**Diabetes Classification**

# 1-Problem Statement

# **Our problem is about to diagnostically classify whether or not a patient has diabetes based on certain diagnostic measurements included in the dataset. The measurements are numerical features. They are Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age and Outcome : Class variable (0 or 1) .**

# **This dataset is 786 records originally from the National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the selection of these instances from a larger database. All patients here are females at least 21 years old.**

# 2-Related Work

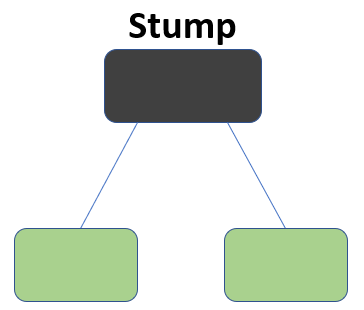
There is a couple of Related work listed below:

* <https://www.news-medical.net/health/What-is-Diabetes.aspx>
* <https://pubmed.ncbi.nlm.nih.gov/8422798/>
* <https://pubmed.ncbi.nlm.nih.gov/7988310/>
* <https://vitalflux.com/python-train-model-logistic-regression/>

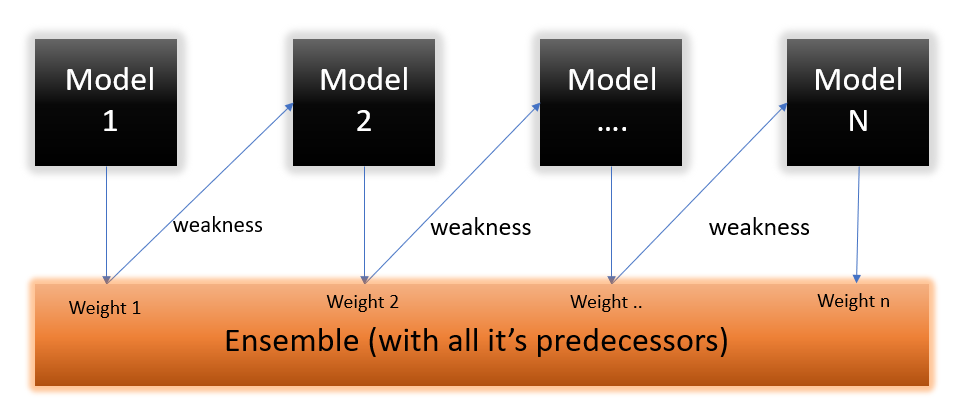
# 3-Model Architecture

## 3.1-AdaBoost

Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. These trees are also called Decision Stumps.



What this algorithm does is that it builds a model and gives equal weights to all the data points. It then assigns higher weights to points that are wrongly classified. Now all the points which have higher weights are given more importance in the next model. It will keep training models until and unless a lower error is received.



* Below are the steps for performing the AdaBoost algorithm:
  1. Initially, all observations are given equal weights.
  2. A model is built on a subset of data.
  3. Using this model, predictions are made on the whole dataset.
  4. Errors are calculated by comparing the predictions and actual values.
  5. While creating the next model, higher weights are given to the data points which were predicted incorrectly.
  6. Weights can be determined using the error value. For instance, the higher the error the more is the weight assigned to the observation.
  7. This process is repeated until the error function does not change, or the maximum limit of the number of estimators is reached.

## 3.2-Random Forest

Random forest is a ***Supervised Machine Learning Algorithm*** that is ***used widely in Classification and Regression problems***. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing ***continuous variables*** as in the case of regression and ***categorical variables*** as in the case of classification. It performs better results for classification problems.



**Steps involved in random forest algorithm:**

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on ***Majority Voting or Averaging***for Classification and regression respectively.

the random forest can generalize over the data in a better way. This randomized feature selection makes random forest much more accurate than a decision tree.

