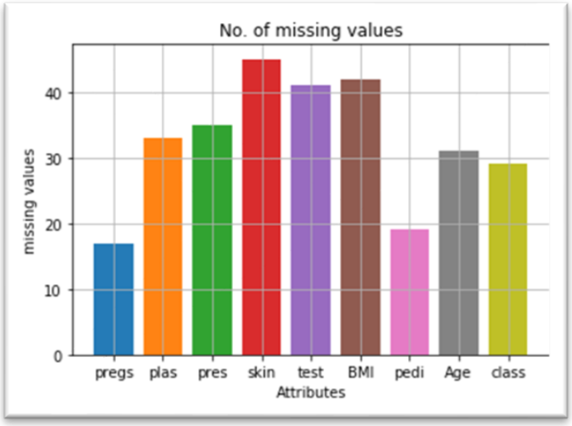
IC272: Lab-2 Report

Jai Prakash Yadav • +91-9306871378 • b19247@students.iitmandi.ac.in

Inference from Task 1

(Task 1 – To plot graph of attributes with missing values in them)

Attribute	Missing values
Pregs	17
Plas	33
Pres	35
Skin	45
Test	41
BMI	42
Pedi	19
Age	31
Class	29



Used the inbuilt function isnull() to calculate missing values for each of the attributes in the data set.
We can see that Attribute – ‘skin’ has the maximum NaN values and Attribute – ‘pregs’ has the minimum.

Inference from task 2.

(Task 2(a) – to delete the tuple having equal to or more than third of attributes with missing values. To print number of tuples deleted and their row number)

Total number of deleted tuples – 39
Rows no of deleted tuple – 1, 39, 40, 53, 54, 83, 89, 103, 125, 136, 145, 210, 211, 212, 213, 249, 250, 254, 280, 281, 284, 314, 321, 335, 429, 430, 449, 450, 451, 471, 472, 473, 474, 718, 719, 720, 721, 753, 766.

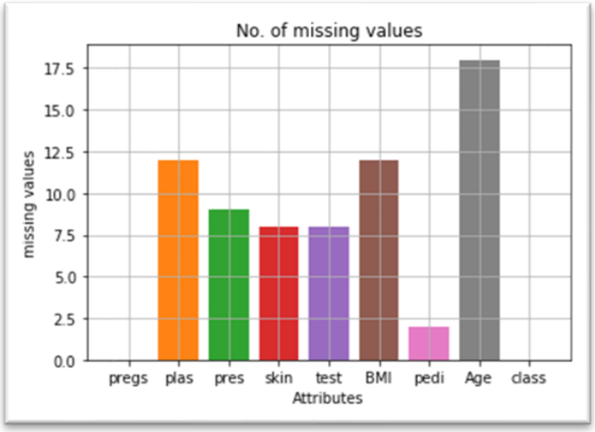
(Task 2(b) – to delete the tuple having missing value in the class attribute. And also, to print number of deleted tuples and also print their row number)

Total number of deleted tuples – 21
Rows number of deleted tuples – 8, 13, 28, 29, 35, 62, 92, 95, 107, 110, 130, 131, 132, 133, 149, 182, 188, 218, 308, 746, 748.

Inference from task 3.

(Task 3 – To find missing values in each attribute. And total number missing values in file.)

Attribute	Missing values
Pregs	0
Plas	12
Pres	9
Skin	8
Test	8
BMI	12
Pedi	2
Age	18
Class	0



Total number of missing values – 69

Inference from task 4.

(Task 4(a) – Replace the missing values by mean of their respective attribute.)
Use the function fillna() to replace the all the null values with other value.

1. (Task 4(a)-i- To compute mean, median, mode and standard deviation for each attribute and compare it with that of original file.)

