Satyam Vatts
Avneet Kaur

PURPOSE

This report provides a detailed description of different techniques to correctly predict the type of Treatment received by patients of Acute Pancreatitis (PEX or Traditional) based on independent predictors. It is a continuation of previous report that dealt with missing data imputation. In this analysis, we have utilized the MICE, AMELIA and MISSFORST Imputed datasets to perform the following analysis:

- Comparing Mean & Standard Deviation of continuous predictors between the two groups of patients
- Comparing the percentage distribution different categories of categorical variables between the two groups
- Discriminant Analysis:
    - Testing the Assumptions of Discriminant Analysis
    - Performing DA and interpreting the results
    - Determining the cause of misclassification of 3 of the misclassified pateints
- K-nearest Neighbour Analysis:
    - Performing KNN Prediction and checking accuracy
    - Comparing accuracy of KNN with DA

Data Description (Re-Cap)

The dataset consists of 165 observations corresponding to patients involved in the study. Among them, 83 were treated with PEX (Plasma Exchange Therapy) treatment while others were given treatment according to Vietnam's Ministry of Health's guidelines in 2015. As we analysed the data, we observed that many variables present in the dataset are redundant and fail to provide any additional information. Also, in the previous analysis we remove some more variables that were not significant in prediction analysis.

Group Summary Statistics Comparison

The summary statistics provide an overview of differences between the two groups of patients – PEX treated and Traditionally treated patients based on various independent variables. The comparison has been made based on Mean and Standard deviations of Numerical Variables and Percentage distribution of categories for categorical data. The results of summary statistics are captured in the tables below:

Numerical Variables:

	AMELIA IMPUTED DATA				MICE IMPUTED DATA				MISS FOREST IMPUTED DATA			
	Group 1: PEX Trated (N=81)		Group 2: Traditionally Treated (N=84)		Group 1: PEX Trated (N=81)		Group 2: Traditionally Treated (N=84)		Group 1: PEX Trated (N=81)		Group 2: Traditionally Treated (N=84)	
Variable/Predictor	**STDEV**	**MEAN**	**STDEV**	**MEAN**	**STDEV**	**MEAN**	**STDEV**	**MEAN**	**STDEV**	**MEAN**	**STDEV**	**MEAN**
Age	8.98	40.27	9.52	41.92	8.98	40.27	9.52	41.92	8.98	40.27	9.52	41.92
rv_ngaydt	3.36	6.96	2.92	6.51	3.36	6.96	2.92	6.51	3.36	6.96	2.92	6.51
ts_ruou_nam	5.32	3.44	8.26	6.73	5.32	3.44	8.26	6.73	5.32	3.44	8.26	6.73
ls_tt_bmi_t0	2.57	22.69	2.62	22.59	2.59	22.66	2.54	22.66	2.57	22.71	2.54	22.66
ls_tn_mach_t0	17.32	107.64	18.06	107	17.52	107.51	18.18	107.01	17.23	107.86	17.96	106.71
ls_tn_nhiet_t0	0.53	37.1	3.71	36.8	0.53	37.1	3.71	36.78	0.48	37.1	3.71	36.78
ls_tn_spo2_t0	1.58	97.4	1.85	96.8	1.61	97.33	1.97	96.69	1.58	97.4	1.83	96.78
ls_diem_apache_t0	3.18	3.9	2.91	2.78	3.15	4.04	3.05	3.17	3.04	3.9	2.62	2.89
ls_diem_ranson_t0	0.95	1.63	1.05	1.28	0.94	1.65	1.05	1.36	0.9	1.62	0.98	1.3
ls_diem_ct_t0	1.54	3.97	1.53	3.53	1.85	4.04	2.1	3.46	1.43	4	1.29	3.49
ls_diem_imrie_t0	0.82	1.53	1	1.19	0.87	1.54	1.08	1.24	0.81	1.51	0.95	1.18
ls_diem_sofa_t0	1.62	1.39	1.59	1.34	1.69	1.48	1.66	1.58	1.61	1.39	1.54	1.51
cls_hh_bc_t0	4.43	11.36	3.14	9.9	4.43	11.37	3.21	9.75	4.41	11.36	3.12	9.87
cls_hh_bc_t30	3.24	8.88	3.59	9.09	3.27	8.26	3.31	8.86	2.31	9.1	2.8	8.94
cls_hh_bc_t54	3.61	10.33	3.68	9.49	4.76	9.04	4.56	10.14	2.59	10.08	2.1	9.23
cls_hh_bc_t72	4.34	11.23	4.09	9.92	6.03	12	5.82	11.01	3.47	11.11	2.63	10.72
cls_hh_hct_t0	0.08	0.39	0.07	0.39	0.08	0.39	0.07	0.39	0.07	0.39	0.07	0.39
cls_hh_hct_t30	0.05	0.33	0.06	0.35	0.08	0.34	0.07	0.35	0.04	0.34	0.06	0.35
cls_hh_hct_t72	0.04	0.31	0.05	0.34	0.06	0.32	0.06	0.34	0.04	0.31	0.03	0.33
cls_hh_pt_t0	22.02	88.05	18.45	85.73	24.02	85.79	18.47	86.52	21.29	87.26	17.38	85.98
cls_hh_pt_t30	13.56	78.78	13.58	81.39	19.71	77.76	17.3	80.86	10.1	79.7	10.72	81.61
cls_hh_pt_t72	14.7	80.72	14.66	81.29	28.61	79.04	27.61	87.83	10.72	82.3	11.99	81.86
cls_hh_aptt_t0	0.55	1.17	0.27	1.12	0.52	1.19	0.24	1.1	0.52	1.19	0.24	1.1
cls_hh_aptt_t30	0.33	1.24	0.23	1.13	0.63	1.4	0.37	1.18	0.29	1.27	0.18	1.14
cls_hh_fib_t0	1.99	5.22	2.05	5.99	2.01	5.27	2.05	6.02	1.91	5.22	2.02	6.02

Satyam Vatts
Avneet Kaur

cls_hh_fib_t30	1.79	5.85	1.99	6.69	2.35	5.28	2.15	6.56	1.14	5.89	1.62	6.62
cls_sh_ure_t0	8.8	5.48	4.33	5.29	8.77	5.58	4.33	5.29	8.77	5.46	4.32	5.31
cls_sh_ure_t30	2.82	3.94	2.74	4.74	2.57	3.94	2.87	4.88	1.24	4.1	2.48	4.54
cls_sh_ure_t72	5.14	7.05	6.21	5.65	6.02	6.69	6.87	6.17	2.71	5.21	4.59	5.03
cls_sh_cre_t0	43.24	74.39	97.16	93.06	42.19	75.17	97.57	93.77	40.92	73.62	97.15	92.47
cls_sh_cre_t30	34.83	65.36	51.95	78.28	39.54	68.8	59.34	80.76	16.8	68.89	48.43	75.81
cls_sh_glu_t0	8.39	11.86	6.53	11.06	8.41	11.47	5.92	11.45	7.88	11.46	5.27	11.23
cls_sh_chol_t0	7.58	15.99	6.26	12.82	7.49	16.29	6.41	13.58	7.26	15.9	5.79	13.24
cls_sh_chol_t30	3.91	6.7	3.79	7.3					1.9	7.76	2.86	7.93
cls_sh_tri_t0	28.18	33.74	13.68	21.76	28.32	33.09	15.14	22.39	28.27	33.17	13.6	21.5
cls_sh_tri_t30	11.53	7.86	7.67	8.38	19.98	12.86	14.13	11.41	9.34	9.95	5.9	8.89
cls_sh_amy_t0	421.57	400.16	333.9	348.94	406.82	442.28	358.28	428.85	354.05	453.19	262.46	353.03
cls_sh_lip_t0	326.09	525.6	322.66	389.7	388.34	644.78	346.71	482.5	289.48	571.65	245.31	466.37
cls_sh_pro_t0	9.74	59.31	9.02	58.96	10.46	59.18	9.71	59.21	9.54	59.33	7.1	59.21
cls_sh_na_t0	5.59	131.08	5.72	132.49	6.16	130.96	6.32	132.11	5.52	131.24	5.61	132.55
cls_sh_na_t30	3.75	136.77	4.33	134.85	5.4	137	4.55	135.02	2.19	137.23	3.87	134.96
cls_sh_ka_t0	0.57	3.71	0.49	3.63	0.58	3.73	0.52	3.6	0.57	3.71	0.48	3.64
cls_sh_ka_t30	0.33	3.43	0.38	3.4	0.55	3.49	0.49	3.44	0.22	3.44	0.35	3.41
cls_km_ph_t0	0.05	7.38	0.07	7.41	0.05	7.37	0.06	7.4	0.05	7.38	0.06	7.41
cls_km_ph_t30	0.08	7.4	0.06	7.41	0.13	7.43	0.11	7.41	0.05	7.4	0.04	7.41
cls_km_paco2_t0	6.65	30.98	7.58	31.92	6.64	31.03	7.45	31.57	6.53	31.11	7.25	32.06
cls_km_paco2_t30	8.45	31.41	7.3	32.85	11.29	35.28	9.81	34.52	4.58	34.4	5.31	33.37
cls_km_pao2_t0	31.67	91.87	29.32	92.85	32.66	93.9	30.08	92.8	31.38	92.02	27.16	92.12
cls_km_pao2_t30	35.4	83.33	35.36	85.79	52.99	90.4	43.55	98.43	15.58	84.84	22.04	87.89
cls_km_hco3_t0	4.8	18.65	5.45	20.39	4.69	18.63	5.58	19.79	4.67	18.83	5.19	20.36
cls_km_hco3_t30	5.03	19.91	4.8	21.37	9.66	19.95	8.13	20.36	3.37	21.34	4.05	21.44
cls_km_be_t0	5.59	-5.79	6.59	-4.22	5.73	-5.77	5.97	-4.53	5.36	-5.58	5.51	-4.06
cls_km_be_t30	13.81	-11.21	16.67	-6.08	24.9	-7.12	26	-8.62	4.63	-4.78	12.1	-4.86
cls_km_pf_t0	121.39	373.66	116.34	316.58	117.59	368.13	126.45	313.28	83.03	361.4	89.52	320.14
cls_km_pf_t30	97.31	297.12	94.45	275.82	151.83	237.11	136.22	251.68	51.74	279.7	74.82	286.21
cls_km_lac_t0	1.45	2.42	1.59	1.87	1.47	2.34	1.59	1.81	1.43	2.41	1.54	1.8
cls_km_lac_t30	0.68	1.47	0.76	1.01	1.43	2.29	1.3	1.54	0.41	1.43	0.61	1.12
dt_dich_bilan_t24	1356.98	2136.61	1527.06	2154.2	1303.46	2130.74	1562.94	2111.07	1245.15	2103.33	1485.23	2126.93
dt_dich_bilan_t48	1595.6	1359.27	1457.58	852.39	1669.06	1331.23	1571.78	902.38	1529.73	1376.24	1365.44	878.96
dt_dich_bilan_t72	1321.63	808.87	1213.21	694.09	1418.82	649.8	1236.32	781.37	1179.01	807.4	1005.94	693.49
dt_nhin_ngay	1.38	1.7	1.66	1.45	1.39	1.68	1.62	1.5	1.37	1.71	1.61	1.52
dt_pex_lan	0.35	1.11	0.37	1.34	0.35	1.11	0.48	1.71	0.35	1.11	0.17	1.44
dt_pex_sauvv	9.14	14.72	8.59	13.01	9.44	14.86	8.67	9.31	8.72	15.35	3.01	16.02
dt_pex_tri_t_lan1	28.98	33.98	19.16	18.14	28.09	32.78	20.6	23.81	27.97	32.69	9.28	29.41
dt_pex_tri_s_lan1	10.54	8.87	12.48	7.93	10.44	8.54	17.72	19.79	10.21	8.92	5.12	15.36
dt_pex_chol_t_lan1	7.33	16.13	4.88	12.41	8.42	16.89	10.46	19.14	7.11	15.99	3.66	15.2
dt_pex_chol_s_lan1	3.56	6.81	3.3	7.1	5.9	8.63	6.42	12.56	3.06	6.8	1.54	8.75
dt_pex_apache_t_lan1	3.19	4.06	2.69	2.68	3.11	3.9	3	2.8	3.02	3.99	1.63	3.72
dt_pex_apache_s_lan1	2	1.92	2.13	0.05	2.14	2.02	2.85	2.54	1.94	1.91	0.85	2.16
dt_pex_imrie_t_lan1	0.81	1.5	0.72	1.14	0.84	1.51	0.94	0.7	0.8	1.51	0.39	1.38
dt_pex_imrie_s_lan1	0.73	0.74	0.68	0.34	0.76	0.78	0.75	0.48	0.71	0.73	0.25	0.8
dt_pex_sofa_t_lan1	1.66	1.39	1.53	1.74	1.71	1.48	1.68	1.64	1.62	1.38	0.92	1.69
dt_pex_sofa_s_lan1	1.55	1.27	1.56	0.99	1.54	1.17	2.12	2.56	1.45	1.18	0.61	1.33
dt_pex_alob_t_lan1	7.56	19.15	7.16	15.54	10.08	20.15	13.55	18.51	6.53	19.9	3.03	19.24
dt_pex_alob_s_lan1	6.01	15.42	4.72	11.17	8.16	18.62	9.61	16.38	4.24	16.66	1.63	16.89

