Applied Machine Learning Internship at NCL NEDUET:

Report By:

Raja Shaharyar Ahmed (BM-21079)

Under Supervision Of:

Professor Abul Hasan

Task:

Given a folder. Within this are subfolders 'Stage0', 'Stage1', 'Stage2', 'Stage3' and 'Stage4'. In these subfolders are .mat and .csv files for features named 'F1', 'F2', 'F3', 'F4', 'F5' and 'F6', each containing a single variable, a matrix of the size the number of subjects at that stage x 38. Train a model to classify Stage 0 and Stage 1 subjects. Test the model using Stage 3 data.

Data Preprocessing And Exploratory Data Analysis:

1. All .csv files (F1, F2, F3, F4, F5 and F6) that are in a shape of (a matrix of the size the number of subjects at that stage x 38) converted into a single column for all the stages (0-4) and merged into a single dataframe with respect to their stages.

A	B	C	D	E	F	G
f1	f2	f3	f4	f5	f6	Stage
1.5172	0.75264	0.096214	0.09336	0.096214	0.008216	0
1.5694	0.87322	0.11167	0.10726	0.11167	0.009138	0
1.3328	0.67703	0.076207	0.075955	0.076207	0.005358	0
0.90531	0.45782	0.087615	0.089115	0.087615	0.009074	0
1.9076	2.114	0.41191	0.55859	0.41191	0.074853	0
1.7696	1.4893	0.19038	0.18568	0.19038	0.016448	0
1.7216	1.6781	0.23316	0.29657	0.23316	0.025486	0
1.3505	0.40313	0.096702	0.069709	0.096702	0.005853	0
0.88295	0.49523	0.10375	0.114	0.10375	0.012445	0
1.9716	2.4283	0.60414	0.92456	0.60414	0.14088	0
2.2752	2.5124	0.38228	0.35859	0.38228	0.03312	0
1.8091	1.8483	0.46548	0.63629	0.46548	0.067665	0
1.4454	0.46889	0.14102	0.09443	0.14102	0.006651	0
1.1138	0.64431	0.30176	0.35472	0.30176	0.04707	0

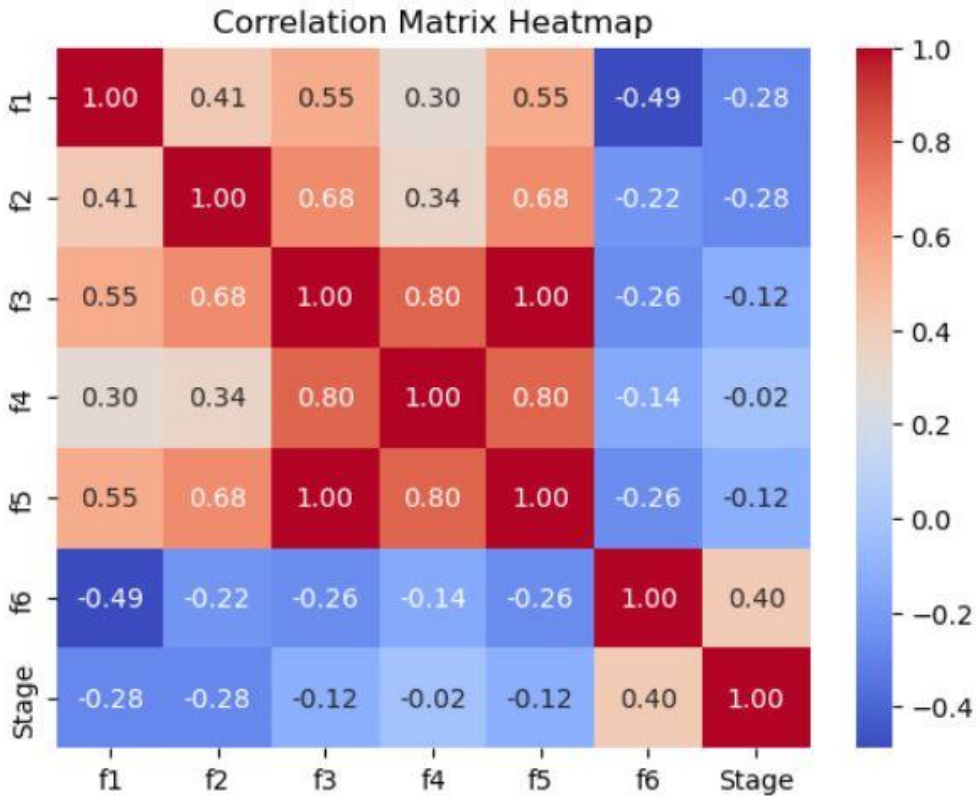
B	C	D	E	F	G	H
f1	f2	f3	f4	f5	f6	Stage
0.037483	0.005767	0.001237	0.003153	0.001237	1.3058	1
0.027802	0.005356	0.001262	0.003502	0.001262	1.3085	1
0.030149	0.005398	0.001218	0.003407	0.001218	1.3081	1
0.014595	0.004123	0.001528	0.003038	0.001528	1.3114	1
0.010791	0.005443	0.002044	0.005053	0.002044	1.3126	1
0.013039	0.005595	0.001605	0.004887	0.001605	1.3127	1
0.012225	0.005908	0.002077	0.005913	0.002077	1.3127	1
0.010048	0.003941	0.000883	0.002626	0.000883	1.313	1
0.007504	0.003339	0.001324	0.002455	0.001324	1.3133	1
0.007186	0.005475	0.002388	0.004875	0.002388	1.3133	1
0.011827	0.005606	0.001674	0.004511	0.001674	1.3131	1
0.00994	0.006312	0.003324	0.005876	0.003324	1.313	1
0.005523	0.004216	0.001174	0.002398	0.001174	1.314	1
0.002203	0.002108	0.000766	0.001341	0.000766	1.3147	1
0.005744	0.005495	0.002246	0.004439	0.002246	1.3137	1
0.010081	0.005935	0.001863	0.004384	0.001863	1.3132	1

