CSI5155 Machine Learning

Final project report

Name: Pavan Balaji Kumar

Student Number: 300169572

Analysis of readmission rate and Medicine dosage change Using Supervised and Semi-supervised Learning

Abstract:

Diabetes is one of the most common ailments people suffer from around the world. Managing the right doses of medicines and monitoring the patients' admission rate can significantly improve the patient's well-being from this ailment. In this project, data from almost 70000 patients consisting of nearly 100000 instances have been analyzed to find the relationship between various features such as patient's age, sex, previous medical conditions, test results, medicines taken etc. and the readmission rate of the patients. Further, how the change in the dosage of a particular medicine is affected by the other features was also examined. A total of 5 Supervised algorithm and 3 Semi supervised algorithms was used to fit the above-mentioned relationships and results were obtained. Important classification metrics, like recall and sensitivity, along with other measures, were calculated to derive meaningful insights.

1 Introduction:

Diabetes is becoming a common disease, especially among young people in their mid-20s and early-30s. Monitoring such patients and controlling their medication can help their well-being and prolong their life without any significant discrepancies. The project address two problems: (i) The impact of the patient's profiles and medical history has on the readmission rate of the patient, (ii) How dosages can be changed based on the patient's medications, test results and other features. These problems were analyzed using five different Supervised learning algorithms namely Multi-Layered perceptron, K nearest neighbours, Decision tree, Gaussian Navies Bayesian and Extreme learning trees. Three semi-supervised learning algorithms, Label propagation, Label spreading, and Self-learning were also used to analyze the first task.

2 Data Preprocessing:

2.1 Dataset Description

The dataset used for the project was obtained from the UCI repository. The data consists of 101766 instances of almost 70000 unique patients and 50 features. In the First task, the feature column "readmitted" is used as the target variable to find the readmitted status, and in the second task, the feature column "**insulin**" is used as the target variable to find the change in dosage. After doing the primary analysis of the Dataset, Feature engineering was done.

2.2 Handling missing data and Feature Engineering

By examining the dataset, it was found that only some columns had a high percentage of missing values. Features like payer_code, weight and medical_speciality had a high percentage of missing values at 39.56%, 49.08% and 96.86%. Feature with that high percentage of missing data cannot be used for the supervised learning task as it may negatively affect the statistical models' performance. When most of the values are missing in a column, filling the missing

values may introduce unwanted bias into the model. Hence columns with such a high percentage of missing values were dropped.

Features	Missing value %
weight	96.86
medical_specialty	49.08
payer_code	39.56
race	2.23
diag_3	1.40
diag_2	0.35
diag_1	0.02

Figure 1. Missing value percentage in each feature

Missing values in columns with minuscule missing value percentages were filled using a Simple imputer from scikit learn by the most frequent occurrence in that particular column. Dropping those rows is not advisable as it may affect the model's performance, and some vital information may be lost. After handling missing values, feature engineering was done. The encounter_id feature column was also dropped because it had a unique value for all the instances in the dataset and may hinder performance as there is no real pattern for the model to learn.

The dataset consisted of 50 features, out of which 37 were categorical, and 13 were numerical. The categorical variables were encoded into numerical variables for using it in a machine learning algorithm. The categorical variables were encoded using Ordina encoder from scikit learn two other custom function to encode the gender and medicine dosage status columns. The custom functions were used to avoid One-hot encoding, which will increase the sparsity of the dataset, increasing the runtime. The gender_encoding function was used to set Male as **1**, Female as **-1** and Other as **0**. The medicine_status_encoding function set Up as **2**, Down as **1**, no as **-1** and steady as **0.** The values were given in such a manner to match any positive or negative change. Once all the categorical variables were converted to numerical variables, the Dataset was Standardised using the Standard Scalar function from scikit learn to improve the convergence time and smooth running of the algorithm.

3 Experiments and Results

After Standardisation was done, the data is now ready to be used with the machine learning algorithms. Three main tasks are going to be addressed in this project which are:

a. Predicting the readmission status using Supervised Learning

b. Prediction of Change in Insulin dosage using Supervised Learning
c. Predicting the readmission status using Semi-supervised Learning

3.1 Predicting the readmission status using Supervised Learning

Multi-Layered perceptron (MLP), K nearest neighbours (KNN), Decision tree, Gaussian Navies Bayesian and Extreme learning trees were used to predict the readmitted status using the dataset set features.