Observations:

- It can be observed from the table above that there are many predictors where the distribution of two classes differ substantially. These are ts_ruou_nam, cls_sh_cre_t0, cls_sh_tri_t0, cls_sh_lip_t0, cls_km_be_t30, dt_pex_tri_t_lan1, dt_pex_chol_t_lan1, etc.
- It can also be observed that for most of the variables, the results are approximately similar for all three datasets.
- For few variables like, dt_pex_imrie_s_lan1, dt_pex_tri_t_lan1, dt_dich_bilan_t72, etc., the standard deviations are quite high representing high variance in the data while for some it is really low.
- Also, it must be noted that all values are at different scales because of different units of measurements.

Since, the datasets do not show a significant distinction between the two classes for many variables, it is expected that the accuracy of the model in detecting the class of patient treatment would not be perfectly accurate.

Satyam Vatts
Avneet Kaur

Categorical Variables:

Variable	Data Values	AMELIA IMPUTED DATA		MICE IMPUTED DATA		MISSFOREST IMPUTED DATA	
		PEX	TRAD	PEX	TRAD	PEX	TRAD
Gender	Nam	0.64	0.73	0.64	0.73	0.64	0.73
	Nu	0.36	0.27	0.36	0.27	0.36	0.27
ts_giadinh	co	0.32	0.25	0.32	0.25	0.32	0.25
	khong	0.68	0.75	0.68	0.75	0.68	0.75
ts_ruou	Co	0.37	0.5	0.37	0.5	0.37	0.5
	Khong	0.63	0.5	0.63	0.5	0.63	0.5
ts_dtd	Co	0.15	0.24	0.15	0.24	0.15	0.24
	Khong	0.85	0.76	0.85	0.76	0.85	0.76
ts_vtc	Co	0.53	0.45	0.53	0.45	0.53	0.45
	Khong	0.47	0.55	0.47	0.55	0.47	0.55
non	0	0.57	0.68	0.57	0.68	0.57	0.68
	1	0.43	0.32	0.43	0.32	0.43	0.32
cls_sa_dichob_t0	Co	0.65	0.68	0.65	0.68	0.65	0.68
	Khong	0.35	0.32	0.35	0.32	0.35	0.32
cls_sa_mat_t0	co	0.94	0.95	0.59	0.76	0.99	0.99
	khong	0.06	0.05	0.41	0.24	0.01	0.01
cls_ct_dichob_lan1	co	0.72	0.58	0.63	0.52	0.84	0.63
	khong	0.28	0.42	0.37	0.48	0.16	0.37
kq	0	0.22	0.18	0.33	0.26	0.16	0.08
	1	0.78	0.82	0.67	0.74	0.84	0.92
bcxa	0	0.02	1	0.02	1	0.02	1
	1	0.98	0	0.98	0	0.98	0

Observations:

- It can be observed that the distribution of categories of categorical variables for two classes of response variable is mostly similar for variables. It means that the distribution of data is even for most of variables between the classes and thus, variables are not significant in detecting distinction between two classes accurately.
- It can be seen that onlyb cls_ct_dichob_lan1 vriable shows substantial difference between distribution.

DISCRIMINANT ANALYSIS

Discriminant Analysis is a statistical technique to develop prediction model for a categorical response based on numerical predictors. Generally, Linear DA is the technique used for most cases but there also exists Quadratic DA that utilizes a quadratic combination of numerical predictors to develop the classification model. The accuracy of this model can be measure using confusion matrix to determine the number of misclassified predictions.

However, the use of DA requires certain assumptions to be satisfied by the dataset. These assumptions are analysed and the results are described below.

Data Preparation

The Discriminant Analysis allows only numerical predictors to be used for developing classification model. Thus, the categorical variables are eliminated from each of the three datasets. After removal of these variables the dataset structure for the three datasets is:

```
[1] "For MiCE/Amelia/Missforest Imputed Data: Distribution of Patients between two Treatements"
[1] "0: Traditional Treatment   1: PEX Treatment"

 0  1
84 81
[1] "Number of Variables in MICE Dataset :  75"
[1] "Number of Variables in Amelia Dataset :  76"
[1] "Number of Variables in MissForest Dataset :  76"
```

It can be observed that the number of variables in each dataset is smaller than the number of observations in each class of response variable. Thus, it is safe to apply DA based on data structure and distribution of classes.

Checking Assumptions

The Discriminant Analysis is valid and provides quality results if the assumptions about the dataset are met. These assumptions must be verified before running the DA. Since for this analysis there are three datasets (MICE, AMELIA & MISSFOREST Imputation), we have tested the assumptions on all these datasets. The results of these validations are given below:

MULTIVARIATE NORMALITY:

Here we are performing Shapiro-Wilik test to check the normality of variables in all three datasets. The results of the test are summarised below:

Satyam Vatts
Avneet Kaur

Variable	AMELIA		MISSFOREST		MICE	
	Test_Statistics	P_Value	Test_Statistics	P_Value	Test_Statistics	P_Value
Age	0.975	0.004227	0.975	0.004227	0.975	0.004227
rv_ngaydt	0.944	4.25E-06	0.944	4.25E-06	0.944	4.25E-06
ts_ruou_nam	0.745	1.27E-15	0.745	1.27E-15	0.745	1.27E-15
ls_tt_bmi_t0	0.977	0.007578	0.976	0.005916	0.977	0.008049
ls_tn_mach_t0	0.975	0.003863	0.976	0.005941	0.973	0.002529
ls_tn_nhiet_t0	0.17	1.49E-26	0.161	1.15E-26	0.165	1.29E-26
ls_tn_spo2_t0	0.789	3.78E-14	0.783	2.34E-14	0.78	1.87E-14
ls_diem_apache_t0	0.925	1.57E-07	0.911	1.86E-08	0.912	1.93E-08
ls_diem_ranson_t0	0.926	1.86E-07	0.905	7.12E-09	0.897	2.68E-09
ls_diem_ct_t0	0.916	3.52E-08	0.86	2.91E-11	0.903	5.95E-09
ls_diem_imrie_t0	0.908	1.13E-08	0.895	1.95E-09	0.883	4.24E-10
ls_diem_sofa_t0	0.832	1.69E-12	0.845	5.97E-12	0.841	3.97E-12
cls_hh_bc_t0	0.984	0.049087	0.983	0.040536	0.984	0.054764
cls_hh_bc_t30	0.978	0.008719	0.963	0.000227	0.99	0.266635
cls_hh_bc_t54	0.986	0.095707	0.949	1.03E-05	0.955	4.00E-05
cls_hh_bc_t72	0.995	0.859962	0.953	2.54E-05	0.967	0.000587
cls_hh_hct_t0	0.988	0.168324	0.987	0.140483	0.988	0.171415
cls_hh_hct_t30	0.993	0.560319	0.971	0.001463	0.981	0.022295
cls_hh_hct_t72	0.992	0.459538	0.975	0.004466	0.892	1.28E-09
cls_hh_pt_t0	0.985	0.070683	0.979	0.01172	0.979	0.011642
cls_hh_pt_t30	0.992	0.455019	0.964	0.000279	0.952	1.84E-05
cls_hh_pt_t72	0.989	0.231021	0.927	2.00E-07	0.873	1.26E-10
cls_hh_aptt_t0	0.589	1.14E-19	0.492	1.29E-21	0.506	2.28E-21
cls_hh_aptt_t30	0.796	6.99E-14	0.645	2.26E-18	0.695	4.37E-17
cls_hh_fib_t0	0.986	0.091103	0.99	0.313072	0.986	0.096242
cls_hh_fib_t30	0.989	0.248142	0.937	1.10E-06	0.958	6.97E-05
cls_sh_ure_t0	0.425	8.15E-23	0.415	5.55E-23	0.421	7.07E-23
cls_sh_ure_t30	0.949	1.12E-05	0.848	8.32E-12	0.892	1.23E-09
cls_sh_ure_t72	0.849	9.22E-12	0.485	9.28E-22	0.627	8.17E-19
cls_sh_cre_t0	0.581	7.67E-20	0.557	2.41E-20	0.578	6.56E-20
cls_sh_cre_t30	0.784	2.58E-14	0.589	1.15E-19	0.739	8.09E-16
cls_sh_glu_t0	0.795	6.00E-14	0.686	2.49E-17	0.726	3.29E-16
cls_sh_chol_t0	0.942	2.84E-06	0.925	1.39E-07	0.944	3.82E-06
cls_sh_chol_t30	0.962	0.000167	0.82	5.29E-13		
cls_sh_tri_t0	0.67	9.79E-18	0.656	4.18E-18	0.669	9.02E-18
cls_sh_tri_t30	0.723	2.75E-16	0.499	1.68E-21	0.561	2.91E-20
cls_sh_amy_t0	0.94	2.02E-06	0.849	9.10E-12	0.876	1.93E-10
cls_sh_lip_t0	0.958	6.65E-05	0.935	7.46E-07	0.933	5.50E-07
cls_sh_pro_t0	0.987	0.144515	0.971	0.001523	0.984	0.056891
cls_sh_na_t0	0.967	0.000508	0.961	0.000139	0.936	1.04E-06
cls_sh_na_t30	0.988	0.152655	0.965	0.000363	0.96	0.000124
cls_sh_ka_t0	0.963	0.000203	0.963	0.000202	0.969	0.000913
cls_sh_ka_t30	0.941	2.31E-06	0.882	3.72E-10	0.884	4.83E-10
cls_km_ph_t0	0.961	0.000153	0.952	2.23E-05	0.964	0.000267
cls_km_ph_t30	0.995	0.884079	0.889	8.61E-10	0.961	0.000129
cls_km_paco2_t0	0.979	0.012624	0.972	0.001832	0.977	0.007572
cls_km_paco2_t30	0.955	4.12E-05	0.927	1.95E-07	0.943	3.58E-06
cls_km_pao2_t0	0.869	7.79E-11	0.857	2.14E-11	0.88	2.89E-10
cls_km_pao2_t30	0.955	3.49E-05	0.74	9.16E-16	0.972	0.002114
cls_km_hco3_t0	0.985	0.068671	0.974	0.00302	0.979	0.011647
cls_km_hco3_t30	0.99	0.320068	0.94	1.97E-06	0.902	5.30E-09
cls_km_be_t0	0.963	0.0002	0.974	0.003153	0.976	0.005608
cls_km_be_t30	0.891	1.19E-09	0.441	1.51E-22	0.55	1.74E-20
cls_km_pf_t0	0.994	0.725902	0.991	0.433841	0.971	0.001472
cls_km_pf_t30	0.98	0.015283	0.9	4.01E-09	0.827	1.02E-12
cls_km_lac_t0	0.842	4.39E-12	0.804	1.35E-13	0.809	2.11E-13
cls_km_lac_t30	0.945	5.53E-06	0.798	7.95E-14	0.866	5.58E-11
dt_dich_bilan_t24	0.987	0.113898	0.977	0.007423	0.981	0.023328
dt_dich_bilan_t48	0.983	0.041331	0.967	0.000551	0.986	0.096298
dt_dich_bilan_t72	0.947	7.81E-06	0.923	1.07E-07	0.941	2.46E-06
dt_nhin_ngay	0.729	4.24E-16	0.699	5.68E-17	0.699	5.64E-17
dt_pex_lan	0.824	7.87E-13	0.798	7.93E-14	0.658	4.86E-18
dt_pex_sauvv	0.979	0.011906	0.949	1.02E-05	0.805	1.49E-13
dt_pex_tri_t_lan1	0.843	5.15E-12	0.732	4.97E-16	0.69	3.24E-17
dt_pex_tri_s_lan1	0.867	6.76E-11	0.764	5.05E-15	0.75	1.88E-15
dt_pex_chol_t_lan1	0.949	1.16E-05	0.933	5.41E-07	0.932	5.08E-07
dt_pex_chol_s_lan1	0.955	4.19E-05	0.909	1.27E-08	0.889	8.67E-10

Satyam Vatts
Avneet Kaur

dt_pex_apache_t_lan1	0.96	0.000124	0.931	4.26E-07	0.893	1.57E-09
dt_pex_apache_s_lan1	0.974	0.003287	0.882	3.67E-10	0.834	2.12E-12
dt_pex_imrie_t_lan1	0.948	9.31E-06	0.908	1.09E-08	0.859	2.56E-11
dt_pex_imrie_s_lan1	0.952	1.88E-05	0.868	6.99E-11	0.751	2.00E-15
dt_pex_sofa_t_lan1	0.913	2.43E-08	0.908	1.23E-08	0.839	3.53E-12
dt_pex_sofa_s_lan1	0.884	4.93E-10	0.816	3.67E-13	0.852	1.29E-11
dt_pex_alob_t_lan1	0.985	0.063729	0.896	2.25E-09	0.861	3.20E-11

Observations:

- It can be observed that majority of vairables for all the three datasets do not follow normal distribution.
- For MICE datatset , 19 vairables follow normality while the rest have p-value less than 0.05.
- For Amelia and Missforest, only 3 and 4 columns respectively follow normality. For the rest the p-value for Shapiro-Wilik Test is less than 0.05 and hence the null hypothesis of normlity gets rejected.

Although, the majority of data is not normal, but it is not fatal for Discriminant Analysis. Although, it will affect the predictions and accuracy of model but the severeness is not high.

Multi-Collinearity:

The Multicollinearity among the predictors adversely affect the results of discriminant analysis. DA function coefficients fail to reliably predict group membership if high correlation exists between variables. Thus, the pooled within-groups correlation matrix is used to detect multicollinearity. The columns where the correlation is greater than 0.8 for all the three datasets is listed below:

AMELIA		MICE		MISSFOR	
related_cols	**cor_coeff**	**related_cols**	**cor_coeff**	**related_cols**	**cor_coeff**
ls_diem_apache_t0 , dt_pex_apache_t_lan1	0.89	ls_diem_apache_t0 , dt_pex_apache_t_lan1	0.85	ls_diem_apache_t0 , dt_pex_apache_t_lan1	0.88
ls_diem_ranson_t0 , ls_diem_imrie_t0	0.85	ls_diem_ranson_t0 , ls_diem_imrie_t0	0.84	ls_diem_ranson_t0 , ls_diem_imrie_t0	0.85
ls_diem_ranson_t0 , dt_pex_imrie_t_lan1	0.8	ls_diem_sofa_t0 , dt_pex_sofa_t_lan1	0.84	ls_diem_imrie_t0 , dt_pex_imrie_t_lan1	0.83
ls_diem_imrie_t0 , dt_pex_imrie_t_lan1	0.87	cls_sh_tri_t0 , dt_pex_tri_t_lan1	0.81	ls_diem_sofa_t0 , dt_pex_sofa_t_lan1	0.87
ls_diem_sofa_t0 , dt_pex_sofa_t_lan1	0.89	cls_km_paco2_t0 , cls_km_hco3_t0	0.81	cls_sh_chol_t0 , dt_pex_chol_t_lan1	0.81
cls_sh_tri_t0 , dt_pex_tri_t_lan1	0.81			cls_sh_tri_t0 , dt_pex_tri_t_lan1	0.81
cls_km_paco2_t0 , cls_km_hco3_t0	0.85			cls_km_paco2_t0 , cls_km_hco3_t0	0.85
cls_km_hco3_t0 , cls_km_be_t0	0.83			cls_km_hco3_t0 , cls_km_be_t0	0.81

Observations & Actions Taken:

1. Since Ranson and IMRIE scores are almost similar, we can drop one of them. For our analysis, we are dropiing the IMRIE Score.

2. km_paco2 and km_hco3 have high correlation coz they provide info about blood gas so we can remove t0 and t30 of one of these. Here we are removing hco3 from datasets. This also eliminates the correlation between km_be & km_hco3.

