A close-up of a logo

Description automatically generated

A logo of a company

Description automatically generated

**Joint Tech Internship Community Program**

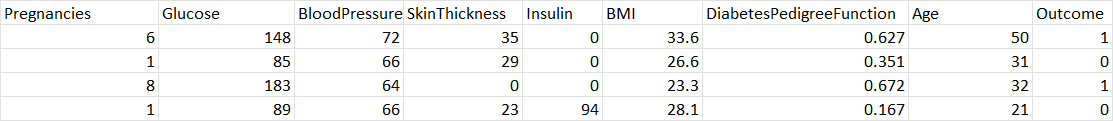
**Assignment 1**

**SUBMITTED BY**

NIDHARSAN V

Candidate ID:2024060092

(nidharsanv.22cse@kongu.edu)



**Features:**

There are some factors used to predict or classify the data or output based on input either in text or image . The output always depends on these factors. Such factors are known as features.

From above example

* Pregnancies
* Glucose
* BloodPressure
* SkinThickness
* Insulin
* BMI
* DiabetesPedigreeFunction
* Age

are the features

**Labels:**

A label is the output are the target variable that the model is being trained to predict .

In supervised learning each training data consist of an input ie features and associated label

From the above example the label is **Outcome**

**Prediction** :

The process of using a pretrained model or custom trained model to predict the output based on new or unseen inputs . It learns the pattern or dependency from the dataset from which It has been trained

**Outlier** :

The outliers are nothing but are the datas or datapoints which are exceptional where it give a different characterstics in the whole data and also different from other datas .

These are mainly occurs when the user enters the wrong data , logical mistakes , or really the data exist In the same ways

In clustering for instance if there are two clusters , the datapoint neither in cluster 1 nor in cluster 2 then such datapoint is said to be outliers

User uploaded image

From above example there is a data which has diabetespedigreefunction more than 1 but it still gives the outcome as 0

Even a single outlier affect the output . to prevent it we can simply remove it in the preprocessing stage or visualization stage

**Test** **Data** :

It is subset of original dataset which was unseen by the model upto training phase completion . After the completion of training the model these data are used as test cases to get the accuracy , precision and other scores

To check whether the model provide the correct output or not after validation test

Like 70% of data for traing and remaining 30% for testing

**Training** **data** :

It is also the subset of the original dataset in which most of the data is taken to train the model and get a fully pretrained model to predict the output from the unseen input such data is called as training data

**Model** :

Model refers to a mathematical or computational representation that learns patterns from data and makes predictions or decisions based on new data

Generally there are many algorithms in machinelearning , a model is an algorithm that has been trained on a specific dataset to perform specific task ie prediction , classification , detection

Each algorithm or model has parameters which defines the relation ship between in input and output

**Validation** **data**:

To determine the best model which has the high validation scores to perform a task . using this we can check whether the model performs well on the traing data but poorly in new data

For instance, if you have multiple versions of a model or different algorithms, you can use validation data to compare their performance and choose the one that performs best.

**Hyper** **parameter** :

These are parameters which are given before the training process actually starts and moniter , controls the learning and training process of the model

Example:

* Learning rate
* No of Epoch
* Batch size
* Weight

These hyper paramters can significantly boost the performance of the model , so we have to alter these parameters still it get more training score

**Epoch**:

Forward and backaward propagation of training dataset

Epoch is nothing but the one complete pass through the entire dataset during the training process of a model

During each epoch the parameters are adjusted to get the best fit model during training

**Loss** **function**:

A loss function is a way to measure how well or poorly a machine learning model is performing. It calculates the difference between the model's predictions and the actual values from the data.

The loss function helps the model learn by providing feedback on its performance. If the model's predictions are wrong, the loss function gives a high score, indicating that the model needs improvement.

By minimizing the loss, the model's parameters (like weights) are adjusted to make better predictions. The goal is to reduce the loss as much as possible to get the most accurate model.

**Learning** **rate**:

It is a numerical data that controls the how much the model changes its predictions based on the mistakes it makes during training

It determines the how the adjustments are done when the model learn from its errors

It is important to adjust model parameters , speed and quality

This learning rate also depends on time

**Overfitting** :

Overfitting is a common problem in machine learning where a model learns the training data too well, including the noise and outliers, resulting in poor performance on new, unseen data.

While training the model captures the pattern along with noise and fluctuations as the result the model do well in training data but very low score in tesing with new unseen data

**Underfitting** :

It is a situation in machine learning where a model is too simplistic to capture the patterns in the training data. This results in poor performance both on the training data and on new, unseen data.

In such case , the model fails to learn enough from the data, leading to irregular and inaccurate predictions.

The root cause of underfitting may be low datasets.

**Regularization**:

Method to reduce the overfitting

Adding penalty to loss function which by shrinking the coefficent estimates to zero

Prevent the model being too complex and restriction thr noise in training data

Helps the model to focus on important features and patterns and make it to perform not only training data but alsp in new unseen data too.

Cross Validation:

It is the statsitical method which is used to evaluvate the performance of a model . It involves the paritioning the dataset to train the model and a test to evaluvate it

It is to test the models ability to generalize to an independent dataset

It is mainly used to avoid overfitting

**Feature Engineering:**

Feature engineering is the process of using domain knowledge to extract features (variables, attributes) from raw data to improve the performance of machine learning algorithms.

It involves creating new features, transforming existing ones, and selecting the most relevant features that contribute to the predictive power of the model.

**Dimensionality Reduction**:

Dimensionality reduction is the process of reducing the number of random variables or features under consideration, by obtaining a set of principal variables.

It is a crucial step in data preprocessing for machine learning and data analysis, especially when dealing with high-dimensional data.

The main goals of dimensionality reduction are to simplify the model, reduce computational cost, and improve model performance by mitigating the curse of dimensionality

**Bias:**

bias refers to the error introduced by approximating a real-world problem, which might be complex, by a simplified model.

Bias can come from various sources, including the data, the model, or the assumptions made during the model training process.

Understanding and managing bias is crucial for building accurate and reliable machine learning models.

When the bias is higher it leads to underfitting

**Variance:**

variance refers to the model's sensitivity to small fluctuations in the training dataset.

A model with high variance pays too much attention to the training data, capturing noise and details that do not generalize well to new data.

This leads to overfitting, where the model performs well on the training data but poorly on unseen data.

The error due to excessive sensitivity to the specific training data.this leads to overfitting