**Abstract:** This study employs machine learning methods to categorize tumor cells as benign or malignant cancer. The dataset used to contains various features related to tumor or cancer cell characteristics, with the goal of predicting the tumor’s class accurately. The study applies supervised ensemble classification methods to train machine learning models on a dataset of 569 instances, each described by 32 features. The primary objective is to forecast whether a tumor is cancerous or benign with a significant amount of accuracy.

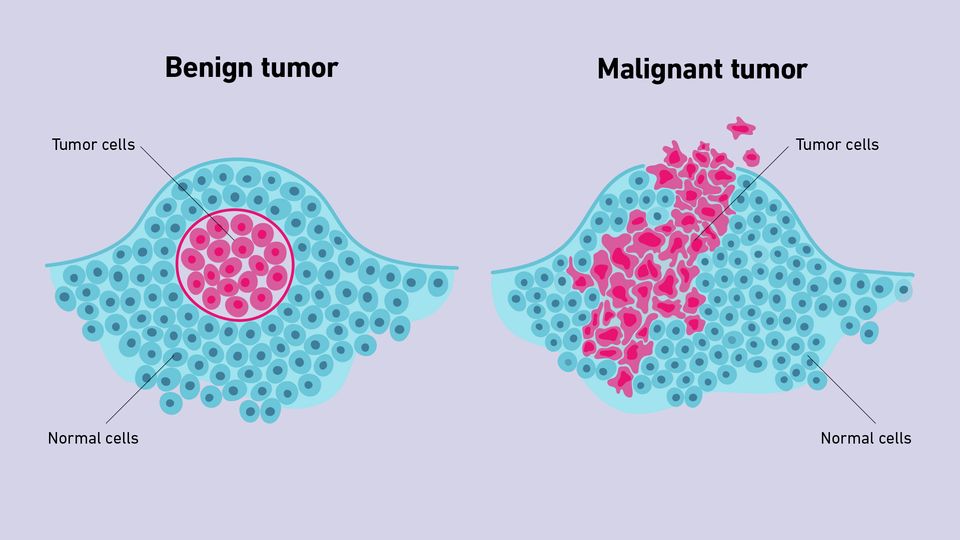
[](https://assets.technologynetworks.com/production/dynamic/images/content/364765/benign-vs-malignant-tumors-364765-640x360.webp?cb=12252466)

Figure-01 (heat map)

**Introduction:**

Tumors are abnormal mases of cells that result from irregular cell division in the body. These cells grow and divided more than they should or don’t die when they should. While benign tumors are benign in nature and don’t spread to other parts of the body, malignant tumors or cancerous cells , develop uncontrollably and have the potential to metastasis for sprade out to other parts of the body. Through the use of machine learning algorithms, this study seeks to different between benign or malignant.

**Materials Used:** Google Colab, Cancer data set (in csv format), CPU core i5 6th gen, Ram: 4GB

**Methodology:**

**Exploratory Data Analysis:** Exploratory Data analysis is one of the data preprocessing technique which extract the vital features those suitable for fit on machine learning model, deep learning model, data science

**Data Collection:** We found[**cancer data**](https://www.kaggle.com/datasets/erdemtaha/cancer-data)on kaggle on csv (coma separated value) format. That Data set have around 569 data and 33 columns.

**Analyze data for cleaning:** Checking null values, noise data, unnecessary features, and duplicate data

**Findings of EDA:** In EDA we found some decisions...

* All features are categorical.
* In all categorical features have lots of unique value but most of these numerical value. So no need to apply any kind of encoding except 'diagnosis'
* Only 'diagnosis' column have binary category and string data type. Here may be perform binary encoding based on data set.
* The 'id' and 'Unnamed 32' detected as noise in whole data. So these should be removed.
* The 'diagnosis' column is the target feature of the data.