3. Others have high correlation since they are the tests being taken for both pex and traditional treatment and have similar scores. Thus, to comply to multicorrelation assumption we are going to drop the vairables corresponding to traditional treatment

After removing the columns that are causing multi-colinearity, the data structure of the three datasets is as follows:

```
[1] "Number of Variables in MICE Dataset :  69"
[1] "Number of Variables in Amelia Dataset :  69"
[1] "Number of Variables in MissForest Dataset :  69"
```

Equality of Variance within Groups:

The DA requires equality of variance-covariance of predictors within groups i.e., the covariance matrix within each group should be equal. Here we have used Levene's Test to determine the equality of variance among the predictors of PEX treated observations & Traditionally Treated observations. The following table describes the result of the test on the three datasets.

Variable	MICE		AMELIA		MISSFOR	
	F_Value	**P_Value**	**F_Value**	**P_Value**	**F_Value**	**P_Value**
Age	0.195	0.659	0.195	0.659	0.195	0.659
rv_ngaydt	3.035	0.083	3.035	0.083	3.035	0.083
ts_ruou_nam	12.321	0.001	12.321	0.001	12.321	0.001
ls_tt_bmi_t0	0.007	0.932	0.093	0.761	0.003	0.957
ls_tn_mach_t0	0.144	0.705	0.121	0.729	0.116	0.734
ls_tn_nhiet_t0	1.51	0.221	1.543	0.216	1.682	0.196
ls_tn_spo2_t0	4.367	0.038	3.69	0.056	4.539	0.035
ls_diem_ranson_t0	2.265	0.134	0.584	0.446	0.708	0.401
ls_diem_ct_t0	1.11	0.294	9.728	0.002	0.001	0.971
cls_hh_bc_t0	9.162	0.003	0.001	0.972	9.027	0.003
cls_hh_bc_t30	0.06	0.807	0.069	0.793	3.63	0.059
cls_hh_bc_t54	0.589	0.444	0.28	0.598	0.789	0.376
cls_hh_bc_t72	0.742	0.39	2.412	0.122	3.323	0.07
cls_hh_hct_t0	2.13	0.146	1.046	0.308	1.707	0.193
cls_hh_hct_t30	1.873	0.173	0.005	0.947	5.432	0.021
cls_hh_hct_t72	6.873	0.01	1.57	0.212	3.636	0.058

cls_hh_pt_t0	3.936	0.049	0.057	0.812	1.537	0.217
cls_hh_pt_t30	3.58	0.06	0.134	0.714	1.632	0.203
cls_hh_pt_t72	0.016	0.899	2.345	0.128	0.126	0.723
cls_hh_aptt_t0	2.397	0.123	1.24	0.267	1.656	0.2
cls_hh_aptt_t30	6.668	0.011	0.012	0.913	0.638	0.426
cls_hh_fib_t0	0.00E+00	0.993	1.256	0.264	0.045	0.832
cls_hh_fib_t30	3.177	0.077	0.213	0.645	4.891	0.028
cls_sh_ure_t0	0.211	0.647	0.149	0.7	0.123	0.726
cls_sh_ure_t30	0.335	0.564	0.285	0.594	13.051	0
cls_sh_ure_t72	0.072	0.789	1.652	0.2	0.264	0.608
cls_sh_cre_t0	2.016	0.158	0.014	0.906	2.21	0.139
cls_sh_cre_t30	1.047	0.308	0.096	0.757	5.728	0.018
cls_sh_glu_t0	0.492	0.484	2.473	0.118	0.76	0.384
cls_sh_chol_t0	1.151	0.285	0.168	0.682	3.332	0.07
cls_sh_tri_t30	1.273	0.261	0.715	0.399	0.108	0.743
cls_sh_amy_t0	0.629	0.429	4.207	0.042	4.779	0.03
cls_sh_lip_t0	1.944	0.165	0.122	0.727	1.228	0.269
cls_sh_pro_t0	0.756	0.386	0.815	0.368	7.095	0.009
cls_sh_na_t0	0.001	0.982	0.003	0.958	0.01	0.921
cls_sh_na_t30	3.008	0.085	1.609	0.206	17.419	0.00E+00
cls_sh_ka_t0	0.644	0.423	1.374	0.243	1.48	0.226
cls_sh_ka_t30	2.719	0.101	0.021	0.884	10.265	0.002
cls_km_ph_t0	0.162	0.688	0.563	0.454	0.015	0.903
cls_km_ph_t30	4.536	0.035	1.426	0.234	0.143	0.706
cls_km_paco2_t0	0.635	0.427	1.222	0.271	0.378	0.539
cls_km_paco2_t30	3.495	0.063	0.055	0.815	2.766	0.098
cls_km_pao2_t0	0.061	0.805	0.023	0.879	0.291	0.59
cls_km_pao2_t30	5.779	0.017	0.116	0.734	2.358	0.127
cls_km_be_t0	0.118	0.731	0.117	0.733	0.126	0.723
cls_km_be_t30	0.144	0.705	0.222	0.638	0.536	0.465
cls_km_pf_t0	0.857	0.356	0.005	0.943	0.596	0.441
cls_km_pf_t30	0.002	0.961	0.204	0.652	4.082	0.045
cls_km_lac_t0	1.459	0.229	0.77	0.382	1.654	0.2
cls_km_lac_t30	2.5	0.116	0.117	0.733	1.022	0.314
dt_dich_bilan_t24	1.408	0.237	0.444	0.506	0.592	0.443
dt_dich_bilan_t48	0.327	0.568	0.502	0.48	1.202	0.275
dt_dich_bilan_t72	0.175	0.676	0.039	0.845	0.693	0.406
dt_nhin_ngay	0.041	0.84	0.004	0.947	0.072	0.788
dt_pex_lan	9.466	0.002	14.801	0	0.007	0.932
dt_pex_sauvv	5.242	0.023	0.302	0.584	43.096	0.00E+00
dt_pex_tri_t_lan1	3.091	0.081	1.444	0.231	12.678	0
dt_pex_tri_s_lan1	29.031	0.00E+00	10.501	0.001	1.708	0.193
dt_pex_chol_t_lan1	8.49	0.004	8.449	0.004	17.85	0.00E+00
dt_pex_chol_s_lan1	0.012	0.913	0.207	0.65	9.902	0.002
dt_pex_apache_t_lan1	0.038	0.845	1.25	0.265	20.391	0.00E+00
dt_pex_apache_s_lan1	10.196	0.002	0.642	0.424	30.33	0.00E+00
dt_pex_imrie_t_lan1	0.007	0.932	0.662	0.417	21.813	0.00E+00
dt_pex_imrie_s_lan1	1.329	0.251	0.453	0.502	44.902	0.00E+00
dt_pex_sofa_t_lan1	0.095	0.758	0.337	0.562	10.904	0.001
dt_pex_sofa_s_lan1	12.973	0	0.006	0.936	14.602	0
dt_pex_alob_t_lan1	2.168	0.143	0	0.983	20.166	0.00E+00
dt_pex_alob_s_lan1	0.746	0.389	2.864	0.092	18.37	0.00E+00

Observations:

- The above table highlighted columns are the ones where the groups do not have equal variances. The p-value of Levene's Test is less than 0.05 indicating rejection of null hypothesis of equal variances.
- It can also be observed that almost all PEX treatment variables for Missforest data show unequal variance between the groups. Also, there are many variables where the results of the test differ between the datasets.

Thus, it can be observed that the Vairance Equality assumption is not being followed by many variables that are being taken as predictors, hence the results can be a bit different for the three datasets.

Building Classification Model

The Discriminant Analysis detects the direction that maximizes the separation between the response variable classes. These directions are the linear combinations of predictor variables also called as Linear Discriminants. These directions are then utilized to predict the class of a new patient whether treated with PEX or Traditional treatment.

The model has been developed using a training dataset obtained by splitting the whole dataset into a ratio of 7:3. The LDA model developed in R is described below for all the datasets.

Satyam Vatts
Avneet Kaur

Prior Probabilities: These represent the proportions of pex and traditionally treated patients in the training set.

	mice_lda$prior <dbl>	amelia_lda$prior <dbl>	missfor_lda$prior <dbl>
0	0.4661017	0.5084746	0.4912281
1	0.5338983	0.4915254	0.5087719

It can be observed that distribution of patients among two treatments is more even in case of Amelia and Missforest dataset with almost equal division while for Mice dataset it is a bit uneven.

