# Topic: AdaBoost- Extreme Gradient Boosting

**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: Nithin Dsouza**

**Batch Id:** 05012021-10AM

**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**

# Note:

**The assignment should be submitted in the following format:**

* **R code**
* **Python code**
* **Code Modularization should be maintained**
* **Documentation of the model building (elaborating on steps mentioned above)**

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:

**Name: Batch Id:**  **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 (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**

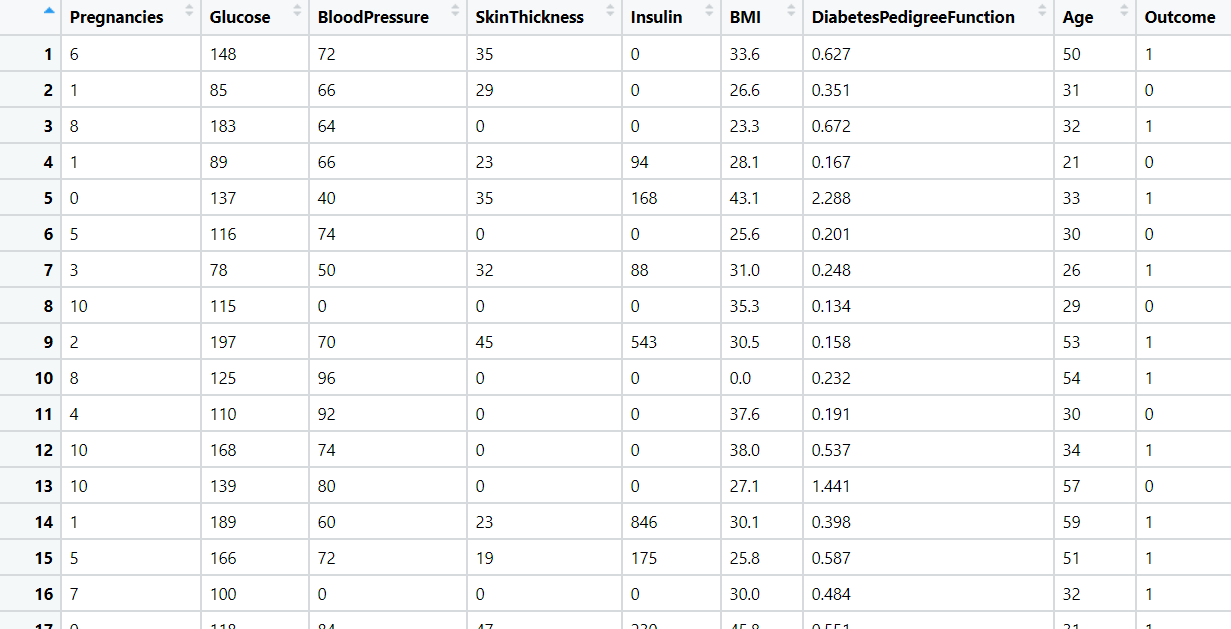
# Note:

**The assignment should be submitted in the following format:**

* **R code**
* **Python code**
* **Code Modularization should be maintained**
* **Documentation of the model building (elaborating on steps mentioned above)**

