CM5

January 29, 2021

1 Data Cleaning of datasets

Interpolation with linear method in forward direction for filling the missing values in both the dataset. We have used this method because it will take into consideration only the values above and below the missing value, thus not taking into the all the values this will, which will keep the test set in unknown to the classifier upto some extent.

1.1 Required Libraries

[1]:
```
import pandas as pd
```

1.2 Data Cleaning of Heart Diseases Dataset

[2]:
```
df_heart= pd.read_csv("heart_disease_missing.csv")
```

[3]:
```
df_heart=df_heart.interpolate(method ='linear', limit_direction ='forward')
df_heart
```

[3]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	\
0	76	0	2	140.102822	197.105970	0	2.0	115.952071	0	
1	43	0	0	132.079599	341.049462	1	0.0	135.970028	1	
2	47	1	2	107.899290	242.822816	0	1.0	152.210039	0	
3	51	1	2	99.934001	288.887585	0	1.0	143.049207	1	
4	57	1	0	110.103508	334.952353	0	1.0	143.099327	1	
..	…	…	..	…	…	…	…	…	…	
207	56	1	2	148.997583	256.189595	1	0.0	141.981335	1	
208	57	0	0	127.981407	302.985611	0	0.0	158.992132	0	
209	54	1	2	150.188534	232.117551	0	0.0	164.977674	0	
210	41	1	2	129.918793	214.008059	0	0.0	167.851493	0	
211	41	0	2	112.075764	268.005496	0	0.0	172.008896	1	

	oldpeak	slope	ca	thal	target
0	1.284822	1.0	0	2.175904	1
1	3.110483	1.0	0	3.082071	0
2	-0.023723	2.0	0	2.020827	0
3	1.195082	1.0	0	2.100312	1
4	3.082052	1.0	1	2.831509	0
..	…	…	..	…	…

```
207  0.606726    1.0   1  0.983927       0
208  1.160978    2.0   1  1.884149       1
209  1.715230    2.0   0  2.970521       1
210  1.992138    1.0   0  1.879487       1
211  0.085251    2.0   0  2.194162       1

[212 rows x 14 columns]
```

1.3 Data cleaning for Iris Dataset

[4]:
```
df_iris= pd.read_csv("iris_dataset_missing.csv")
```

[5]:
```
df_iris=df_iris.interpolate(method ='linear', limit_direction ='forward')
df_iris
```

[5]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.045070	2.508203	3.018024	1.164924	Iris-versicolor
1	6.325517	2.115481	4.542052	1.413651	Iris-versicolor
2	5.257497	3.814303	1.470660	0.395348	Iris-setosa
3	6.675168	3.201700	5.785461	2.362764	Iris-virginica
4	5.595237	2.678166	4.077750	1.369266	Iris-versicolor
..	…	…	…	…	…
100	4.874848	3.217348	1.592887	0.123588	Iris-setosa
101	5.564197	2.771731	3.483588	1.074754	Iris-versicolor
102	5.548047	4.249211	1.453466	0.214527	Iris-setosa
103	5.510482	2.652867	4.276817	1.298032	Iris-versicolor
104	4.538713	3.056142	1.545136	0.241424	Iris-setosa

[105 rows x 5 columns]

1.4 References

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.interpolate.html