Coefficient & Group Means:

Vairable	MICE DATASET			AMELIA DATASET			MISSFOREST DATASET		
	Coefficients	Group Means			Group Means			Group Means	
		Traditional	PEX	Coefficients	Traditional	PEX	Coefficients	Traditional	PEX
Age	0.054	41.418	39.873	0.009	42.35	40.207	0.006	42.393	40.483
rv_ngaydt	-0.053	6.909	7.048	0.178	6.333	7.103	0.178	6.643	6.828
ts_ruou_nam	-0.111	6.909	3.714	0.019	7.65	2.569	-0.107	6.571	3.069
ls_tt_bmi_t0	0.034	22.637	22.662	-0.177	22.504	22.628	0.262	22.531	22.764
ls_tn_mach_t0	-0.01	108.109	108.048	-0.037	107.5	107.454	-0.042	107.041	105.118
ls_tn_nhiet_t0	0.185	36.667	37.133	-0.802	37.228	37.137	0.376	36.628	37.098
ls_tn_spo2_t0	0.1	96.709	97.175	-0.131	96.642	97.404	-0.055	96.815	97.461
ls_diem_ranson_t0	0.216	1.364	1.619	0.579	3.547	3.874	1.569	1.381	1.545
ls_diem_ct_t0	-0.042	3.636	4.254	0.207	9.546	11.151	0.14	3.412	3.995
cls_hh_bc_t0	0.097	9.699	11.557	-0.202	8.813	8.958	0.178	10.024	11.513
cls_hh_bc_t30	0.105	9.086	8.28	-0.1	9.406	9.94	-0.102	8.975	9.146
cls_hh_bc_t54	-0.056	9.951	8.853	0.078	10.102	10.747	-0.071	9.338	10.041
cls_hh_bc_t72	0.063	10.321	11.947	2.731	0.387	0.392	0.531	10.543	11.138
cls_hh_hct_t0	-9.099	0.384	0.384	4.102	0.348	0.325	-11.609	0.382	0.385
cls_hh_hct_t30	3.126	0.346	0.343	-15.162	0.339	0.303	9.23	0.347	0.332
cls_hh_hct_t72	-1.396	0.338	0.324	-0.024	85.014	86.664	-42.617	0.327	0.312
cls_hh_pt_t0	-0.025	85.665	85.022	0.007	82.219	78.794	-0.04	85.507	87.682
cls_hh_pt_t30	0.012	80.765	78.775	0.007	81.243	79.01	0.031	81.385	78.595
cls_hh_pt_t72	-0.011	91	81.54	-0.247	1.129	1.216	0.044	82.44	82.515
cls_hh_aptt_t0	0.252	1.092	1.221	0.614	1.125	1.212	0.267	1.092	1.208
cls_hh_aptt_t30	-0.449	1.19	1.35	-0.185	6.196	5.009	0.805	1.148	1.236
cls_hh_fib_t0	-0.197	6.219	5.296	0.276	6.841	5.714	-0.398	5.96	5.19
cls_hh_fib_t30	-0.24	6.811	5.366	0.038	5.684	6.071	-0.403	6.58	5.96
cls_sh_ure_t0	-0.031	6.038	4.94	-0.393	4.977	3.943	-0.094	5.355	5.871
cls_sh_ure_t30	0.12	5.129	4.216	0.219	5.412	6.871	-0.16	4.496	4.124
cls_sh_ure_t72	0.018	6.011	7.141	-0.001	93.572	71.291	0.006	5.151	5.275
cls_sh_cre_t0	-0.001	100.989	73.852	-0.007	82.854	67.019	0.006	90.729	71.333
cls_sh_cre_t30	0.013	79.164	69.349	0.037	11.465	11.562	0.032	73.793	68.988
cls_sh_glu_t0	-0.003	11.664	11.736	-0.038	12.901	14.967	0.125	11.281	11.742
cls_sh_chol_t0	0.117	13.948	16.839	-0.059	7.045	6.904	0.403	7.558	7.619
cls_sh_tri_t30	0.005	12.332	12.62	0.052	7.54	8.241	-0.101	8.041	10.13
cls_sh_amy_t0	-0.001	434.204	449.708	-0.001	336.699	407.396	0.004	347.981	464.227
cls_sh_lip_t0	0.001	450.958	660.74	0.001	381.618	570.56	-0.001	480.839	585.638
cls_sh_pro_t0	0.02	59.589	59.892	0.048	58.228	59.689	0.057	58.996	58.617
cls_sh_na_t0	0.024	132.073	131.238	-0.05	132.684	131.522	-0.032	133.054	131.53
cls_sh_na_t30	0.069	135.455	136.905	0.089	135.184	136.417	0.103	135.568	137.584
cls_sh_ka_t0	0.265	3.555	3.724	0.25	3.653	3.695	-0.436	3.59	3.678
cls_sh_ka_t30	-1.099	3.456	3.493	0.45	3.432	3.437	0.042	3.439	3.438
cls_km_ph_t0	-11.102	7.407	7.373	-6.243	7.404	7.378	-18.499	7.411	7.38
cls_km_ph_t30	3.258	7.419	7.413	11.531	7.417	7.4	14.504	7.415	7.403
cls_km_paco2_t0	-0.065	32.231	30.856	0.015	32.549	31.634	-0.166	31.911	31.543
cls_km_paco2_t30	0.059	34.847	35.363	0.055	32.979	31.33	0.162	33.508	34.623
cls_km_pao2_t0	-0.005	90.787	95.341	-0.01	90.29	89.491	-0.004	94.384	87.736
cls_km_pao2_t30	-0.015	96.482	89.184	0.019	85.931	81.811	-0.031	89.278	84.632
cls_km_be_t0	-0.062	-4	-6.075	-0.029	-4.073	-5.078	0.026	-2.998	-5.174
cls_km_be_t30	-0.008	-8.493	-8.152	-0.094	-3.804	-11.901	-0.158	-3.616	-4.291
cls_km_pf_t0	-0.001	312.28	383.679	0.003	289.807	367.137	0.002	319.789	352.476
cls_km_pf_t30	0	254.545	250.841	-0.005	264.898	288.794	-0.015	280.973	270.893
cls_km_lac_t0	-0.118	1.844	2.373	-0.048	1.942	2.418	-0.582	1.791	2.539
cls_km_lac_t30	0.247	1.529	2.267	-0.787	1.035	1.463	-1.61	1.127	1.406
dt_dich_bilan_t24	0	2168	2198.889	0	2073.043	2145.962	0	2185.195	2116.91
dt_dich_bilan_t48	0	749.091	1360.159	0.00E+00	808.056	1492.24	-0.001	885.502	1309.791
dt_dich_bilan_t72	0	838.636	622.254	0	891.753	993.003	0.001	659.108	962.277
dt_nhin_ngay	0.135	1.436	1.73	0.249	1.25	1.755	0.144	1.627	1.774
dt_pex_lan	-2.464	1.727	1.127	-2.631	1.337	1.12	-3.379	1.434	1.109

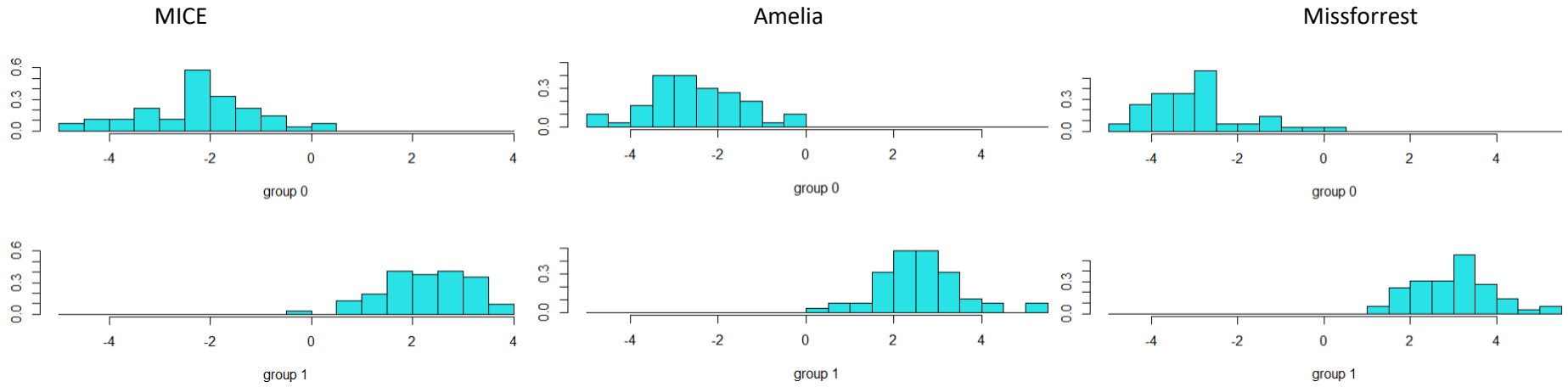
Satyam Vatts
Avneet Kaur

dt_pex_sauvv	0.02	8.642	14.428	-0.038	13.183	14.916	0.003	15.609	15.167
dt_pex_tri_t_lan1	-0.001	24.363	34.503	0.004	17.072	29.792	0.023	28.977	32.533
dt_pex_tri_s_lan1	0.007	20.071	9.087	0.025	7.334	7.768	-0.114	15.441	7.509
dt_pex_chol_t_lan1	-0.056	18.849	17.402	0.121	12.451	15.187	0.001	14.849	15.756
dt_pex_chol_s_lan1	-0.009	12.251	8.754	-0.041	7.204	6.614	-0.358	8.74	6.5
dt_pex_apache_t_lan1	0.07	2.957	4.01	-0.064	2.727	3.863	0.257	3.757	4.013
dt_pex_apache_s_lan1	-0.253	2.527	2.175	0.46	-0.002	1.893	-1.082	2.092	1.83
dt_pex_imrie_t_lan1	0.179	0.727	1.476	-0.88	1.159	1.365	-1.545	1.417	1.488
dt_pex_imrie_s_lan1	0.156	0.527	0.778	1.348	0.346	0.693	1.58	0.794	0.749
dt_pex_sofa_t_lan1	0.068	1.709	1.286	-0.023	1.713	1.413	-0.378	1.683	1.432
dt_pex_sofa_s_lan1	-0.131	2.673	1.063	-0.363	1.127	1.268	0.248	1.314	1.162
dt_pex_alob_t_lan1	0.027	18.518	19.754	0.109	16.362	19.176	0.093	19.38	19.775
dt_pex_alob_s_lan1	0.032	17.055	19.04	0.096	11.07	15.21	0.149	16.819	15.795

Observations:

- The above table provides the coefficients of all the predictors of response variable for Linear Discriminant Analysis. It can be observed from the coefficients that variables dt_dich_bilan and cls_km_pf_t30 have 0 as coefficients showing that they do not contribute towards the LDA function.
- It can also be observed that group means for almost all the variables are similar. This indicates that each variable contributes a small amount towards predicting the class of target variable
- The LDA function can be written as (for MICE dataset):
  $LD = 0.054\ X_1 - 0.053\ X_2 - 0.111\ X_3 \ldots$
  where X1 : Age; X2: rv_ngayat etc.

