PROJECT DESIGN PHASE - I

AI BASED DIABETES PREDICTION SYSTEM

ARTIFICIAL INTELLIGENCE:

Artificial Intelligence(AI) is the intelligence of machines or software, as opposed to the intelligence of humans or animals. It is also the field of study in computer science that develops and studies intelligent Machines. “AI” may also refer to the machines themselves. AI technology is Widely used throughout industry, government and science. Some high-profile applications are advances web search engines, recommendation systems, understanding human speech, self-driving cars etc., The various sub-fields of AI research are centered around particular goals and the use of particular tools. The traditional goals of AI research include reasoning, knowledge representation, planning, learning, natural language processing, perception and support for robotics. There are thousands of successful AI applications used to solve specific problems for specific industries or institutions. In a 2017 survey, one in five companies reported they had incorporated “ AI “ in some offering or processes. A few examples are energy storage, medical diagnosis, military logistics, applications that predict the result of judicial decision, foreign policy of supply chain management. In the early 2020s, generative AI gained widespread prominence. ChatGPT, based on GPT-3, and other language models were tried by 14% of American adults.

Tools Used:

TensorFlow: TensorFlow applications can be run conveniently on your local machine, cloud, android and iOS devices. As it is built on a deployable scale, it runs on CPU and GPU.

PyTorch: PyTorch is similar to TensorFlow in terms of the nature of the projects chosen. However, when the priority is for faster development, PyTorch is the better choice. TensorFlow is gone in case the project involves larger and more complex project.

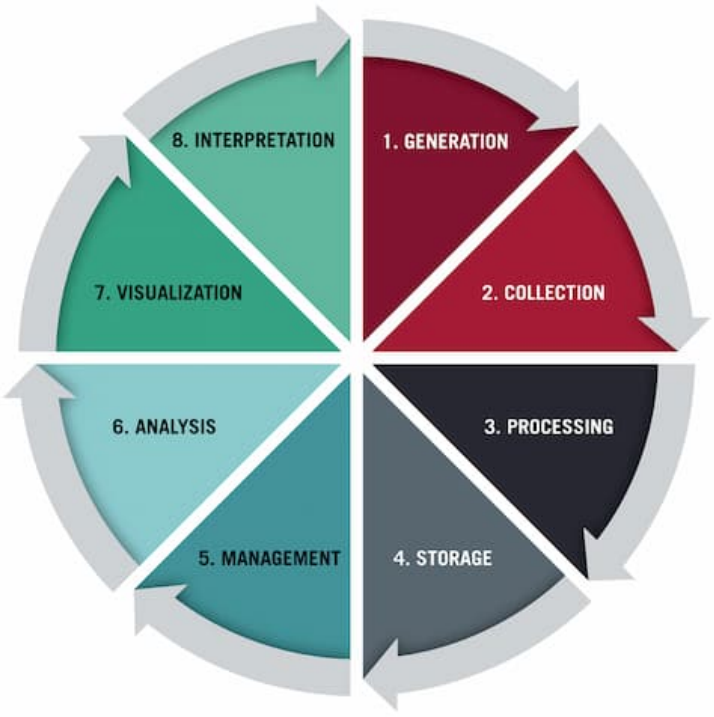
Scikit Learn: Scikit-learn is a widely praised Artificial Intelligence tool that simplifies the complexities of machine learning tasks. It boasts an intuitive and use-friendly interface that caters to learners across different proficiency levels. Scikit-learn equips users with the means to construct and deploy machine learning models effortlessly.

Natural Language Processing: It focuses on developing natural interactions between humans and computers. Specialized software helps machines process human language, create understandable words, and interact with humans through language.

Figstack: Figstack provides a comprehensive set of artificial intelligence tools designed to support developers in comprehending and documenting code more effectively. Its diverse array of features is geared towards simplifying the coding process, featuring a natural language interpreter capable of understanding code in nearly any programming language.

Data Collection:

Data collection is the process of collecting and analyzing information on relevant variables in a predetermined, methodical way so that one can respond to specific research questions, test hypotheses, and assess results.