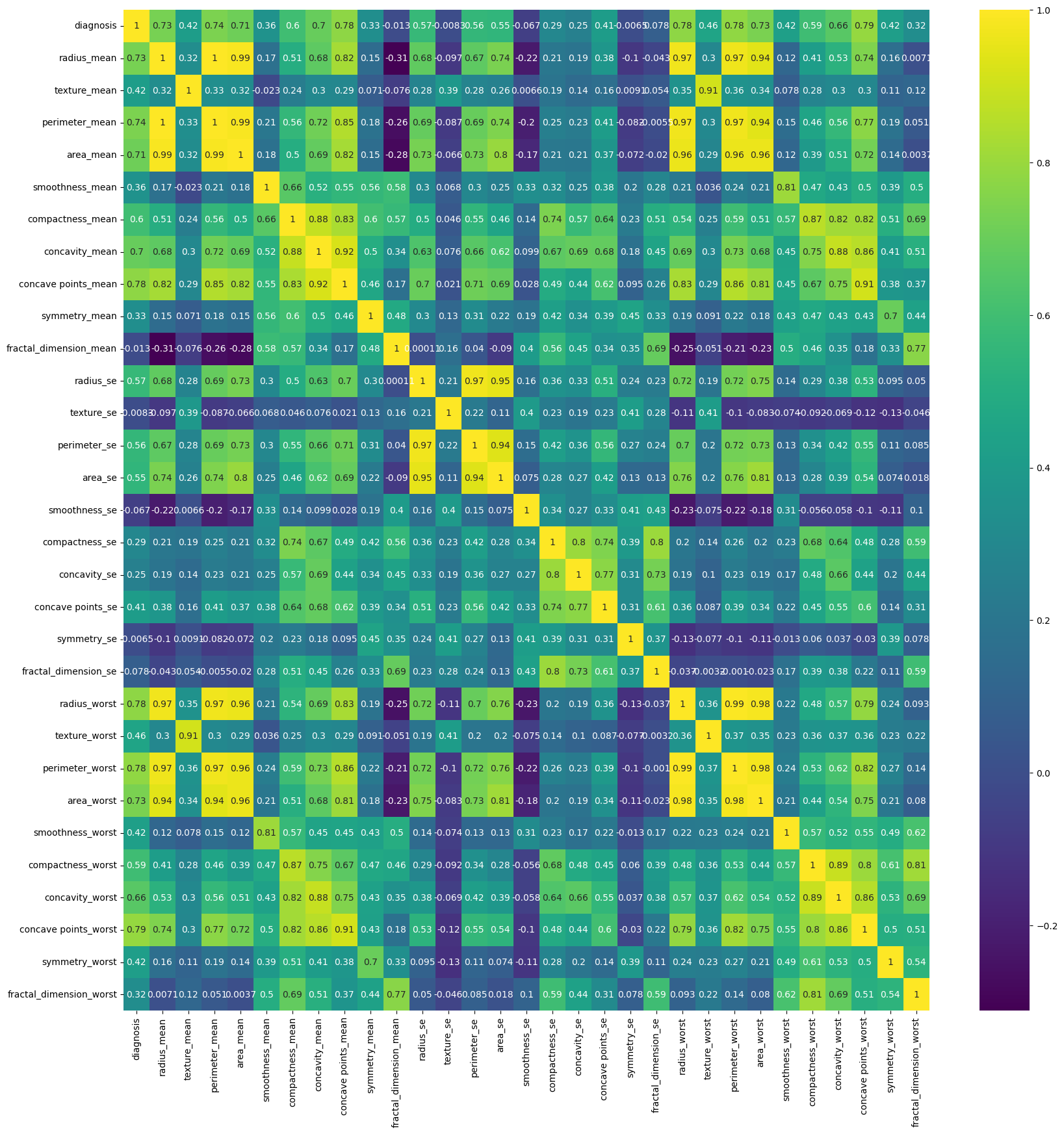
**Feature Engineering**: Feature Engineering is a machine learning technique that transforms available data sets into sets of figures essential for a specific task. It reduce data inconsistencies (such as incomplete, incorrect, anomalies) form data set**.**

**What is** [**features encoding**](https://www.elastic.co/guide/en/machine-learning/current/ml-feature-encoding.html#:~:text=Machine%20learning%20models%20can%20only,process%20is%20called%20feature%20encoding.): Machine learning model can only works with numerical values. For this reason it is necessary to transform the categorical value into numerical value

**Common types of encoding:**

* Label encoding: Label encoding is used on multiclass classifications of features.
* Binary encoding: Binary encoding is used on binary classifications of features. It presents either 0 or 1
* Tergate\_mean\_encoding: replace categorical values with mean values of target features.
* Frequency encoding: How many times a given categorical value is present in relation with a feature

**Actions of above EDA findings:**

* Here on the data set Id and unname32 deleted.
* ‘diagnosis’ column is selected as target columns
* The ‘diagnosis’ column is encoded by label binarizer. Because ‘diagnosis’ column have only two unique value. Which is binary classification. So we encoded it by LebelBinaraizer. LebelBinaraizer is an object of sklearn.preprocessing.
* Finally we find the data full of numerical value.

