# Ensemble Techniques

**Instructions:**

A screenshot of a cell phone

Description automatically generatedPlease share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: Shahina Athar**

**Batch ID:** 10122020

**Topic: Ensemble Techniques**

**Grading Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline.
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline.

(OR)

* + Less than 30% of problems in the assignments are submitted before the deadline.
* **Grade F: (< 50):** No submission (or) malpractice.

1. **Business Problem**
   1. **What is the business objective?**
   2. **Are there any constraints?**
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. And if not relevant, provide reasons and a description of the feature.**

**Using R and Python codes perform:**

1. **Data Pre-processing**

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

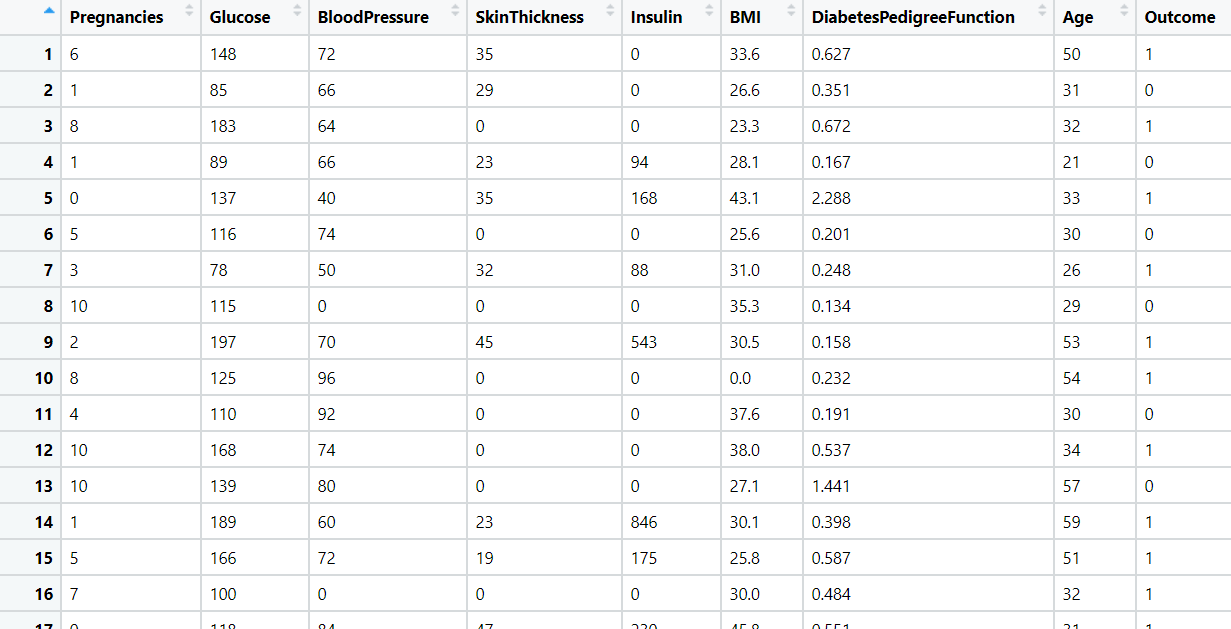
**3.2 Outlier Treatment.**

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.**
3. **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**
4. **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 model and compare accuracies by building a confusion matrix and use different hyperparameters. Also use GridSearchCV to improve your model performance.**
   4. **Briefly explain the model output in the documentation.**

1. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Problem Statements:**

1. Given is the diabetes dataset. 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. 



1. **Business Problem**
   1. **What is the business objective?**

**Ans:** Given is the diabetes dataset. 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.

* 1. **Briefly explain the model output in the documentation.**

**Ans:** First, we load the data and then split it into train and test dataset with (80-20), we apply "Bagging" method we get test accuracy 81% & train accuracy 100% then we apply "AdaBoostClassifier" we get test accuracy 79% and train accuracy 84% then we apply "GradientBoostingClassifier" we get test accuracy 80% and train accuracy 91% then we give Hyperparameters - to improve the accuracy we get test accuracy 79% and train accuracy 80% and then we apply "xgboost" we get test accuracy 77% and train accuracy 100% then we apply "GridSearchCV" to improve model prediction and we get 78% of test accuracy, and also we get "{'colsample\_bytree': 0.8, 'gamma': 0.3,'max\_depth': 9,