Data Preprocessing :

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model.

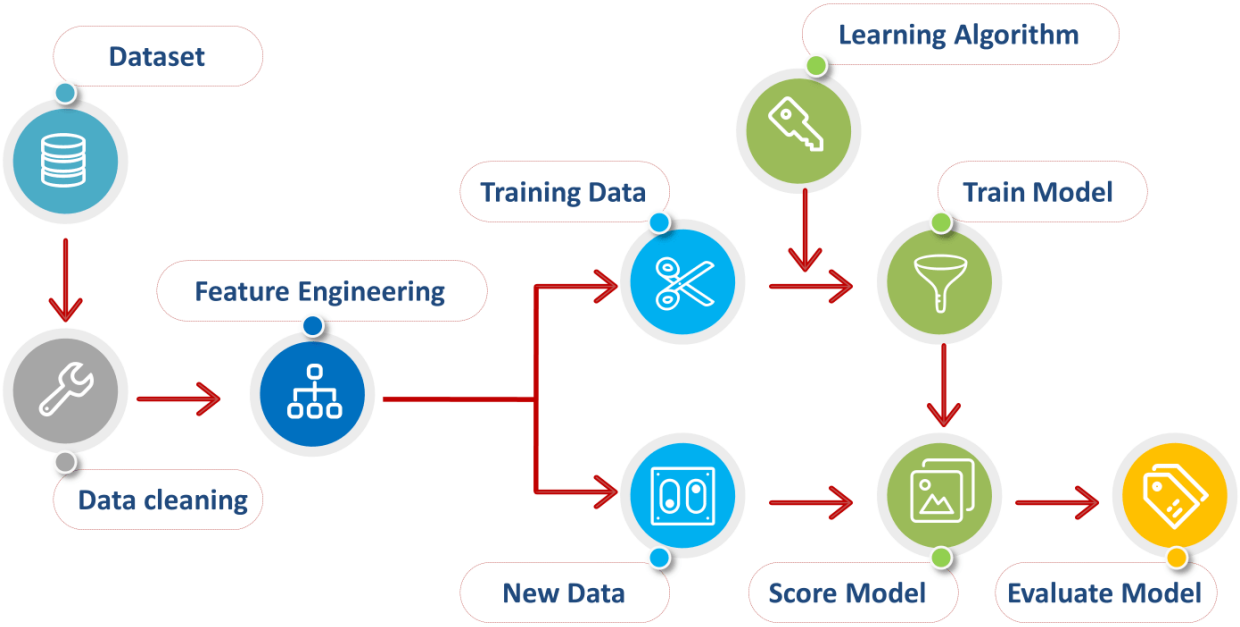
It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across clean and formatted data.

And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

It involves those steps:

1) Getting the Dataset
2) Importing Libraries
3) Importing datasets
4) Finding missing data
5) Encoding categorical data
6) Splitting dataset into training and test set
7) Feature Scaling



Feature Selection:

While developing the machine learning model, only a few variables in the dataset are useful for building the model, and the rest features are either redundant or irrelevant.

If we input the dataset with all these redundant and irrelevant features, it may negatively impact and reduce the overall performance and accuracy of the model.

Hence it is very important to identify and select the most appropriate features from the data and remove the irrelevant or less important features, which is done with the help of feature selection in machine learning.

1)Wrapper methods (forward, backward, and stepwise selection)
2) Filter methods (ANOVA, Pearson correlation, variance thresholding)
3)Embedded methods (Lasso, Ridge, Decision Tree)

Filter Method:

These methods are generally used while doing the pre-processing step. These methods select features from the dataset irrespective of the use of any machine learning algorithm. In terms of computation, they are very fast and inexpensive and are very good for removing duplicated, correlated, redundant features but these methods do not remove multicollinearity.

Selection of features is evaluated individually which can sometimes help when features are in isolation (don't have a dependency on other features) but will lag when a combination of features can lead to increase in the overall performance of the model.