**Data visualization:**

In this project we use four type of plot by using matplotlib and seaborn such as heatmap, boxplot, histlot, scatterplot. We try to heatmap we find the data correlations with each other and try to understand what’s the correlation is with the target column. We try to understand given certain features are suitable for correct prediction. If there have some features less important to predict target category that should be removed from features though sometimes it change the tiny state of prediction accuracy but it is useful.

Figure-02 (heat map)

In figure-02 we see that corresponding to diagnoses 'fractal\_dimension\_mean', 'texture\_se', 'smoothness\_se', and ‘symmetry\_se' these features correlations are not good at all. According to color bar scale there score are -0.013, -0.0083, -0.067 and 0.0065 respectively. So we decided that from above analysis we might remove these features to show batter data correlation to predict batter. Sometimes data correlation may depend on data density and low outliers. We try to show data outliers by box plot. Create box plot for every features and in figure-03 boxplot there have a several number of outliers on each features as displayed bellow:

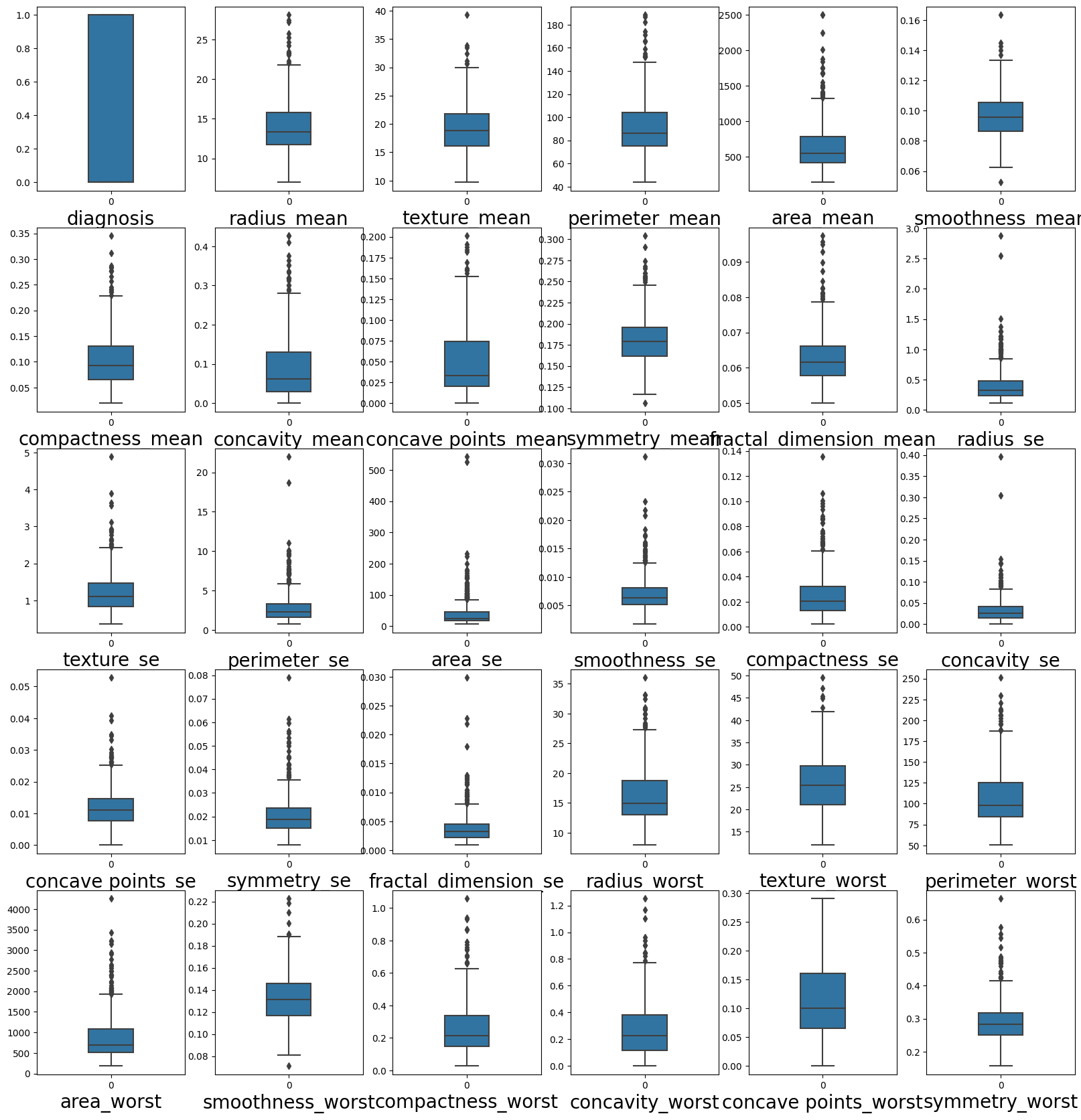


Figure-03 (heat map)