B	C	D	E	F	G	H
f1	f2	f3	f4	f5	f6	Stage
0.002321	0.002892	0.007489	3.3231	0.007489	0.43312	2
0.001744	0.001896	0.006694	3.8999	0.006694	0.43167	2
0.001427	0.001383	0.005468	4.1798	0.005468	0.39473	2
0.001007	0.00146	0.004395	2.8853	0.004395	0.57437	2
0.000433	0.000609	0.002515	2.8381	0.002515	0.52947	2
0.000809	0.001073	0.00411	3.9631	0.00411	0.37305	2
0.000756	0.001049	0.004526	4.3867	0.004526	0.3641	2
0.000797	0.000956	0.003978	3.336	0.003978	0.44059	2
0.001835	0.002638	0.007406	3.086	0.007406	0.51475	2
0.001149	0.001486	0.003687	3.1267	0.003687	0.52614	2
0.000629	0.000959	0.00395	3.8937	0.00395	0.37576	2
0.000831	0.001255	0.004643	4.7789	0.004643	0.38869	2
0.000737	0.001063	0.003867	3.4115	0.003867	0.46995	2
0.000932	0.001371	0.004139	2.8881	0.004139	0.50055	2
0.001164	0.001599	0.004068	3.4452	0.004068	0.53537	2
0.000721	0.001298	0.004856	4.545	0.004856	0.34307	2

B	C	D	E	F	G	H
f1	f2	f3	f4	f5	f6	Stage
2.614	1.3786	0.62649	0.5893	0.62649	0.091245	3
2.371	1.3185	0.57093	0.48182	0.57093	0.062394	3
2.5129	1.5011	0.55678	0.45168	0.55678	0.049866	3
1.6486	0.84756	0.45594	0.56133	0.45594	0.096232	3
1.8814	1.8023	1.0421	1.2517	1.0421	0.23719	3
2.6766	2.7674	1.2093	1.2214	1.2093	0.17244	3
1.5737	1.6209	0.79383	0.83416	0.79383	0.13756	3
1.4312	0.8819	0.45459	0.47661	0.45459	0.063815	3
1.2434	0.72543	0.51323	0.58299	0.51323	0.095424	3
1.0051	1.2509	0.94225	1.5948	0.94225	0.37476	3
2.6148	3.1074	1.3113	1.3569	1.3113	0.19778	3
0.94549	1.2883	0.90853	1.2192	0.90853	0.29827	3
0.67544	0.3072	0.21828	0.23677	0.21828	0.041342	3
1.4192	0.98994	0.89284	1.0275	0.89284	0.17415	3