Set of all features → Selecting the best subset → Learning algorithm → Performance

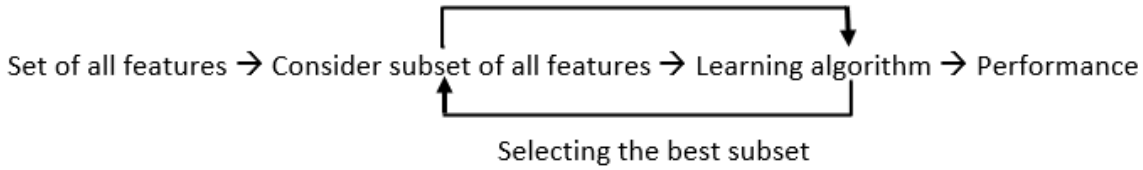
Wrapper Method:

Wrapper methods, also referred to as greedy algorithms train the algorithm by using a subset of features in an iterative manner.

Based on the conclusions made from training in prior to the model, addition and removal of features takes place.

Stopping criteria for selecting the best subset are usually pre-defined by the person training the model such as when the performance of the model decreases or a specific number of features has been achieved.

The main advantage of wrapper methods over the filter methods is that they provide an optimal set of features for training the model, thus resulting in better accuracy than the filter methods but are computationally more expensive.

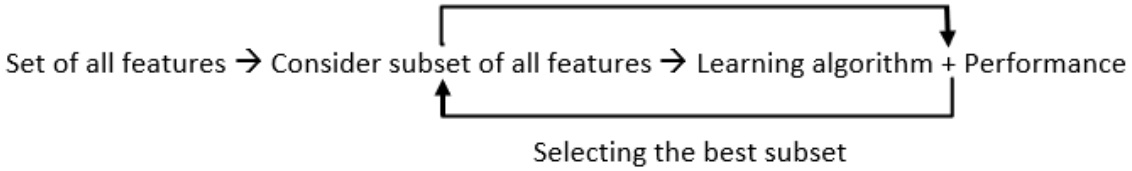


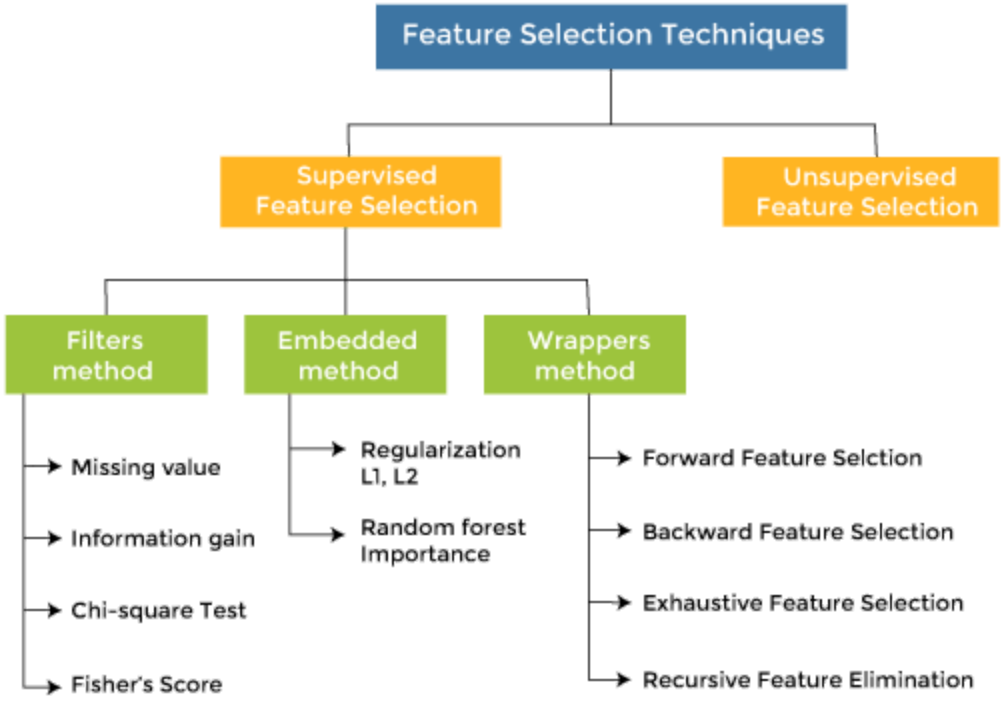
Embedded Method:

In embedded methods, the feature selection algorithm is blended as part of the learning algorithm, thus having its own built-in feature selection methods.