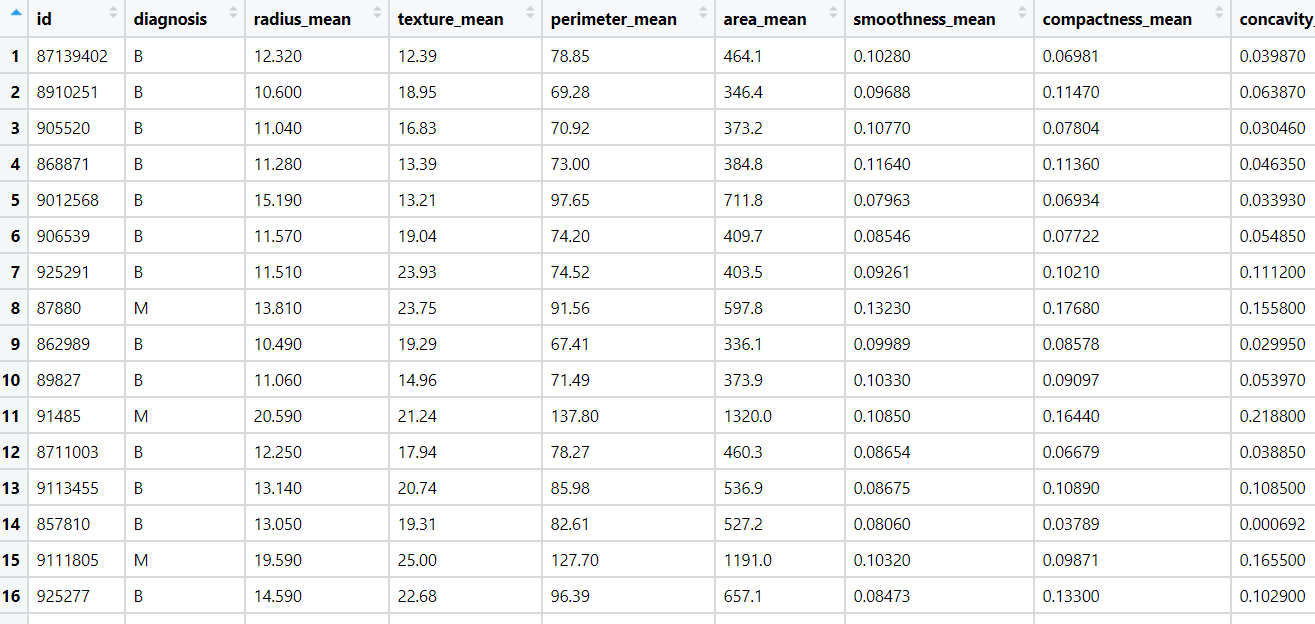
 'rag\_alpha': 0.01,'subsample': 0.9}" after that we apply "stacking", in stacking Our both individual models scores an accuracy of 68%, 76%, and our Stacked model got an accuracy of nearly 72%. By Combining two individual models we got a significant performance improvement. Then we apply "Voting" technique in that we get 68% of accuracy on hard voting and 72% of accuracy on Soft voting.

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

**Ans:** In this diabetes dataset, we Build an ensemble model to correctly classify the outcome variable (who suffers from diabetes or not) with different methods and improve our model prediction by using GridSearchCV. We apply Bagging, Boosting, Stacking, and Voting to the dataset.