3.1.1 Model Traning

The MLP model was built with three hidden layers with sizes 128,64, and 32, respectively. The neural network was run for 500 iterations with a learning rate of 0.003. The KNN model was trained with k set to 20. The Naive Bayesian classifier and Extreme learning tree were used in their base parameters.The Decision tree algorithm was trained with its max depth set to 15 and all other parameters the same as the base ones.

All the algorithms were trained using cross-validation for ten folds using the training dataset, and the results are shown in the below figure.

Algorithm / Folds	MLP	KNN	DT	NB	ELT
1	0.592690	0.603301	0.609459	0.459976	0.638019
2	0.580506	0.602122	0.609983	0.460107	0.640115
3	0.596096	0.592297	0.597537	0.583519	0.633434
4	0.602908	0.602777	0.613127	0.460238	0.637757
5	0.590540	0.597615	0.610980	0.460430	0.642820
6	0.595126	0.602332	0.614256	0.460037	0.642427
7	0.587526	0.592374	0.601808	0.583333	0.629062
8	0.597877	0.604036	0.609932	0.585430	0.643999
9	0.590540	0.600629	0.615959	0.460037	0.643606
10	0.589361	0.596960	0.613732	0.459906	0.637841
Average Accuracy(mean)	0.592317	0.599444	0.609677	0.497301	0.638908
Standard deviation	0.005856	0.004178	0.005477	0.056822	0.004553

Figure 2. Cross-Validation results; Prediction of readmission status

It can be seen from Figure 2 that the algorithm that performs the best in this task is the extreme learning tree ensemble algorithm with an accuracy of almost 64%. Naive Bayesian algorithm is the worst performer with an average accuracy, not even at 50%. The final estimator from this was obtained, and the further measure was collected using the test dataset.

3.1.2 Model Evaluation

All five models were evaluated on the test dataset. Sensitivity(recall) and specificity, some of the critical measures for the medical field was calculated. Also, F1-measure, the combination of Precision and recall, accuracy on the test dataset, average runtime models and the area under the roc curve is calculated. The measures calculated can be seen in the below figure.

	MLP	KNN	DT	NB	ELT
f1_score	0.5321	0.4958	0.5707	0.6337	0.5614
Recall	0.5127	0.4244	0.5663	0.9997	0.5111
Specificity	0.6416	0.7514	0.6383	0.0007	0.732
runtime	11 min 50 secs	01 min 23 secs	00 min 02 secs	00 min 01 secs	00 min 25 secs
accuracy	0.5818	0.5997	0.6049	0.464	0.6295
AUC	0.5987	0.6317	0.622	0.6373	0.6802

Figure 3. Evaluation Results; Prediction of readmission status

Figure 3 shows that the MLP model has a very high runtime compared to the other models, around 9 mins 4 secs, but the accuracy is less than that of the Extreme learning trees, which has a runtime of 17 secs. So is it not feasible to use MLP for this classification task. The Extreme learning tress model also has a higher F1-score, Specificity and accuracy than others suggesting that this is the best classifier for this task. This is further supported in figure 4, where the ROC curve for Extreme learning tress is higher than the others.