Embedded methods encounter the drawbacks of filter and wrapper methods and merge their advantages.

These methods are faster like those of filter methods and more accurate than the filter methods and take into consideration a combination of features as well.





MODEL SELECTION

Choose suitable artificial machine learning algorithms

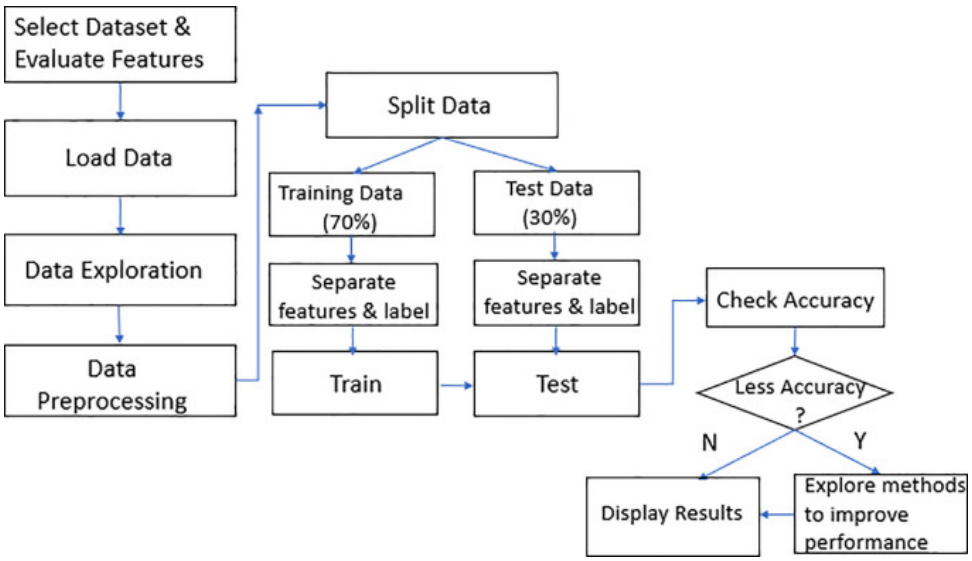
Example :

* **Logistic regression**
* **Random Forest**
* **Gradient boosting**

LOGISTIC REGRESSION

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. To ensure high accuracy detection, two main method are used to clean the data. The mean-based method Deals with missing values, and the clustering-based Method deals with outliers.

Extensive experiments are conducted to train and test The proposed classifier using a standard database. Import the relevant libraries



RANDOM FOREST

Random forest is called a Random Forest because we use Random subsets of data and features and we end up building a Forest of decision trees (many trees). Random Forest is also a classic example of a bagging approach as we use different subsets of data in each model to make predictions

Implementation of Random Forest Algorithm

The procedure of the random forest algorithm execution is done there are several steps involved; first, there is a requirement togather the information and to store the information. The gathered information are in the form of data set in an excel sheet. In data exploration, the entire data set checked and removed the unnecessary data that is present. However, the data which is further treated using a random forest algorithm in two ways by using train data set and then using the test data set.

GRADIENT BOOSTING

Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is

trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent.

* Import the necessary libraries.
* Setting SEED for reproducibility.
* Load the digit dataset and split it into train and test.
* Instantiate Gradient Boosting classifier and fit the model.
* Predict the test set and compute the accuracy score.