New data –

	Mean	Median	Mode 1	Mode 2	Standard deviation
Pregs	3.885	3.000	1.000	NaN	3.373
Plas	120.66	118.000	99.000	100.000	30.990
Pres	69.001	72.000	70.000	NaN	19.691
Skin	20.348	23.000	0.000	NaN	15.946
Test	77.814	36.000	0.000	NaN	110.607
BMI	32.009	32.009	32.000	NaN	7.764
Pedi	0.476	0.382	0.254	0.258	0.333
Age	33.094	29.000	22.000	NaN	11.519
Class	0.343	0.000	0.000	NaN	0.475

Original data –

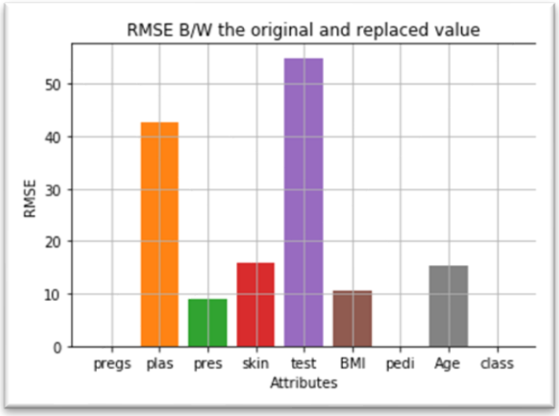
	Mean	Median	Mode 1	Mode 2	Standard deviation
Pregs	3.845	3.000	1.000	NaN	3.369
Plas	120.894	117.000	99.000	100.000	31.972
Pres	69.105	72.000	70.000	NaN	19.355
Skin	20.536	23.000	0.000	NaN	15.952
Test	79.799	30.500	0.000	NaN	115.244
BMI	31.992	32.000	32.000	NaN	7.884
Pedi	0.471	0.372	0.254	0.258	0.331
Age	33.240	29.000	22.000	NaN	11.760
Class	0.348	0.000	0.000	NaN	0.476

From the above two table it is clearly visible that centre of the new data does not change much from the original data. Because we only replace the missing values with the mean, which does not affect the centre of the data.

(Task 4(a)-ii- To compute root mean square error between the original and replaced values for each attribute.)

Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data.

Attribute	RMSE
Pregs	0.000
Plas	42.643
Pres	8.950
Skin	15.839
Test	54.969
BMI	10.450
Pedi	0.046
Age	15.365
Class	0.000



(Task 4(b) – Replace the missing value in each attribute using linear interpolation technique.)

1. (Task 4(b)-i- To compute mean, median, mode and standard deviation for each attribute and compare it with that of original file.)

New data

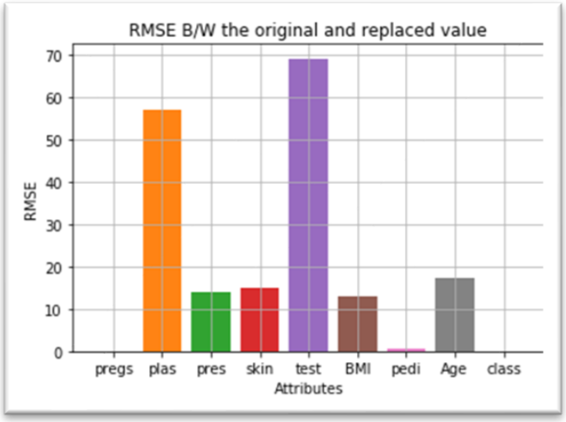
	Mean	Median	Mode 1	Mode 2	Standard deviation
Pregs	3.885	3.000	1.000	NaN	3.373
Plas	120.349	1117.000	99.000	100.000	31.274
Pres	69.109	72.000	70.000	NaN	19.735
Skin	20.392	23.000	0.000	NaN	15.975
Test	77.355	27.000	0.000	NaN	110.755
BMI	32.046	32.250	32.000	NaN	7.792
Pedi	0.477	0.382	0.254	0.258	0.334
Age	33.216	29.000	22.000	NaN	11.652
Class	0.343	0.000	0.000	NaN	0.475

Original data

	Mean	Median	Mode 1	Mode 2	Standard deviation
Pregs	3.845	3.000	1.000	NaN	3.369
Plas	120.894	117.000	99.000	100.000	31.972
Pres	69.105	72.000	70.000	NaN	19.355
Skin	20.536	23.000	0.000	NaN	15.952
Test	79.799	30.500	0.000	NaN	115.244
BMI	31.992	32.000	32.000	NaN	7.884
Pedi	0.471	0.372	0.254	0.258	0.331
Age	33.240	29.000	22.000	NaN	11.760
Class	0.348	0.000	0.000	NaN	0.476

2. (Task 4(b)-ii- To compute root mean square error between the original and replaced values for each attribute.)