1. Most cancers form a lump called a tumour. But not all lumps are cancerous. Doctors extract a sample from the lump and examine it to find out if it’s cancer or not. Lumps that are not cancerous are called benign (be-NINE). Lumps that are cancerous are called malignant (muh-LIG-nunt). Obtaining incorrect results (false positives and false negatives) especially in a medical condition such as cancer is dangerous. So, perform Bagging, Boosting, Stacking, and Voting algorithms to increase model performance and provide your insights in the documentation.

A screenshot of a cell phone

Description automatically generated

1. **Business Problem**
   1. **What is the business objective?**

**Ans:** Most cancers form a lump called a tumour. But not all lumps are cancerous. Doctors extract a sample from the lump and examine it to find out if it’s cancer or not. Lumps that are not cancerous are called benign (be-NINE). Lumps that are cancerous are called malignant (muh-LIG-nunt). Obtaining incorrect results (false positives and false negatives) especially in a medical condition such as cancer is dangerous. So, perform Bagging, Boosting, Stacking, and Voting algorithms to increase model performance and provide your insights in the documentation.

* 1. **Briefly explain the model output in the documentation.**

**Ans:** First, we load the data and then split it into train and test dataset with (80-20), we apply "Bagging" method we get test accuracy 95% & train accuracy 100% then we apply "AdaBoostClassifier" we get test accuracy 96% and train accuracy 100% then we apply "GradientBoostingClassifier" we get test accuracy 96% and train accuracy 100% then we give Hyperparameters - to improve the accuracy we get test accuracy 96% and train accuracy 96% and then we apply "xgboost" we get test accuracy 97% and train accuracy 100% then we apply "GridSearchCV" to improve model prediction and we get 96% of test accuracy, and also we get "{'colsample\_bytree': 0.8,'gamma': 0.3,'max\_depth': 7,'rag\_alpha': 0.01,'subsample': 0.9}" after that we apply "stacking", in stacking Our both individual models scores an accuracy of 95%, 95%, and our Stacked model got an accuracy of nearly 95%. By Combining two individual models we got a significant performance improvement. Then we apply "Voting" technique in that we get 95% of accuracy on hard voting and 96% of accuracy on Soft voting.

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

**Ans:** In this tumour dataset, we Build an ensemble model to correctly classify that if it’s cancer or not with different methods and improve our model prediction by using GridSearchCV. We apply Bagging, Boosting, Stacking, and Voting to the dataset.

A screenshot of a cell phone

Description automatically generated

1. A sample of global companies and their ratings are given for the cocoa bean production along with the location of the beans being used. Identify the important features in the analysis and accurately classify the companies based on their ratings and draw insights from the data. Build ensemble models such as Bagging, Boosting, Stacking, and Voting on the dataset given.

**A screenshot of a computer

Description automatically generated**

1. **Business Problem**
   1. **What is the business objective?**

**Ans:** A sample of global companies and their ratings are given for the cocoa bean production along with the location of the beans being used. Identify the important features in the analysis and accurately classify the companies based on their ratings and draw insights from the data. Build ensemble models such as Bagging, Boosting, Stacking, and Voting on the dataset given.

* 1. **Briefly explain the model output in the documentation.**

**Ans: This Problem can be solved as Regression & Classification, I solved in Both ways.so, here I explained with both ways 1st I explain with Regression Method:**

First, we load the data, remove unwanted columns and do labelencoing and then split it into train and test dataset with (80-20), we apply "Bagging" method we get test mean squared error is 0.183, test r2\_score is 0.118 & train mean squared error is 0.025, train r2\_score is 0.890,then we apply "AdaBoostClassifier" we get test mean squared error is 0.207, test r2\_score is 0.002 & train mean squared error is 4.35, train r2\_score is 0.999 then we apply "GradientBoostingClassifier" we get test mean squared error is 0.186, test r2\_score is 0.106 & train mean squared error is 0.143, train r2\_score is 0.999 and then we apply "xgboost" we get we get test mean squared error is 0.187, test r2\_score is 0.100 & train mean squared error is 0.145, train r2\_score is 0.377 then we apply "GridSearchCV" to improve model prediction and we get Residual sum of squares is 0.18, and also we get "{'colsample\_bytree': 0.8, 'gamma': 0.1,'max\_depth': 3,'rag\_alpha': 0.01, 'subsample': 0.8}" after that we apply "stacking", we get -0.195 score. Then we apply "Voting" technique in that we get [3.50614177 3.24197456 3.18824146 ... 3.35151055 3.33372764 3.22221436].