B	C	D	E	F	G	H
f1	f2	f3	f4	f5	f6	Stage
0.63697	0.17937	0.096075	0.042899	0.096075	0.002486	4
0.84353	0.2622	0.14283	0.059473	0.14283	0.002657	4
0.36331	0.13431	0.084454	0.039397	0.084454	0.002924	4
0.91621	0.53061	0.31088	0.12215	0.31088	0.004591	4
1.0655	0.49515	0.27045	0.11148	0.27045	0.003881	4
0.3176	0.18206	0.14032	0.068836	0.14032	0.006855	4
0.50074	0.4008	0.21242	0.081737	0.21242	0.003377	4
0.71456	0.5025	0.2718	0.10223	0.2718	0.003178	4
0.29955	0.1842	0.1811	0.083846	0.1811	0.008278	4
0.48294	0.45762	0.24996	0.084191	0.24996	0.003496	4
0.45036	0.42079	0.20429	0.076759	0.20429	0.005297	4
0.61055	0.70003	0.38128	0.10716	0.38128	0.004182	4
0.67414	0.73371	0.3751	0.12584	0.3751	0.00725	4
0.84195	0.99866	0.51455	0.15611	0.51455	0.007601	4
0.9772	1.2038	0.63446	0.1734	0.63446	0.007306	4
0.72987	0.7827	0.4339	0.15296	0.4339	0.00906	4

2. Outliers from the data are removed.
3. No null values are detected in the datasets.
4. Correlation is drawn to study relation of features with target variable.



It can be seen that feature "f4" has nearly no effect on "Stage".

Model Training:

Model is trained with [f1, f2, f3, f5 and f6] as features and target variables are [0 and 1] where 0 represent stage 0 and 1 represents stage 1.

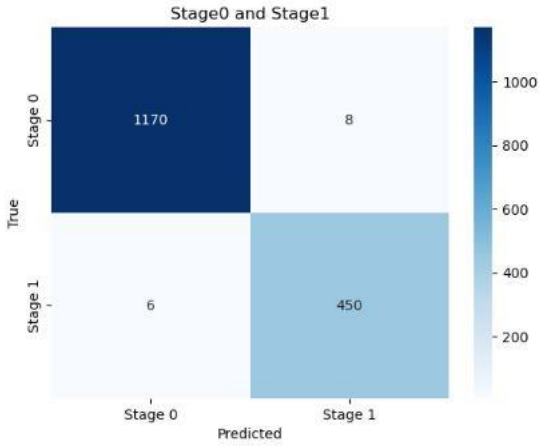
F4 is neglected because without it model works more accurate due to its negative relation with target variable.

Model is tested with K Fold Cross on 8 splits, given are the accuracies using different models.

Number of splits	Logistic Regression	Support Vector Machine	Random Forest Classifier	Gradient Booster
1	84	83	94	91
2	75	76	90	88
3	78	78	88	86
4	83	83	92	91
5	78	78	87	87
6	83	84	91	91
7	77	77	89	86
8	80	83	91	88

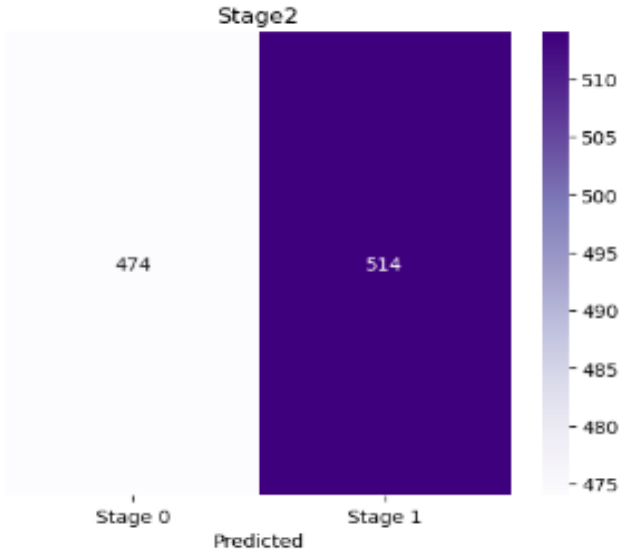
Finally, model is trained using Random Forest Classifier giving average 8 folds accuracy of 90%. But I trained Random Forest Classifier on datapoints where k-fold accuracy is 94%.