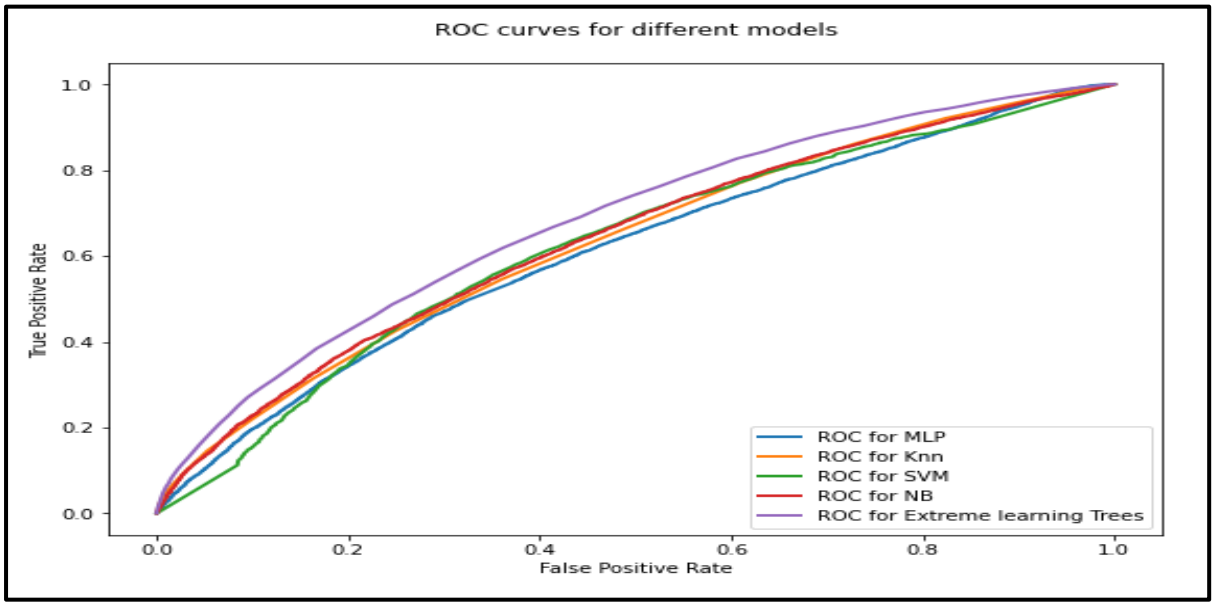

Figure 4. ROC curves for Different modes; Prediction of readmission rate

3.1.3 Statistical Significance

A student's t-test was performed on the accuracy obtained from the ten-fold cross-validation performed earlier to find the statistical significance between the different algorithms. P-value was calculated, and the 95% significance level was tested. The results of the statistical test can be seen in Figure 5.

Accuracy Difference Fold	MLP-KNN	MLP-DT	MLP-NB	MLP-ELT	KNN-DT	KNN-NB	KNN-ELT	DT-NB	DT-ELT	NB-ELT
1	-0.010612	-0.016769	0.132713	-4.532949e-02	-0.006157	0.143325	-3.471767e-02	0.149483	-2.856020e-02	-0.178043
2	-0.021617	-0.029477	0.120398	-5.960959e-02	-0.007861	0.142015	-3.799293e-02	0.149876	-3.013232e-02	-0.180008
3	0.003799	-0.001441	0.012577	-3.733788e-02	-0.005240	0.008778	-4.113717e-02	0.014018	-3.589676e-02	-0.049915
4	0.000131	-0.010219	0.142670	-3.484868e-02	-0.010350	0.142539	-3.497969e-02	0.152889	-2.462990e-02	-0.177519
5	-0.007075	-0.020440	0.130110	-5.227987e-02	-0.013365	0.137186	-4.520440e-02	0.150550	-3.183962e-02	-0.182390
6	-0.007206	-0.019130	0.135089	-4.730084e-02	-0.011923	0.142296	-4.009434e-02	0.154219	-2.817086e-02	-0.182390
7	-0.004848	-0.014282	0.004193	-4.153564e-02	-0.009434	0.009041	-3.668763e-02	0.018475	-2.725367e-02	-0.045729
8	-0.006158	-0.012055	0.012448	-4.612159e-02	-0.005896	0.018606	-3.996331e-02	0.024502	-3.406709e-02	-0.058569
9	-0.010089	-0.025419	0.130503	-5.306604e-02	-0.015330	0.140592	-4.297694e-02	0.155922	-2.764675e-02	-0.183569
10	-0.007600	-0.024371	0.129455	-4.848008e-02	-0.016771	0.137055	-4.088050e-02	0.153826	-2.410901e-02	-0.177935
Average Accuracy(mean)	-0.007128	-0.017360	0.095016	-4.659097e-02	-0.010233	0.102143	-3.946346e-02	0.112376	-2.923062e-02	-0.141607
Standard Deviation	0.006398	0.007848	0.056109	7.061985e-03	0.003847	0.059007	3.213324e-03	0.061205	3.617124e-03	0.059156
p-value	0.008629	0.000095	0.000663	9.955019e-09	0.000023	0.000569	3.961615e-11	0.000376	1.653888e-09	0.000052

Figure 5. Statistical Test Results; Prediction of readmission rate

From the results shown in figure 5, it can bee is seen that there is a statistical difference between all the algorithms

3.2 Prediction of Insulin Dosage change Using supervised Learning

The same algorithms used in the first task was also used for this task. The target variable used in this task consists of four classes Up, down, no and steady. The second task can be considered as a multi-class classification problem.