Model Accuracy



It can be observed from the histograms above that some obervations are missclassifies in case of MICE and MISSFOREST datasets while for AMELIA, the distinction between the group 0 and 1 is clear.

To quantify the model accuracy, the predictions are performed on Test Dataset for all the three datatsets using respective LDA models. The results of the predictions are described below:

Patient Number	Actual Treatment	MICE Dataset			AMELIA Dataset		
		Predicted	Posterior Probab.		Predicted	Posterior Probab.	
			0	1		0	1
1	1	1	0	1	1	0	1
5	0	0	0.72	0.28	1	0.06	0.94
7	1	1	0	1	1	0	1
8	1	1	0	1	0	1	0
16	0	0	1	0	0	0.91	0.09
18	0	0	1	0	0	1	0
19	1	0	1	0	0	0.56	0.44
21	0	0	1	0	1	0	1
22	0	0	1	0	0	1	0
30	0	0	1	0	0	0.81	0.19
36	0	1	0.19	0.81	1	0	1
38	0	0	1	0	0	1	0
40	1	1	0	1	1	0	1
42	0	0	1	0	0	0.84	0.16
44	1	1	0	1	1	0	1
47	0	0	0.94	0.06	0	1	0
50	0	0	1	0	1	0	1
53	0	1	0.06	0.94	1	0	1
62	0	1	0	1	0	0.95	0.05
69	1	0	0.93	0.07	0	1	0
71	0	0	1	0	1	0	1

Patient Number	MISSFOREST Dataset		
	Predicted	Posterior Probab.	
		0	1
1	1	0	1
2	1	0.47	0.53
5	1	0	1
11	1	0	1
12	0	1	0
21	0	1	0
24	1	0.31	0.69
26	0	1	0
28	0	1	0
30	0	1	0
34	0	0.98	0.02
37	1	0	1
39	0	0.99	0.01
44	1	0	1
49	0	1	0
55	1	0.3	0.7
58	0	1	0
59	0	0.86	0.14
61	1	0	1
64	1	0	1
74	1	0	1

73	0	1	0	1	1	0.01	0.99
77	0	0	1	0	1	0	1
78	1	1	0	1	0	1	0
80	1	1	0	1	1	0	1
81	0	0	1	0	0	1	0
83	0	0	0.97	0.03	1	0.18	0.82
86	1	1	0	1	1	0	1
88	0	0	1	0	0	1	0
95	1	1	0	1	0	1	0
96	0	0	0.85	0.15	1	0.32	0.68
97	0	0	1	0	1	0	1
100	1	1	0	1	0	0.89	0.11
105	1	1	0	1	1	0	1
110	0	0	1	0	0	1	0
112	1	1	0	1	0	1	0
113	0	0	1	0	1	0	1
118	1	1	0	1	0	1	0
122	0	0	0.78	0.22	0	1	0
126	1	1	0	1	1	0.01	0.99
136	1	1	0	1	1	0	1
140	0	0	1	0	1	0	1
146	0	0	1	0	0	1	0
149	0	1	0.23	0.77	0	1	0
151	1	1	0	1	0	1	0
163	0	0	1	0	1	0	1
165	0	0	1	0	1	0	1

75	0	1	0
76	0	1	0
77	0	1	0
79	0	1	0
81	1	0	1
82	1	0	1
84	0	1	0
85	0	1	0
86	0	1	0
91	1	0	1
99	0	1	0
100	1	0	1
101	1	0	1
102	1	0	1
107	0	0.53	0.47
108	0	1	0
111	0	1	0
119	1	0	1
122	0	1	0
129	0	1	0
137	0	1	0
138	1	0	1
143	1	0	1
144	1	0	1
148	0	1	0
152	0	0.98	0.02
154	0	1	0
155	1	0	1
160	0	1	0
163	0	1	0

```
[1] "MICE DATASET CONFUSION MATRIX"
         Actual
Predicted  0  1
        0 24  2
        1  5 16
[1] "MICE DATA Accuracy :  0.851063829787234"
```

```
[1] "AMELIA DATASET CONFUSION MATRIX"
         Actual
Predicted  0  1
        0 19  4
        1  5 19
[1] "AMELIA DATA Accuracy :  0.808510638297872"
```

```
[1] "MISSFOREST DATASET CONFUSION MATRIX"
         Actual
Predicted  0  1
        0 20  9
        1  8 14
[1] "MISSFOREST DATA Accuracy :  0.666666666666667"
```

Observations:

- It can be observed that MICE Dataset has the highest accuracy among the three dataset with an accuracy of 85.1% for predicting the correct class of patients.
- There are 7 patients that are misclassified in MICE dataset, 9 in Amelia Dataset and 17 in Missforest Dataset.
- Although, many assumptions were not met, the model performs sufficiently good to be used for prediction of new observations.

K NEAREST NEIGHBOUR (KNN) ANALYSIS

KNN algorithm is one of the fundamental supervised machine learning algorithms that can be used both for classification as well as regression predictive problems. To classify a new datapoint, KNN algorithm checks the similarity measures of the previously stored data points where K is the number of neighbours the algorithm used by the algorithm in classification. KNN uses Euclidean distance to calculate the nearest neighbour.

In this report, KNN algorithm is used to predict the variable **'pex**' on three datasets imputed in the previous assignment – **Amelia. Mice** and **Missforest.**

Data Preparation

Since KNN works with numerical data, to handle non – numeric data, factorization has been done to give the categories numerical values. For example, in Gender variable (Nu, Nam) has been converted to (0 , 1). Similarly, all the categorical variables have been assigned **numeric labels** in order to perform KNN. The entire dataset is split into **train** and **test** data in **70: 30** ratios.

Choosing the right K

To choose K, we have followed the below steps:

- **Square root** of the number of observations in the dataset which came out to be :

```
10.723805294763608
```

- **Odd value** of K is usually taken to avoid any confusion between the classes. Hence we took **K** as **10 – 1 = 9.**

Satyam Vatts
Avneet Kaur

KNN on AMELIA Dataset

Confusion Matrix and Accuracy Score for Test Dataset

```
[[12  7]
 [23  8]]

Accuracy score =  0.4
```

True Positive = 12
False Positive= 23 (Type 1 error)
False Negative = 7 (Type 2 error)
True Negative = 8
In our Test data out of 50 observations, only 20 values of 'pex' variable were correctly classified and 30 values were misclassified. Hence the accuracy is 40%

The test model does not perform well and has an accuracy of just 40%. The reason can be seen from the accuracy of model on the Training dataset below. It can be seen the accuracy of Train dataset is itself low (64 %).

Confusion Matrix and Accuracy Score for Training Dataset

```
[[54 11]
 [30 20]]

0.6434782608695652
```

KNN on MICE Dataset

Confusion Matrix and Accuracy Score

```
[[13  6]
 [24  7]]

Accuracy score =  0.4
```

True Positive = 13
False Positive= 24(Type 1 error)
False Negative = 6 (Type 2 error)
True Negative = 7
In our Test data out of 50 observations, only 20 values of 'pex' variable were correctly classified and 30 values were misclassified. Hence the accuracy is 40%

The test model does not perform well and has an accuracy of just 40%. The reason can be seen from the accuracy of model on the Training dataset below. It can be seen the accuracy of Train dataset is itself low (62 %).

Confusion Matrix and Accuracy Score for Training Dataset

```
[[53 12]
 [31 19]]

0.6260869565217392
```

KNN on MISSFOREST Dataset

Confusion Matrix and Accuracy Score

```
[[11  8]
 [23  8]]

Accuracy score =  0.38
```

True Positive = 11
False Positive= 23(Type 1 error)
False Negative = 8 (Type 2 error)
True Negative = 8
In our Test data out of 50 observations, only 19 values of 'pex' variable were correctly classified and 31 values were misclassified. Hence the accuracy is 38%

The test model does not perform well and has an accuracy of just 38%. The reason can be seen from the accuracy of model on the Training dataset below. It can be seen the accuracy of Train dataset is itself low (63 %).

Confusion Matrix and Accuracy Score for Training Dataset

```
[[52 13]
 [29 21]]

0.6347826086956522
```

COPARISON OF KNN & DISCRIMINANT ANALYSIS

The Analysis done so far indicates the following comparison between the three datasets and the prediction models developed on them:

- The Discriminant Analysis showed relatively better results than the KNN model. For MICE dataset the DA achieved an accuracy of 85% while KNN could only get 40% of accuracy. Thus, a lot of patients have been misclassified by KNN model. The reason can be the relationship between the predictors. A linear relationship has proved more insightful rather than clustering of data based on distance from center.
- It can also be observed that MICE dataset has outperformed the imputed data from other techniques i.e., Amelia and Missforest. Thus, for this dataset, MICE imputation can be considered as a better imputation technique.
- It can also be observed that the tendency of misclassification is more for Traditional Treatment than for the PEX treatment. As the number of observations of False Positive i.e., where 0 has been classified as 1 is more than the other.

