glucose concentration | Plasma glucose concentration a 2 hours in an oral glucose tolerance test | Ratio | Relevant |
| Diastolic blood pressure | Diastolic blood pressure (mm Hg) | Ratio | Relevant |
| Triceps skin fold thickness | Triceps skin fold thickness (mm) | Ratio | Relevant |
| 2-Hour serum insulin | 2-Hour serum insulin (mu U/ml) | Ratio | Relevant |
| Body mass index | Body mass index (weight in kg/(height in m)^2) | Ratio | Relevant |
| Diabetes pedigree function | Diabetes pedigree function | Ratio | Relevant |
| Age (years) | Age of patient | Ratio | Relevant |
| Classvariable | Variable for patiest having diabetes or not | Nominal | Relevant |

**Data Pre-Processing:** the complete data can be used for preparing the models of Ensemble Techniques.

**Ensemble Models:** I have segregated the data into training and testing data set as 70% and 30% of the total data respectively. I have applied all the ensemble techniques of voting, bagging, boosting and stacking in all the techniques I have got the accuracy of approximately 75% for all the techniques.

**Problem Statement: -**

Most cancers form a lump called as tumour or a growth. But not all lumps are cancer. Doctors take out a piece of the lump and look at it to find out if it’s cancer. Lumps that are not cancer are called benign (be-NINE). Lumps that are cancer are called malignant (muh-LIG-nunt). There are some cancers, like leukemia (cancer of the blood), that don’t form tumour. They grow in the blood cells or other cells of the body. For instance, If a doctor tends to wrongly diagnose a benine tumour as a malignant tumour can a cause a overwhelming anxiety in patient which can lead to depression or much worse, a wrong diagnosis is a major problem in our health care sector, to improve their analysis build an ensemble model on the dataset which can accurately classify benine and Malignant tumours on the dataset given. Perform Bagging, Boosting, Stacking, Voting algorithm and provide your insights in the documentation.





**A screenshot of a cell phone

Description automatically generatedSol:**

**Business Objective:** To predict the cancer disease for the new patient using the ensemble techniques.

**Constraints:** Lack of analysis of the patient’s data.

**Data Types:** the complete data is about the cancer patients and it effects and measure in the body.

**Data Pre-Processing:** I have identified that id could not be used for the analysis on the data so I have dropped the column and remaining all the columns are used for the analysis.

**Ensemble Models:** I have segregated the data into training and testing data set as 70% and 30% of the total data respectively. I have applied all the ensemble techniques of voting, bagging, boosting and stacking in all the techniques I have got the accuracy of approximately 73% for all the technique’s.

**Problem Statement: -**

A sample of global companies and their ratings are given for the cocoa bean production along with the location of bean being used by the companies. Identify the important features in the analysis and accurately classify the companies based on their ratings and draw insights from your model. Perform Ensemble methodology such as Bagging, Boosting, Stacking, voting algorithms on the dataset given.

**A screenshot of a computer

Description automatically generated**

**Sol:**

**Business Objective:** Segregating the companies data based on the ratings.

**Constraints:** Lack of analysis of the companie’s data.

**Data Types:** given data and its data types are shown below

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Data type | Relevance |
| Company | Name of the comany | Nominal | Relevant |
| Name | Name of the person | Nominal | Relevant |
| REF | Reference id | Nominal | Relevant |
| Review | Review year | Ordinal | Relevant |
| Cocoa\_Percent | Cocoa percentage in the product | Ratio | Relevant |
| Company\_Location | Location of the company | Nominal | Relevant |
| Ratings | Ratings given by the company | Ratio | Relevant |
| Bean\_Type | Type of bean used | Nominal | Relevant |
| Origin | Origin of the company | Nominal | Relevant |

**Data Pre-Processing:** I have identified the categorical data in the data given for which I have converted numeric data by using factor function in R and label encoding in Python. For this data to do the analysis I have segregated the data based on ratings for which I have considered the ratings as high if it is greater than 3 and others as less respectively.

**Ensemble Models:** I have segregated the data into training and testing data set as 70% and 30% of the total data respectively for doing the analysis. I have applied all the ensemble techniques of voting, bagging, boosting and stacking in all the techniques I have got the accuracy of approximately 70% for all the techniques.

**Problem Statement: -**

Data privacy has been and is always an important factor that websites are very critical about, to safe guard their customers details from ethical hackers and other unsolicited misuse of data, users are required to use alpha numeric characters while creating their account for the first time for password strength. Perform Ensemble technique to classify the user’s password strength of users, use Bagging, Boosting, Stacking, voting algorithms on the dataset given.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

**Sol:**

**Business Objective:** Segregating the passwords of the user as strong and weak.

**Constraints:** Lack of analysis of the password data.

**Data Types:** given data and its data types are shown below

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Data type | Relevance |
| Characters | Password of the user | Nominal | Relevant |
| Password strength | Password strength of the user | Nominal | Relevant |

**Data Pre-Processing:** I have segregated the password strength as strong and weak based on the characters used by the user and the final data is used for doing the analysis.

**Ensemble Models:** I have segregated the data into training and testing data set as 70% and 30% of the total data respectively for doing the analysis. I have applied all the ensemble techniques of voting, bagging, boosting and stacking in all the techniques I have got the accuracy of approximately 68% for all the techniques.