3.2.1 Model Traning

All the models were trained with parameters that set in the first using the new training dataset, including the target variable from the previous task as a feature. The new target variable in the insulin dosage status consisting of four classes. The models were trained using ten-fold cross-validation. The cross-validation results are shown in figure 6.

Figure 6 shows that the Extreme Learning Tree performs the best on the training dataset with an accuracy of almost 80%, which is higher than all the other algorithms under consideration.

Algorithm / Folds	MLP	KNN	DT	NB	ELT
1	0.762872	0.763527	0.775318	0.491026	0.794052
2	0.765361	0.763265	0.783571	0.494039	0.791301
3	0.768112	0.767195	0.781868	0.487358	0.791301
4	0.776497	0.776628	0.784750	0.494301	0.803485
5	0.771751	0.769130	0.777254	0.498428	0.791798
6	0.761923	0.763758	0.783412	0.492925	0.792977
7	0.767165	0.763889	0.776992	0.493580	0.795597
8	0.766640	0.767820	0.775419	0.508124	0.790356
9	0.767951	0.769261	0.781709	0.482573	0.794549
10	0.756551	0.769523	0.775681	0.487421	0.789177
Average Accuracy(mean)	0.766482	0.767399	0.779597	0.492977	0.793459
Standard deviation	0.005168	0.003925	0.003604	0.006610	0.003833

Figure 6. Cross-validation results; Predication of insulin dosage change

All the other algorithms are similar in performance except the Naïve Bayesian Classifier, which has a low accuracy of 49.2 %. It is clear that the Naive Bayesian classifier is not working on this dataset from tasks one and two.

3.2.2 Model Evaluation

The Five supervised learning models were evaluated based on the recall, specificity, f1-score, area under roc, runtime and accuracy on the test dataset. Since this is a Multi-class classification problem, classification measures such as F1-score, the area under roc, recall and specificity were calculated using the **macro** average method to combine the measures obtained for one vs all method.

	MLP	KNN	DT	NB	ELT
f1_score	0.7646	0.7675	0.7873	0.4959	0.7964
Recall	0.7646	0.7675	0.7873	0.4959	0.7964
Specificity	0.922	0.9162	0.9264	0.8278	0.9292
runtime	06 min 08 secs	00 min 54 secs	00 min 01 secs	00 min 01 secs	00 min 12 secs
accuracy	0.7646	0.7675	0.7873	0.4959	0.7964
AUC	0.9275	0.9233	0.9037	0.8139	0.9455

Figure 7. Results; Prediction of insulin dosage change

From Figure 7, it can be seen that the algorithms Decision tree and Extreme learning trees perform well on this dataset. They have high f1-measure, recall, specificity and accuracy. This is further supported by the Roc curves for the models shown in the figure 8.