EVALUATION

Evaluation metrics are tied to machine learning tasks. There are different metrics for the tasks of classification and regression. Some metrics, like precision-recall, are useful for multiple tasks. Classification and regression are examples of supervised learning, which constitutes a majority of machine learning applications. Using different metrics for performance evaluation, we should be able to improve our model's overall predictive power before we roll it out for production on unseen data. Without doing a proper evaluation of the Machine Learning model by using different evaluation metrics, and only depending on accuracy, can lead to a problem when the respective model is deployed on unseen data and may end in poor predictions.

ACCURACY

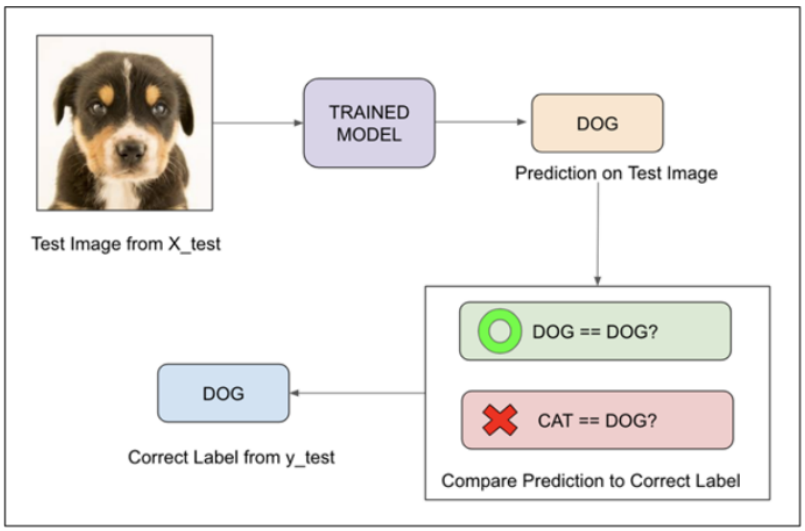
Accuracy simply measures how often the classifier correctly predicts. We can define accuracy as the ratio of the number of correct predictions and the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

When any model gives an accuracy rate of 99%, you might think that model is performing very well but this is not always true and can be misleading in some situations. I am going to explain this with the help of an example.

Example

Consider a binary classification problem, where a model can achieve only two results, either model gives a correct or incorrect prediction. Now imagine we have a classification task to predict if an image is a dog or cat as shown in the image. In a supervised learning algorithm, we first fit/train a model on training data, then test the model on testing data. Once we have the model's predictions from the X_test data, we compare them to the true y_values (the correct labels).



Precision

It explains how many of the correctly predicted cases actually turned out to be positive. Precision is useful in the cases where False Positive is a higher concern than False Negatives. The importance of *Precision is in music or video recommendation systems, e-commerce websites, etc. where wrong results could lead to customer churn and this could be harmful to the business.*

Precision for a label is defined as the number of true positives divided by the number of predicted positives.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Recall (Sensitivity)

It explains how many of the actual positive cases we were able to predict correctly with our model. Recall is a useful metric in cases where False Negative

is of higher concern than False Positive. It *is important in medical cases where it doesn't matter whether we raise a false alarm but the actual positive cases should not go undetected!*

Recall for a label is defined as the number of true positives divided by the total number of actual positives.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

F1 Score

It gives a combined idea about Precision and Recall metrics. It is maximum when Precision is equal to Recall.

F1 Score is the harmonic mean of precision and recall.

$$F1 = 2. \frac{Precision \times Recall}{Precision + Recall}$$

The F1 score punishes extreme values more. F1 Score could be an effective evaluation metric in the following cases:

- When FP and FN are equally costly.
- Adding more data doesn't effectively change the outcome
- True Negative is high

AUC-ROC

The Receiver Operator Characteristic (ROC) is a probability curve that plots the TPR(True Positive Rate) against the FPR(False Positive Rate) at various threshold values and separates the 'signal' from the 'noise'.

The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes. From the graph, we simply say the area of the curve ABDE and the X and Y-axis.

CONCLUSION

With the introduction of more innovative and new generation AI tools, healthcare is more advanced in the sense of more awareness, efficiency in delivering care, identification of developing complications, accurate diagnosis of diseases ahead of time, and most recent approaches for interventions.