Attribute	RMSE
Pregs	0.000
Plas	57.055
Pres	13.771
Skin	14.875
Test	68.984
BMI	12.819
Pedi	0.508
Age	17.399
Class	0.000



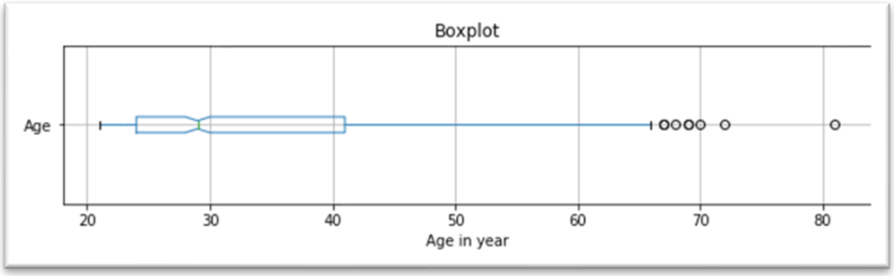
The RMSE value of interpolation Technique is more than that of Mean filling technique. So, from these RMSE value we can conclude that Mean Filling Technique is better or more efficient in this case.

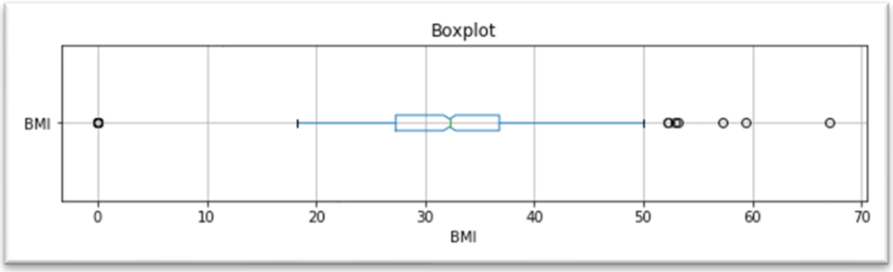
Inference from task 5.

(Task 5(a) – To find the outliers in the attribute “Age” and “BMI”.)

Outliers in Age: 69.0 67.0 72.0 81.0 67.0 70.0 68.0 69.0
Outliers in BMI: 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 53.2 67.1 52.3 52.3 52.9 59.4 57.3

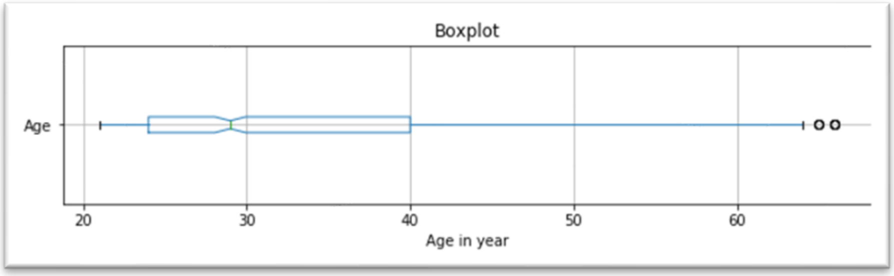
	Min. value	Max. value	Q1-1.5*IQR	Q3+1.5*IQR
Age	21	81	-1.5	66.5
BMI	0	67.1	13.05	51.05

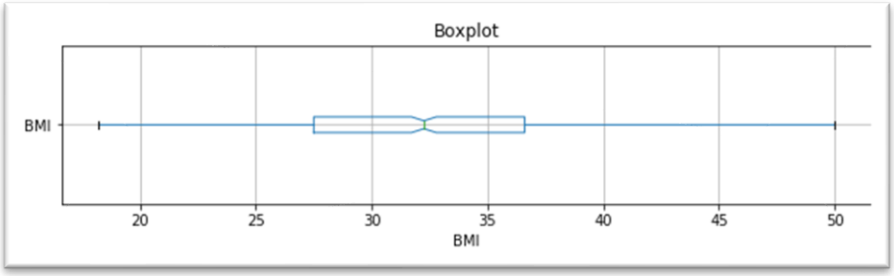




(Task 5(b) – To replace outliers by median of attribute.)

	Min. value	Max. value	Q1-1.5*IQR	Q3+1.5*IQR
Age	21	66	0.0	64.0
BMI	18.2	50	13.85	50.25





After replacing the outliers with the median value, the value of whiskers shift.
Even though the number of outliers had reduced significantly, few outliers are still present in the Age attribute but not in BMI attribute. This is due to the change in value of whiskers (whiskers slightly shift towards median