Summary

In this analysis, we performed Discriminant Analysis and KNN Techniques to develop classification prediction models for detecting the treatment received by Acute Pancreatitis patients. The two treatments that are being classified are Traditional & PEX treatments. In this report, we have provided our findings by performing the following tasks:

- Determining differences between the two groups of patients by analyzing their symptoms, test results and treatments. It is performed by analyzing the mean and standard deviations of numerical variables and percentage distribution of different categories for categorical variables.
- Checking the Assumptions of Multi-Colinearity, Equal Variances within groups and multivariate normality
- Developing LDA prediction model on training dataset after processing the datasets and carrying out predictions on test dataset.
- Evaluating Accuracy of the LDA prediction using confusion matrices.
- Developing KNN prediction model on training dataset after processing categorical variables as factors and carrying out predictions on test dataset.
- Evaluating Accuracy of the KNN prediction using confusion matrices.
- Comparing the results of Discriminant Analysis and KNN Analysis

Appendix:

R-CODE:

```
---
title: "Assignment 3"
author: "Satyam Vatts & Avneet Kaur"
date: "01/11/2021"
output: word_document
---
```

Satyam Vatts
Avneet Kaur

DISCRIMINANAT ANALYSIS & KNN PREDICTION COMPARISONS

Importing Libraries

```{r}
library(dplyr)
library(car)
library(tidyverse)
library(MASS)
library(klaR)
```

```{r}
library(MASS)  #load the package for lda functions
library(DiscriMiner) #load the package for lda functions
library(ggplot2)  #visualization
library(dplyr) #data manipulation
library(gridExtra) #visualization
library(car) #multvariate test
library(psych)
library(corrplot) #visualization for correlation
library(Hmisc) #run descriptive analysis
```

Loading Imputed Datasets: Mice, Amelia, MissForest

```{r}
mice_data <- read.csv('mice_imputed.csv')[-c(1,2,46)]
amelia_data <- read.csv('amelia_imputed.csv')[-c(1,2)]
missfor_data <- read.csv('missforest_imputed.csv')[-c(1,2)]

cat_vars <- c(2,4,5,7,8,9,19,20,21,22,86,87,88)
cat_vars_a <- c(2,4,5,7,8,9,19,20,21,85,86,87)
cat_vars_m <- c(2,4,5,7,8,9,19,20,21,22,85,86,87)


mice_data[cat_vars_m] = lapply(mice_data[cat_vars_m], as.factor)
missfor_data[cat_vars] = lapply(missfor_data[cat_vars], as.factor)
amelia_data[cat_vars_a] = lapply(amelia_data[cat_vars_a], as.factor)
```

Extrating Groups of Data based on response target variable 'pex' so that various tests can be performed.

Since Pex variable does not has a blank, we can use the same filter for all datasets

```{r}
pex_treatment <- mice_data$pex == 1
mice_pex <- subset(mice_data,pex_treatment)
mice_trad <- subset(mice_data,!pex_treatment)
```

Satyam Vatts
Avneet Kaur

amelia_pex <- subset(amelia_data,pex_treatment)

amelia_trad <- subset(amelia_data,!pex_treatment)

missfor_pex <- subset(missfor_data,pex_treatment)

missfor_trad <- subset(missfor_data,!pex_treatment)

```

```{r}

mice_summary_pex <- rename(as.data.frame(cbind(as.data.frame(t(mice_pex %>%  summarise_if(is.numeric,sd)))$V1,

as.data.frame(t(mice_pex %>%  summarise_if(is.numeric,mean)))$V1)), PEX_STDEV = V1, PEX_MEAN=V2)

mice_summary_trad <- rename(as.data.frame(cbind(as.data.frame(t(mice_trad %>%  summarise_if(is.numeric,sd)))$V1,

as.data.frame(t(mice_trad %>%  summarise_if(is.numeric,mean)))$V1)), TRAD_STDEV = V1, TRAD_MEAN=V2)

mice_summary <- data.frame(cbind(mice_summary_pex,mice_summary_trad), row.names = colnames(mice_pex[-cat_vars_m]))

write.csv(mice_summary, "mice_summary.csv")

mice_summary

```

```{r}

amelia_summary_pex <- rename(as.data.frame(cbind(as.data.frame(t(amelia_pex %>%  summarise_if(is.numeric,sd)))$V1,

as.data.frame(t(amelia_pex %>%  summarise_if(is.numeric,mean)))$V1)), PEX_STDEV = V1, PEX_MEAN=V2)

amelia_summary_trad <- rename(as.data.frame(cbind(as.data.frame(t(amelia_trad %>%  summarise_if(is.numeric,sd)))$V1,

as.data.frame(t(amelia_trad %>%  summarise_if(is.numeric,mean)))$V1)), TRAD_STDEV = V1, TRAD_MEAN=V2)

amelia_summary <- data.frame(cbind(amelia_summary_pex,amelia_summary_trad), row.names = colnames(amelia_pex[-cat_vars_a]))

write.csv(amelia_summary, "amelia_summary.csv")

amelia_summary

```

```{r}

missfor_summary_pex <- rename(as.data.frame(cbind(as.data.frame(t(missfor_pex %>%  summarise_if(is.numeric,sd)))$V1,

as.data.frame(t(missfor_pex %>%  summarise_if(is.numeric,mean)))$V1)), PEX_STDEV = V1, PEX_MEAN=V2)

missfor_summary_trad <- rename(as.data.frame(cbind(as.data.frame(t(missfor_trad %>%  summarise_if(is.numeric,sd)))$V1,

as.data.frame(t(missfor_trad %>%  summarise_if(is.numeric,mean)))$V1)), TRAD_STDEV = V1, TRAD_MEAN=V2)

missfor_summary <- data.frame(cbind(missfor_summary_pex,missfor_summary_trad), row.names = colnames(missfor_pex[-cat_vars]))

write.csv(missfor_summary, "missfor_summary.csv")

missfor_summary

```

```{r}

mice_cat_per <- data.frame(Treatement = c('Values', 'PEX', 'TRAD'))

```
for (i in 1:ncol(mice_pex[cat_vars_m])){
  props_pex <- rename(as.data.frame(prop.table(table(mice_pex[cat_vars_m][,i]))),
                      Data_Values = Var1, Percentage = Freq)
  props_trad <- rename(as.data.frame(prop.table(table(mice_trad[cat_vars_m][,i]))),
                      Data_Values = Var1, Percentage = Freq)
  props_all <- rbind(t(props_pex), t(props_trad)[2,])
  mice_cat_per <- cbind(mice_cat_per, props_all)
}
amelia_cat_per <- data.frame(Treatement = c('Values', 'PEX', 'TRAD'))
for (i in 1:ncol(amelia_pex[cat_vars_a])){
  props_pex <- rename(as.data.frame(prop.table(table(amelia_pex[cat_vars_a][,i]))),
                      Data_Values = Var1, Percentage = Freq)
  props_trad <- rename(as.data.frame(prop.table(table(amelia_trad[cat_vars_a][,i]))),
                      Data_Values = Var1, Percentage = Freq)
  props_all <- rbind(t(props_pex), t(props_trad)[2,])
  amelia_cat_per <- cbind(amelia_cat_per, props_all)
}
missfor_cat_per <- data.frame(Treatement = c('Values', 'PEX', 'TRAD'))
for (i in 1:ncol(missfor_pex[cat_vars])){
  props_pex <- rename(as.data.frame(prop.table(table(missfor_pex[cat_vars][,i]))),
                      Data_Values = Var1, Percentage = Freq)
  props_trad <- rename(as.data.frame(prop.table(table(missfor_trad[cat_vars][,i]))),
                      Data_Values = Var1, Percentage = Freq)
  props_all <- rbind(t(props_pex), t(props_trad)[2,])
  missfor_cat_per <- cbind(missfor_cat_per, props_all)
}
write.csv(mice_cat_per, 'mice_sum_cat.csv')
write.csv(amelia_cat_per, 'amelia_sum_cat.csv')
write.csv(missfor_cat_per, 'missfor_sum_cat.csv')
```

Discriminant Analysis:

Preparing the dataset

```{r}
da_mice <- mice_data[-cat_vars_m[-13]]
da_amelia <- amelia_data[-cat_vars_a[-12]]
da_missfor <- missfor_data[-cat_vars[-13]]
```

1.

Checking Sample Size Per Category

```{r}
print('For MiCE/Amelia/Missforest Imputed Data: Distribution of Patients between two Treatements')
print('0: Traditional Treatment   1: PEX Treatment')
```

Satyam Vatts
Avneet Kaur

```
tableV1 <- table(mice_data$pex)

tableV1

print(paste('Number of Variables in MICE Dataset : ', ncol(da_mice)))

print(paste('Number of Variables in Amelia Dataset : ', ncol(da_amelia)))

print(paste('Number of Variables in MissForest Dataset : ', ncol(da_missfor)))
```

Since, DA is performed only on numerical variables, so after remving all the categorical variables, we have 75 vairables. SO, the number of observations in each group is more than the number of variables.

2.