**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.

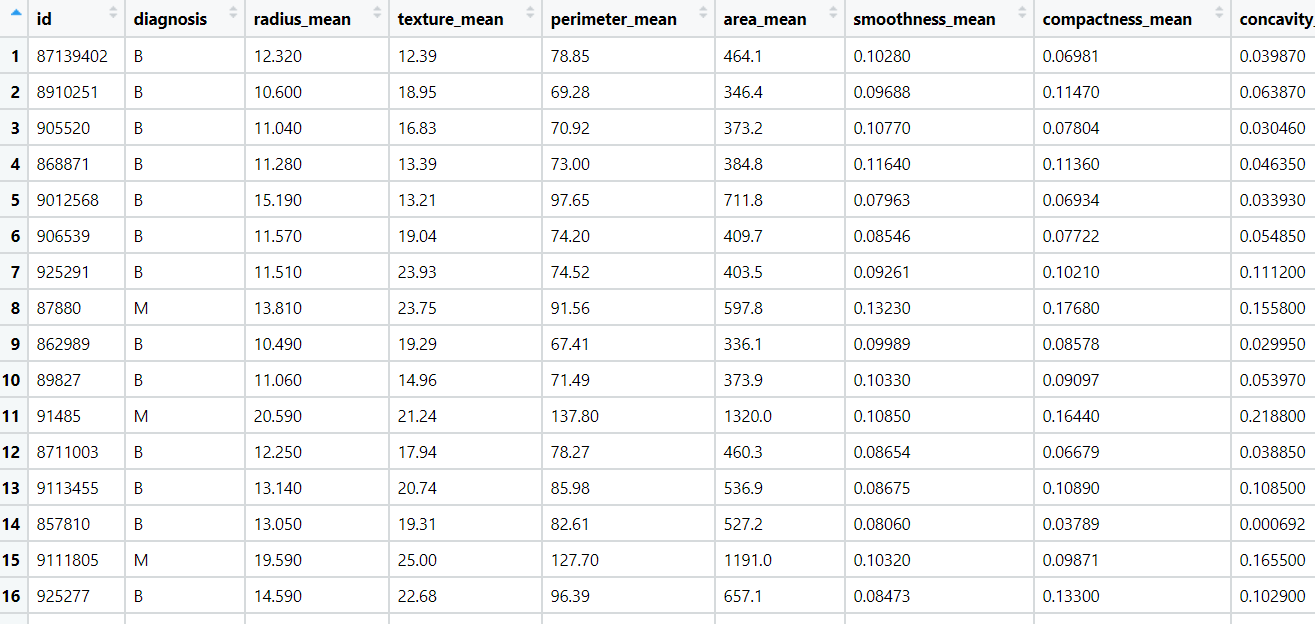


**Business Objective:** Maximize accuracy of target variable (Class variable).

**Business Constraint:** Minimize computational time.

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Type | Relevance |
| Number of times pregnant | Number of times pregnant | Ordinal | It is not important |
| Plasma glucose concentration | Glucose level | Ratio | it provides useful information |
| Diastolic blood pressure | Blood pressure | Interval | it provides useful information |
| Triceps skin fold thickness | Thickness of skin | Ratio | It is not important |
| 2-Hour serum insulin | Insulin | Ratio | Useful information |
| Body mass index | Mass index of body | Ratio | It is required |
| Diabetes pedigree function | Pedigree function | Ratio | It is required |
| Age (years) | Age of pregnant women | Ratio | it is not that useful information |
| Class variable | Boolean type | Ordinal | Targeted variable |

**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.



**Objective:** Maximize the accuracy for classification as Benign and malignant.

Minimize the error in classification.

**Constraints:** Computation Complexity involved.

|  |  |  |  |
| --- | --- | --- | --- |
| Name of the Feature | Description | Type | Relevance |
| id | Patient ID | Nominal | Irrelevant, does not provide useful information |
| diagnosis | Result of all features (B-Benign & M-Malignant) | Ordinal | Relevant, can be used as Response variable |
| radius\_mean | Radius Of Lump | Continuous | Relevant, Provide useful information |
| texture\_mean | Mean Texture of Lump | Continuous | Relevant, Provide useful information |
| perimeter\_mean | Mean Perimeter of Lump | Continuous | Relevant, Provide useful information |
| area\_mean | Mean area of lump | Continuous | Relevant, Provide useful information |
| smoothness\_mean | Mean smoothness of lump | Continuous | Relevant, Provide useful information |
| compactness\_mean | Mean compactness of lump | Continuous | Relevant, Provide useful information |
| concavity\_mean | Mean concavity of lump | Continuous | Relevant, Provide useful information |
| points\_mean | mean points of lump | Continuous | Relevant, Provide useful information |
| symmetry\_mean | mean symmetry of lump | Continuous | Relevant, Provide useful information |
| dimension\_mean | mean dimension of lump | Continuous | Relevant, Provide useful information |
| radius\_se | Standard error radius | Continuous | Relevant, Provide useful information |
| texture\_se | Standard error texture | Continuous | Relevant, Provide useful information |
| perimeter\_se | Standard error perimeter | Continuous | Relevant, Provide useful information |
| area\_se | Standard error area | Continuous | Relevant, Provide useful information |
| smoothness\_se | Standard error smoothness | Continuous | Relevant, Provide useful information |
| compactness\_se | Standard error compactness | Continuous | Relevant, Provide useful information |
| concavity\_se | Standard error concavity | Continuous | Relevant, Provide useful information |
| points\_se | Standard error points | Continuous | Relevant, Provide useful information |
| symmetry\_se | Standard error symmetry | Continuous | Relevant, Provide useful information |
| dimension\_se | Standard error dimension | Continuous | Relevant, Provide useful information |
| radius\_worst | worst radius measure | Continuous | Relevant, Provide useful information |
| texture\_worst | worst texture measure | Continuous | Relevant, Provide useful information |
| perimeter\_worst | worst perimeter measure | Continuous | Relevant, Provide useful information |
| area\_worst | worst area measure | Continuous | Relevant, Provide useful information |
| smoothness\_worst | worst smoothness measure | Continuous | Relevant, Provide useful information |
| compactness\_worst | worst compactness measure | Continuous | Relevant, Provide useful information |
| concavity\_worst | worst concavity measure | Continuous | Relevant, Provide useful information |
| points\_worst | worst points measure | Continuous | Relevant, Provide useful information |
| symmetry\_worst | worst symmetry measure | Continuous | Relevant, Provide useful information |
| dimension\_worst | worst dimension measure | Continuous | Relevant, Provide useful information |

**A screenshot of a cell phone

Description automatically generatedProblem 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**

**Objective:** Maximize the accuracy for classification.

Minimize the error in classification.

**Constraints:** Computation Complexity involved.

|  |  |  |  |
| --- | --- | --- | --- |
| Name of the Feature | Description | Type | Relevance |
| Company | Name of the company | Nominal | Irrelevant, does not provide much of information |
| Name | Company Location name | Nominal | Irrelevant, does not provide much of information |
| REF | Reference Number | Nominal | Irrelevant, does not provide much of information |
| Review | year of review | Ordinal | Relevant, can be used to categorize |
| Cocoa\_Percent | Percentage of cocoa present in bean production | Continuous, ratio | Relevant, it provides useful info |
| Company\_Location | Location of bean being used by company | Nominal | Irrelevant, does not provide much of information |
| Rating | Score for company out of 10 | Ordinal | Relevant, can be used as response variable |
| Bean\_Type | Name of bean type | Nominal | Irrelevant, does not provide much of information |
| Origin | Origin of raw bean imported | Nominal | Irrelevant, does not provide much of information |

**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 generated

A screenshot of a cell phone

Description automatically generated**Business Objective**: Maximize accuracy of target variable(character\_strenght) and minimize classification error.

**Business Constraint**: Minimize computational time.

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Type | Relevance |
| characters | Set of special characters in dataset | Nominal | Input variable |
| characters\_strength | Strength of characters | Ordinal | Target variable |