Results Of Model Testing On Stage 0 and Stage 1 Dataset:



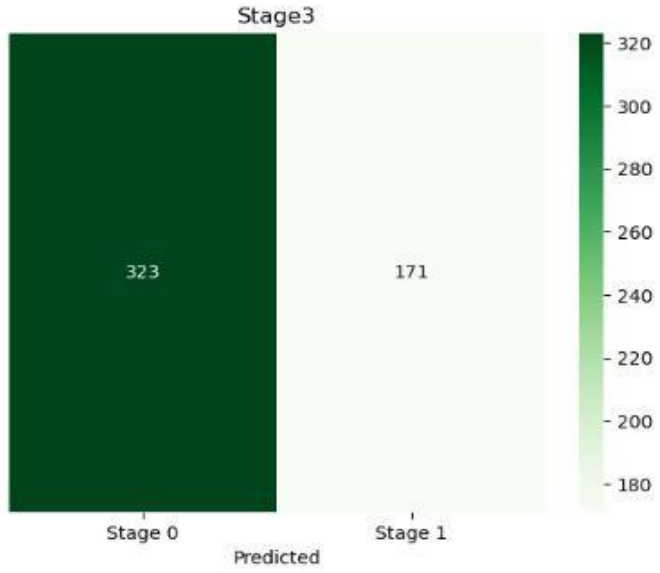
- **Out of 1178 stage 0 entries model accurately predicted 1170 entries.**
- **Out of 456 stage 1 entries model accurately predicted 450 entries.**

Results Of Model Testing On Stage 2 Dataset:



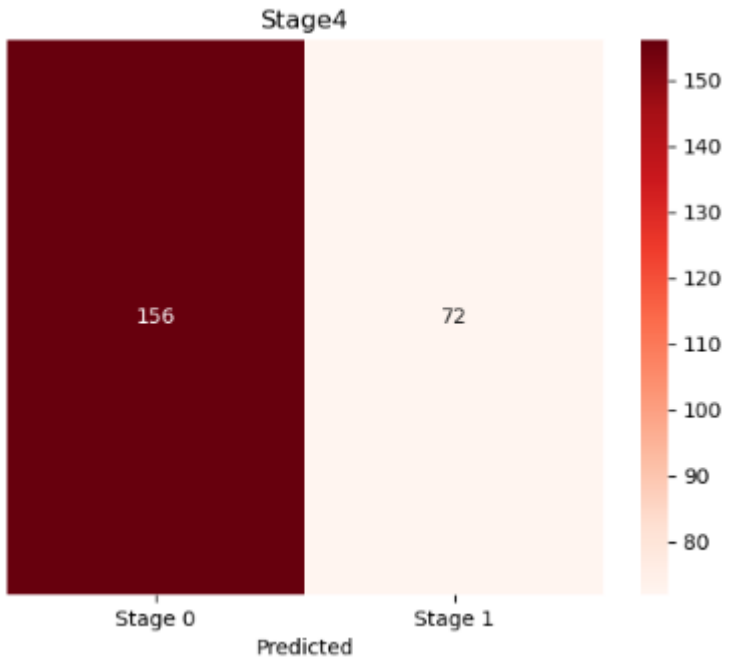
- **Total Stage 2 Entries:** 988
- **Classified as Stage 0:** 474
- **Classified as Stage 1:** 514

Results Of Model Testing On Stage 3 Dataset:



- **Total Stage 3 Entries:** 494
- **Classified as Stage 0:** 323
- **Classified as Stage 1:** 171

Results Of Model Testing On Stage 4 Dataset:



- **Total Stage 4 Entries:** 228
- **Classified as Stage 0:** 156
- **Classified as Stage 1:** 72