Checking Multi-Variate Normality

```{r}
for (i in 1:ncol(da_mice[-75])){
  qqPlot(unlist(da_mice[,i]), ylab = colnames(da_mice[i]),main = paste("QQ Plot of" , colnames(da_mice[i])),col = 'orange')
}
```

```{r}
sh_df <- data.frame(Variable = c(), Test_Statistics = c(), P_Value = c())
for (i in 1:ncol(da_mice[-75])){
  sh_t <- shapiro.test(unlist(da_mice[-75][,i]))
  sh_df <- rbind(sh_df, data.frame(Variable=c(colnames(da_mice[-75])[i]), Test_Statistics = c(sh_t$statistic), P_Value = c(sh_t$p.value)))
}
write.csv(sh_df, 'sh_mice.csv')
sh_df1 <- data.frame(Variable = c(), Test_Statistics = c(), P_Value = c())
for (i in 1:ncol(da_amelia[-76])){
  sh_t <- shapiro.test(unlist(da_amelia[-76][,i]))
  sh_df1 <- rbind(sh_df1, data.frame(Variable=c(colnames(da_amelia[-75])[i]), Test_Statistics = c(sh_t$statistic), P_Value = c(sh_t$p.value)))
}
write.csv(sh_df1, 'sh_amelia.csv')
sh_df2 <- data.frame(Variable = c(), Test_Statistics = c(), P_Value = c())
for (i in 1:ncol(da_missfor[-76])){
  sh_t <- shapiro.test(unlist(da_missfor[-76][,i]))
  sh_df2 <- rbind(sh_df2, data.frame(Variable=c(colnames(da_missfor[-76])[i]), Test_Statistics = c(sh_t$statistic), P_Value = c(sh_t$p.value)))
}
write.csv(sh_df2, 'sh_missfor.csv')
```

3. Checking Multi-Colinearity

```{r}
correlations <- data.frame(related_cols = c('A, B'), cor_coeff = c(0))
crrelat =cor(da_mice[,-75])
for (i in 1:nrow(crrelat)){
```

```
  for (j in i:ncol(crrelat )){
    if (i!=j && crrelat [i,j]>0.8){
      correlations <- correlations %>% add_row(related_cols = paste(colnames(da_mice[,-75])[i]," , ", colnames(da_mice[,-75])[j]), cor_coeff
=crrelat[i,j])
    }
  }
}
correlations <- correlations[-1,]
correlations[order(-correlations$cor_coeff),]
write.csv(correlations, 'multi-norm-check_mice.csv')
```

```{r}
correlations1 <- data.frame(related_cols = c('A, B'), cor_coeff = c(0))
crrelat =cor(da_amelia[,-76])
for (i in 1:nrow(crrelat)){
  for (j in i:ncol(crrelat )){
    if (i!=j && crrelat [i,j]>0.8){
      correlations1 <- correlations1 %>% add_row(related_cols = paste(colnames(da_amelia[,-76])[i]," , ", colnames(da_amelia[,-76])[j]),
cor_coeff =crrelat[i,j])
    }
  }
}
correlations1 <- correlations1[-1,]
correlations1[order(-correlations1$cor_coeff),]
write.csv(correlations1, 'multi-norm-check_ame.csv')
```

```{r}
correlations2 <- data.frame(related_cols = c('A, B'), cor_coeff = c(0))
crrelat =cor(da_missfor[,-76])
for (i in 1:nrow(crrelat)){
  for (j in i:ncol(crrelat )){
    if (i!=j && crrelat [i,j]>0.8){
      correlations2 <- correlations2 %>% add_row(related_cols = paste(colnames(da_missfor[,-76])[i]," , ", colnames(da_missfor[,-76])[j]),
cor_coeff =crrelat[i,j])
    }
  }
}
correlations2 <- correlations2[-1,]
correlations2[order(-correlations2$cor_coeff),]
write.csv(correlations2, 'multi-norm-check_missf.csv')
```

From the above observations, the following decisions are made:

1. Since Ranson and IMRIE scores are almost similar, we can drop one of them. We are dropping imrie

2. km_paco2 and km_hco3 have high correlation coz they provide info about blood gas so we can remove t0 and t30 of one of these. We are removing hco3

3. Others have high correlation since they are the tests being taken for both pex and traditional treatment and have similar scores. Thus, to comply to multicorrelation assumption we are going to drop the vairables corresponding to raditional treatment

Removing multicolinearity issue columns

```{r}
da_mice <- da_mice[-c(8, 12, 34,11, 49, 50)]
da_amelia <- da_amelia[-c(8,9,11,12,35,50,51)]
da_missfor <- da_missfor[-c(8,11,12,33,35,50,51)]
print(paste('Number of Variables in MICE Dataset : ', ncol(da_mice)))
print(paste('Number of Variables in Amelia Dataset : ', ncol(da_amelia)))
print(paste('Number of Variables in MissForest Dataset : ', ncol(da_missfor)))
```

4.Testing Equality of Covariance Matrices between two groups

```{r}
lv_df <- data.frame(Variable = c(), F_Value = c(), P_Value = c())
for (i in 1:ncol(da_mice[-69])){
  lv_t <- leveneTest(unlist(da_mice[-69][,i]) ~ unlist(da_mice[,69]))
  lv_df <- rbind(lv_df, data.frame(Variable = c(colnames(da_mice[-69])[i]), F_Value = c(lv_t$`F value`[1]), P_Value = c(lv_t$`Pr(>F)`[1])))
}
lv_df1 <- data.frame(Variable = c(), F_Value = c(), P_Value = c())
for (i in 1:ncol(da_amelia[-69])){
  lv_t <- leveneTest(unlist(da_amelia[-69][,i]) ~ unlist(da_amelia[,69]))
  lv_df1 <- rbind(lv_df1, data.frame(Variable = c(colnames(da_amelia[-69])[i]), F_Value = c(lv_t$`F value`[1]), P_Value = c(lv_t$`Pr(>F)`[1])))
}
lv_df2 <- data.frame(Variable = c(), F_Value = c(), P_Value = c())
for (i in 1:ncol(da_missfor[-69])){
  lv_t <- leveneTest(unlist(da_missfor[-69][,i]) ~ unlist(da_missfor[,69]))
  lv_df2 <- rbind(lv_df2, data.frame(Variable = c(colnames(da_missfor[-69])[i]), F_Value = c(lv_t$`F value`[1]), P_Value = c(lv_t$`Pr(>F)`[1])))
}
lv <- cbind(lv_df, lv_df1, lv_df2)
write.csv(lv, 'lv_all.csv')
```

##LDA Modelling:

Data Partitioning

```{r}
set.seed(0)
micediv <- sample(2, nrow(da_mice),
                  replace = TRUE,
                  prob = c(0.7, 0.3))
ameliadiv <- sample(2, nrow(da_amelia),
```

```
            replace = TRUE,
            prob = c(0.7, 0.3))
missfordiv <- sample(2, nrow(da_missfor),
            replace = TRUE,
            prob = c(0.7, 0.3))
mice_train <- da_mice[micediv == 1,]
mice_test <- da_mice[micediv == 2,]
amelia_train <- da_amelia[ameliadiv == 1,]
amelia_test <- da_amelia[ameliadiv == 2,]
missfor_train <- da_missfor[missfordiv == 1,]
missfor_test <- da_missfor[missfordiv == 2,]
```

```{r}
#View a ratio of sample size per category
print('MICE Dataset')
table(mice_train$pex)
prop.table(table(mice_train$pex))
print('AMELIA Dataset')
table(amelia_train$pex)
prop.table(table(amelia_train$pex))
print('MISSFOREST Dataset')
table(missfor_train$pex)
prop.table(table(missfor_train$pex))
```

Building Model

```{r}
#the '.' here means including all variable
mice_lda <- lda(pex~., mice_train)
amelia_lda <- lda(pex~., amelia_train)
missfor_lda <- lda(pex~., missfor_train)
```

```{r}
cbind(as.data.frame(mice_lda$counts),as.data.frame(amelia_lda$counts), as.data.frame(missfor_lda$counts))
cbind(as.data.frame(mice_lda$prior),as.data.frame(amelia_lda$prior), as.data.frame(missfor_lda$prior))
lda_coeff <- cbind(rename(as.data.frame(mice_lda$scaling), MICE_COEF=LD1),rename(as.data.frame(amelia_lda$scaling), AMELIA_COEF=LD1), rename(as.data.frame(missfor_lda$scaling), MISSFOR_COEF=LD1))
```

```{r}
g_means_am <- rename(as.data.frame(t(as.data.frame(amelia_lda$means))), AMELIA_TRAD = `0`, AMELIA_PEX= `1`)
g_means_mi <- rename(as.data.frame(t(as.data.frame(mice_lda$means))), MICE_TRAD = `0`, MICE_PEX= `1`)
```

Satyam Vatts
Avneet Kaur

```
g_means_ms <- rename(as.data.frame(t(as.data.frame(missfor_lda$means))), MISSFOR_TRAD = `0`, MISSFOR_PEX= `1`)

lda_sum<-cbind(lda_coeff, g_means_am, g_means_mi, g_means_ms)

lda_sum

write.csv(lda_sum, 'LDA_SUMMARY.csv')
```

```{r}
#2 discriminant functions the plot

plot(mice_lda)

plot(amelia_lda)

plot(missfor_lda)
```

```{r}
#use the model to predict on test dataset

mice_pred <- predict(mice_lda, mice_test)

amelia_pred <- predict(amelia_lda, amelia_test)

missfor_pred <- predict(missfor_lda, missfor_test)

pred_sum1 <- cbind(data.frame(Actual = mice_test$pex, Mice_Predited = mice_pred$class, Amelia_Predited = amelia_pred$class),
as.data.frame(mice_pred$posterior), as.data.frame(amelia_pred$posterior))

write.csv(pred_sum1, 'Prediction_Sum1.csv')
```

```{r}
pred_sum2 <- cbind(data.frame(Missfor_Predited = missfor_pred$class), as.data.frame(missfor_pred$posterior))

write.csv(pred_sum2, 'Prediction_Sum2.csv')
```

```{r}
print('MICE DATASET CONFUSION MATRIX')

mice_con_mat <- table(Predicted = mice_pred$class, Actual = mice_test$pex)

mice_con_mat

print(paste('MICE DATA Accuracy : ',sum(diag(mice_con_mat))/sum(mice_con_mat)))
```

```{r}
print('AMELIA DATASET CONFUSION MATRIX')

amelia_con_mat <- table(Predicted = amelia_pred$class, Actual = amelia_test$pex)

amelia_con_mat

print(paste('AMELIA DATA Accuracy : ',sum(diag(amelia_con_mat))/sum(amelia_con_mat)))
```

```{r}
print('MISSFOREST DATASET CONFUSION MATRIX')

missfor_con_mat <- table(Predicted = missfor_pred$class, Actual = missfor_test$pex)

missfor_con_mat
```

Satyam Vatts
Avneet Kaur

```
print(paste('MISSFOREST DATA Accuracy : ',sum(diag(missfor_con_mat))/sum(missfor_con_mat)))
```