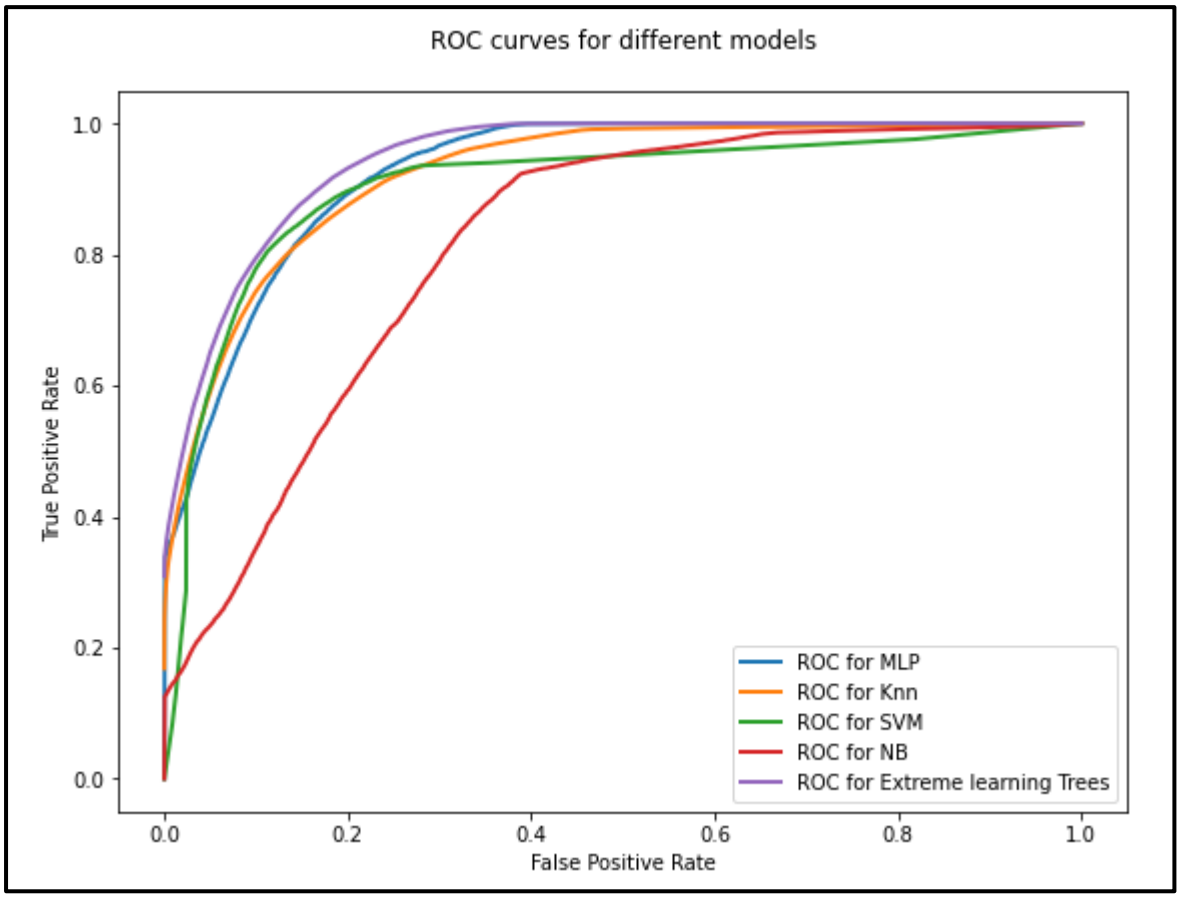

Figure 8. ROC curves for Different models; Prediction of Insulin dosage change

3.2.3 Statistical Significance

Accuracy Difference Fold	MLP-KNN	MLP-DT	MLP-NB	MLP-ELT	KNN-DT	KNN-NB	KNN-ELT	DT-NB	DT-ELT	NB-ELT
1	-0.0007	-0.0124	0.2718	-0.0312	-0.0118	0.2725	-0.0305	0.2843	-0.0187	-0.3030
2	0.0021	-0.0182	0.2713	-0.0259	-0.0203	0.2692	-0.0280	0.2895	-0.0077	-0.2973
3	0.0009	-0.0138	0.2808	-0.0232	-0.0147	0.2798	-0.0241	0.2945	-0.0094	-0.3039
4	-0.0001	-0.0083	0.2822	-0.0270	-0.0081	0.2823	-0.0269	0.2904	-0.0187	-0.3092
5	0.0026	-0.0055	0.2733	-0.0200	-0.0081	0.2707	-0.0227	0.2788	-0.0145	-0.2934
6	-0.0018	-0.0215	0.2690	-0.0311	-0.0197	0.2708	-0.0292	0.2905	-0.0096	-0.3001
7	0.0033	-0.0098	0.2736	-0.0284	-0.0131	0.2703	-0.0317	0.2834	-0.0186	-0.3020
8	-0.0012	-0.0088	0.2585	-0.0237	-0.0076	0.2597	-0.0225	0.2673	-0.0149	-0.2822
9	-0.0013	-0.0138	0.2854	-0.0266	-0.0124	0.2867	-0.0253	0.2991	-0.0128	-0.3120
10	-0.0130	-0.0191	0.2691	-0.0326	-0.0062	0.2821	-0.0197	0.2883	-0.0135	-0.3018
Average Accuracy(mean)	-0.0009	-0.0131	0.2735	-0.0270	-0.0122	0.2744	-0.0261	0.2866	-0.0139	-0.3005
Standard Deviation	0.0044	0.0049	0.0074	0.0038	0.0047	0.0077	0.0037	0.0084	0.0039	0.0079
p-value	0.5431	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Figure 9. Statistical Test Results; Prediction of readmission rate

The same statistical test as the first one was performed at the same significance level to find out which algorithms are significantly different. Based on the results, all the algorithms are significantly different except MLP and knn.