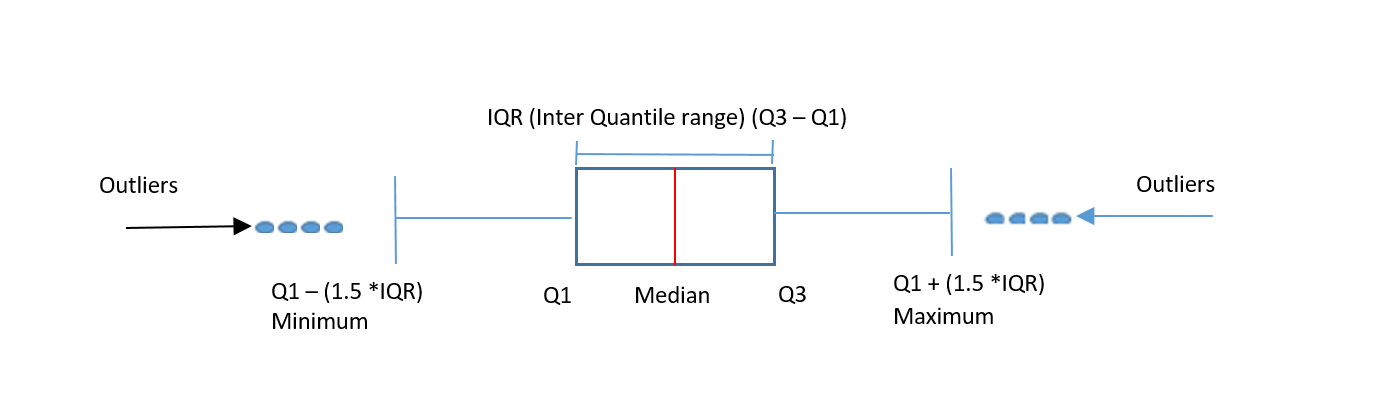
**Handel outliers:**

There are some statistical technique to resolve outliers from data set. From those most tow useful technique are:

* z-score
* Inter Quantile Range

Here we use Inter Quantile Range to reduce outliers. In this technique where whole data is divided into 3 quantile such as 25% quantile, median, 75% quantile.

**Mathematical intrusion about outliers handling using Inter Quantile Range:**



In figure 04 IQR stands for Inter Quantile Range and it is calculated as:

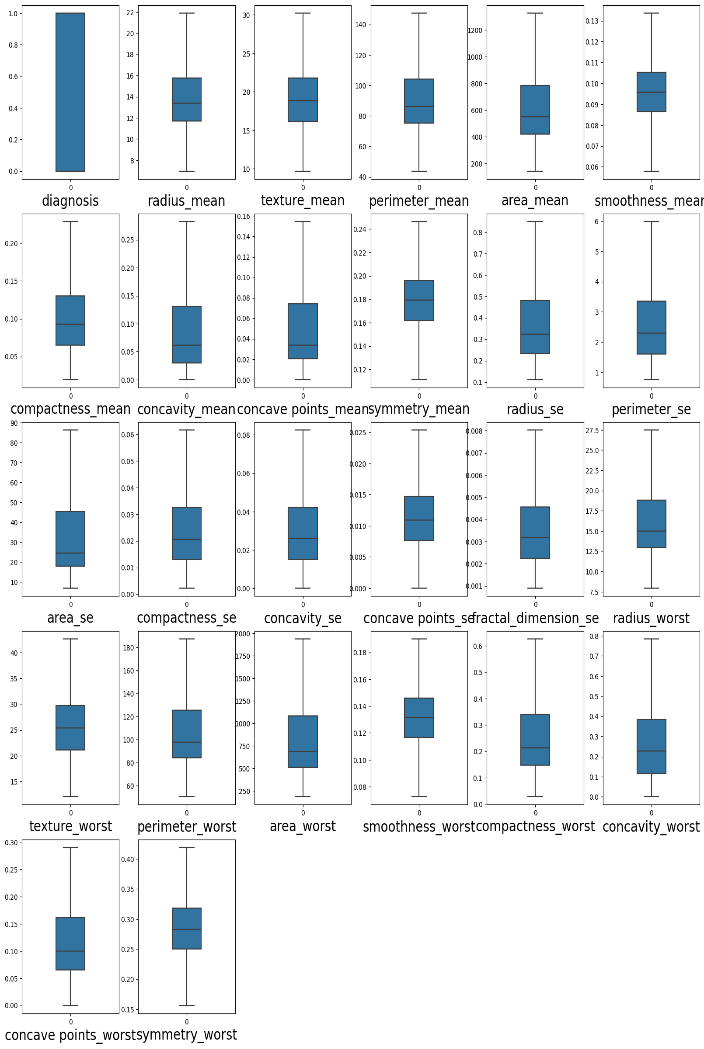
IQR = Q3 – Q1

lower\_limit = Q1 - (1.5\*IQR)

upper\_limit = Q3 + (1.5\*IQR)