**2nd I explained with Classification Method:** First, we load the data, remove unwanted columns and crete dummy variables, make rating column into 0’s and 1’s and then split it into train and test dataset with (75-25), we apply "Voting" technique in that we get 58% of accuracy on hard voting and 100% of accuracy on Soft voting, after that we apply "stacking", in stacking Our individual models scores an accuracy - "lr(LogisticRegression) 62 % , "knn(KNeighborsClassifier) 60%, "cart(DecisionTreeClassifier) 63% ,"svm 58%,"naïve bayes 65%, and our Stacked model got an accuracy of nearly stacking 63%. By Combining two individual models we got a significant performance improvement. Then we apply "Bagging" method we get test accuracy 67% & train accuracy 96%, then we apply "AdaBoostClassifier" we get test accuracy 66% and train accuracy 80%, then we apply "GradientBoostingClassifier" we get test accuracy 60% and train accuracy 69%, and then we apply "xgboost" we get test accuracy 63% and train accuracy 85% then we apply "GridSearchCV" to improve model prediction and we get 63% of test accuracy, and also we get "{'colsample\_bytree': 0.8, 'gamma': 0.2, 'max\_depth': 3, 'rag\_alpha': 0.01, 'subsample': 0.8}"

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

**Ans:** In this Coca\_Rating dataset, we Build an ensemble model to correctly classify the companies on their ratings with different methods and improve our model prediction by using GridSearchCV. We apply Bagging, Boosting, Stacking, and Voting to the dataset

1. Data privacy is always an important factor to safeguard their customers' details. For this, password strength is an important metric to track. Build an ensemble model to classify the user’s password strength.A screenshot of a cell phone

   Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

1. **Business Problem**
   1. **What is the business objective?**

**Ans:** Data privacy is always an important factor to safeguard their customers' details. For this, password strength is an important metric to track. Build an ensemble model to classify the user’s password strength.

* 1. **Briefly explain the model output in the documentation.**

**Ans:** First, we load the data, create dummy variables, and then split it into train and test dataset, we apply "Voting" technique in that we get 85% of accuracy on hard voting and 85% of accuracy on Soft voting, after that we apply "stacking", in stacking Our individual models scores an accuracy - "lr(LogisticRegression) 85% , "knn(KNeighborsClassifier) 85%, "cart(DecisionTreeClassifier) 85% ,"svm 85%,"naïve bayes 14%, and our Stacked model got an accuracy of nearly stacking 85%. By Combining two individual models we got a significant performance improvement. Then we apply "Bagging" method we get test accuracy 85% & train accuracy 100%, then we apply "AdaBoostClassifier" we get test accuracy 85% and train accuracy 100%, then we apply "GradientBoostingClassifier" we get test accuracy 85% and train accuracy 85%, and then we apply "xgboost" we get test accuracy 85% and train accuracy 85% then we apply "GridSearchCV" to improve model prediction and we get 85% of test accuracy, and also we get "{'colsample\_bytree': 0.8, 'gamma': 0.2, 'max\_depth': 3, 'rag\_alpha': 0.01, 'subsample': 0.8}"

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

**Ans:** In this Password Strength dataset, we Build an ensemble model to correctly classify the the user’s password strength. with different methods and improve our model prediction by using GridSearchCV. We apply Bagging, Boosting, Stacking, and Voting to the dataset.