3.3 Predicting the readmission status using Semi-supervised Learning

In Semi-supervised Learning, three algorithms were used, namely Label propagation, Label spreading and Self-learning, on five different datasets with different levels of unlabelled data.

3.3.1 Model Traning and Evaluation

The three models were trained on five different datasets with a different percentage of unlabelled data like 10,20,50,90, and 95 percentage. The Labels are first obtained using the three algorithms. These labelled data is then used to train a random forest classifier for ten-fold cross-validation to get the performance measures. The classifiers are then obtained from the cross_validation results.

Metrics like F1-measure, recall, specificity, accuracy, the area under roc and runtime are calculated for the classifiers. This is shown in the Figure below.

Algorithms	Label propagation					Label Spreading					Self-learning				
Unlabelled percentage	10 %	20 %	50 %	90 %	95 %	10 %	20 %	50 %	90 %	95 %	10 %	20 %	50 %	90 %	95 %
f1_score	0.5763	0.5722	0.5489	0.4997	0.4634	0.5755	0.5685	0.5462	0.5015	0.4679	0.5728	0.568	0.5309	0.5484	0.5309
Recall	0.5267	0.5177	0.4805	0.4205	0.3767	0.5252	0.5122	0.475	0.4223	0.3812	0.5228	0.5082	0.4524	0.4796	0.4524
Specificity	0.7396	0.7476	0.7663	0.7729	0.7844	0.7406	0.7496	0.7715	0.7735	0.7852	0.7383	0.7568	0.7821	0.7669	0.7821
runtime	00 min 11 secs	00 min 11 secs	00 min 14 secs	00 min 14 secs	00 min 14 secs	00 min 11 secs	00 min 11 secs	00 min 14 secs	00 min 14 secs	00 min 14 secs	00 min 14 secs	00 min 16 secs	00 min 15 secs	00 min 13 secs	00 min 07 secs
accuracy	0.6409	0.641	0.6337	0.6095	0.5953	0.6409	0.641	0.6337	0.6095	0.5953	0.6384	0.6415	0.6408	0.6337	0.6292
AUC	0.6945	0.6952	0.6818	0.6446	0.6264	0.6938	0.6934	0.6835	0.6463	0.6265	0.6922	0.6946	0.689	0.6946	0.6735

Figure 10. Test Results; Prediction of readmission rate Semi supervised

Figure 10 show that the random forest approach classifier achieved higher performance than the supervised approach when the data from Label propagation for the output of 20% unlabelled.

3.3.2 Statistical significance

Since we need to compare the algorithm's performance difference on five different datasets, the Friedmans test was used to find the statistically significant difference in the algorithms. From the results shown in the table below, it was clear that there was a statistical difference between the algorithms' performance as the Friedman statistic obtain,10. was higher than the values from Friedman's table.

	Label propagation	Label Spreading	Self-learning
10 %	0.6508	0.6508	0.6614
20 %	0.6533	0.6533	0.6964
50 %	0.675	0.675	0.7723
90 %	0.7265	0.7265	0.8575
95 %	0.7439	0.7439	0.8565

Figure 11. Statistic Test Results; Prediction of readmission rate Semi supervised

4 Conclusion and Future Work

From the analysis and results, it can be concluded that the Extreme learning tree, which is an ensemble-based learning algorithm, works best when it comes to predicting the readmission rate and how the dosage should be changed for the insulin. The semi-supervised algorithm works better than the supervised algorithm in some cases. Hence it can be concluded the Learning on its own has an advantage in machine learning.

More analysis can be done in the future to predict how the dosage changes can be done to all the medicines in the dataset. How diabetes affects each gender or a particular age group can also be analyzed