## 3.3-KNN

Yes, this is the principle behind K Nearest Neighbors. Here, nearest neighbors are those data points that have minimum distance in feature space from our new data point. And K is the number of such data points we consider in our implementation of the algorithm. Therefore, distance metric and K value are two important considerations while using the KNN algorithm. Euclidean distance is the most popular distance metric. You can also use Hamming distance, Manhattan distance, Minkowski distance as per your need. For predicting class/ continuous value for a new data point, it considers all the data points in the training dataset. Finds new data point’s ‘K’ Nearest Neighbors (Data points) from feature space and their class labels or continuous values.



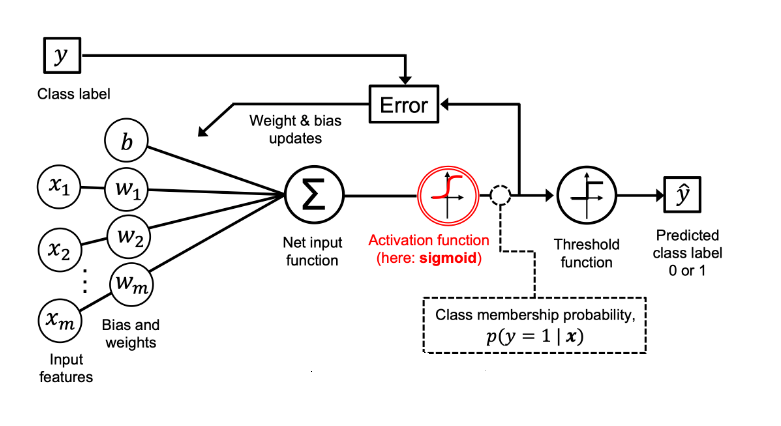
Then:

For classification: A class label assigned to the majority of K Nearest Neighbors from the training dataset is considered as a predicted class for the new data point.

For regression: Mean or median of continuous values assigned to K Nearest Neighbors from training dataset is a predicted continuous value for our new data point

## 3.4-Logistic Regression

Logistic Regression is a supervised machine learning Algorithm that is used widely in Classification and Regression problems. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. Logistic regression could be used to predict whether a person is diabetic or not. These binary outcomes allow straightforward decisions between two alternatives. Logistic regression has become an important tool in the discipline of machine learning. It allows algorithms used in machine learning applications to classify incoming data based on historical data. As additional relevant data comes in, the algorithms get better at predicting classifications within data sets.

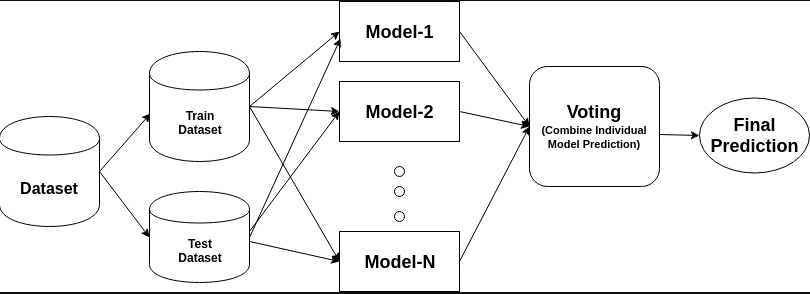


**Steps involved in Logistic Regression** **algorithm:**

* Data Pre-processing step.
* Fitting Logistic Regression to the Training set.
* Predicting the test result.
* Test accuracy of the result (Creation of Confusion matrix)
* Visualizing the test set result.

## 3.5-Ensemble Classifier

Ensemble learning is a way of generating various base classifiers from which a new classifier is derived which performs better than any constituent classifier. These base classifiers may differ in the algorithm used, hyperparameters, representation or the training set.



For every iteration,

1. A base model is created on each of these samplings.
2. The models run in parallel and are independent of each other.
3. The final predictions are determined by combining the predictions from all the models.