Those data are out of the range lower\_limit and upper\_limit these called outliers. So these outliers are replace by lower\_limit and upper\_limit respectively.

Figure-04 (heat map)

**After doing actions based on EDA below show the data correlations by heatmap and show removed outliers by boxplot:**

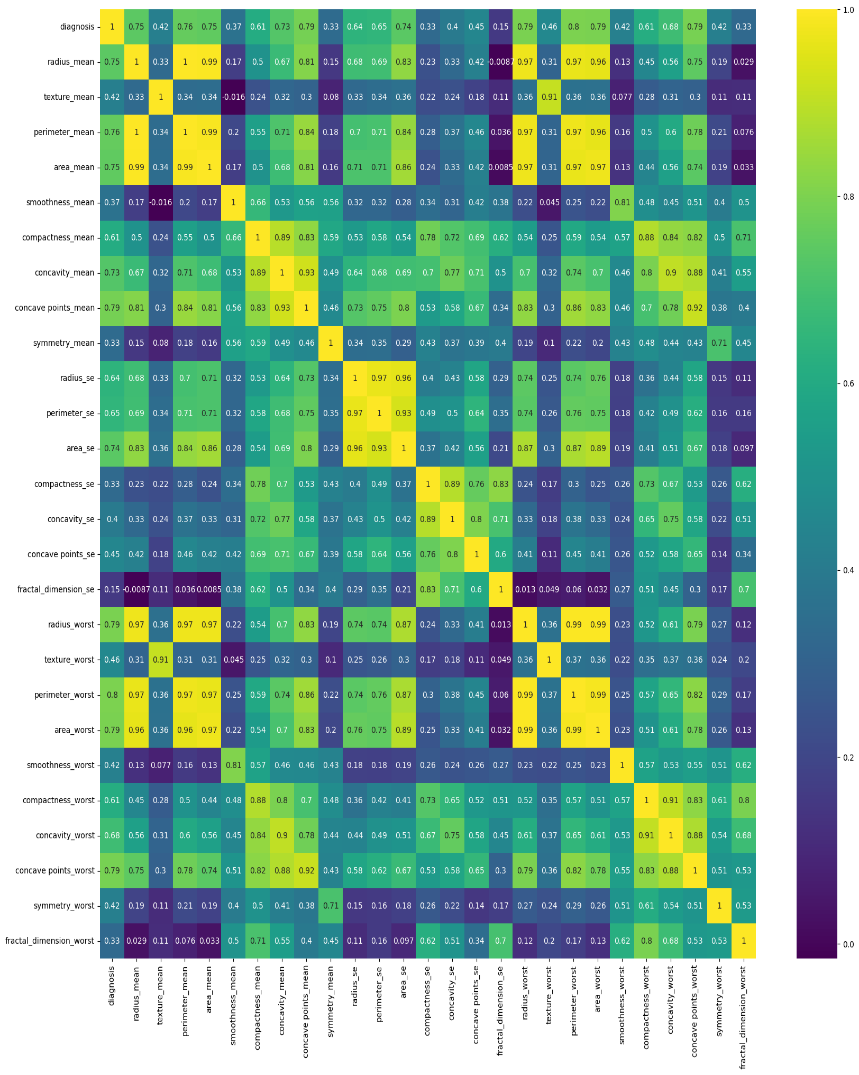


Figure-06 (heat map)

Figure-05 (heat map)

This is displayed that after doing these actions outliers has been gone and data correlations is plotted as batter corresponding to target column as ‘diagnosis’. So data may be predict batter.

**Model training and applying algorithms:**

**Proposed prediction model: This study and analysis is shows that our targeted feature for prediction has binary class classification. Describing scatter plot and correlation of data with target feature it is assumed that this data was fit linear model. There have binary classification on targeted column so we try with three classification algorithms these are: SVC (Support vector classifier, Logistic regression and voting classifier:**

**SVC(Support vector Classifier):**