These models collectively form a higher graded model to produce more accuracy. The final model is averaged by:

e= (**Σ**eᵢ)/n

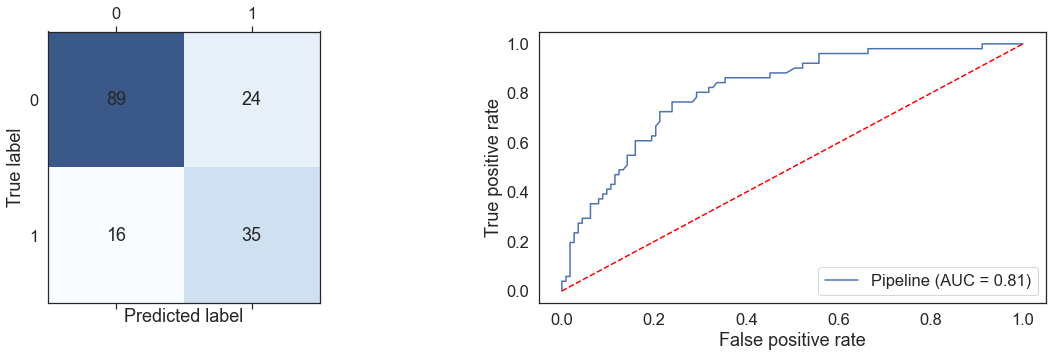
where e₁,e₂…..eₙ = base classifier

e = final classifier

# 4-Evaluation & Results

## 4.1-AdaBoost

Recall: 0.69



## 4.2- Random Forest

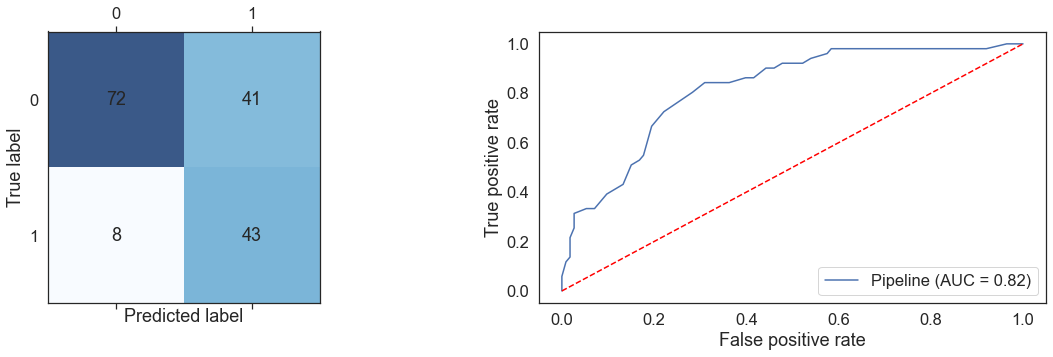
Recall: 0.73

Chart, scatter chart

Description automatically generated

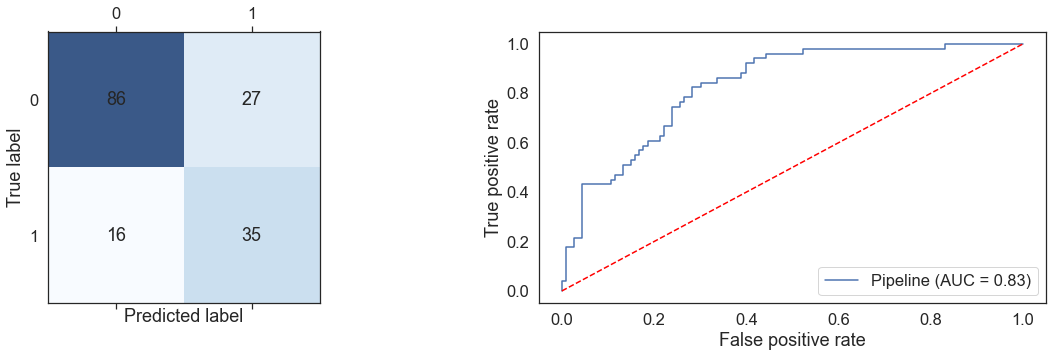
## 4.3-KNN

Recall: 0.84



## 4.4-Logistic Regression

Recall: 0.69



## 4.5-Ensemble Classifier

Recall: 0.76

Chart, line chart

Description automatically generated

## 4.6- The best Model

The best model is KNN because KNN has the highest Recall among all the models.

Chart, line chart, histogram

Description automatically generated

# 

# 5-References

<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

<https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5>