Vee Nis Ling | vnl226 | Human-Centered Data Science Methods for Diabetes Diagnosis

The code and data files for this project can be found here.

Research Question

How can we use human-centered data science to most effectively determine whether ICU patients have been diagnosed with Diabetes Mellitus?

Background and Related Work

In terms of applying machine learning and/or data science to Diabetes Mellitus, extensive work has already been done (Abhari et. al, Ellahham, Rigla et. al). According to Dankwa-Mullan et. al, the main areas to which artificial intelligence has been applied in diabetes care include predictive population risk stratification, clinical decision support, patient self-management tools, and retinal screening. Out of these four, all but predictive population risk stratification focus on patients either diagnosed with or at risk for diabetes, and less on the screening and diagnostic process that I investigated through my analysis. With regard to the work done on predictive population risk stratification, studies have been conducted to investigate optimal methods for diabetes screening and diagnosis. These studies mostly focused on artificial neural networks, logistic regression, random forests, support vector machines, and the k-nearest neighbors algorithm — all of which I investigated in Stage 2 of this analysis. However, the analysis that I conducted is unique in that it targets ICU patients, and so was bound by time and computing constraints to a degree that prior analyses were not. Since the utility of this analysis lay in its ability to provide information quickly in the absence of verified medical histories or records, it was crucial for it to be able to generate results in clinical settings, where computing resources are scarce relative to AI research labs and time is of the essence. Previous work in this area paid little attention to model explainability and/or interpretability, which is an issue that I address in this project.

Motivation and Problem Statement

In the early days of the COVID-19 pandemic, flaws in healthcare systems around the world were made painfully apparent. Healthcare workers worldwide struggled as hospitals were overloaded by patients in critical condition, and lives that could have been saved were lost. On top of this, many factors prevented, and still continue to prevent, hospitals and/or intensive care units (ICUs) from administering treatment immediately. The WiDS team recognized that "...ICUs often lack verified medical histories for incoming patients. A patient in distress or a patient who is brought in confused or unresponsive may not be able to provide information about chronic conditions such as heart disease, injuries, or diabetes. Medical records may take days to transfer, especially for a patient from another medical provider or system." As such, any information that can be provided in a timely manner about chronic conditions such as diabetes can help healthcare workers make better clinical decisions and consequently lead to an improvement in patient outcomes, potentially saving lives.

In addition to the obstacles mentioned above, research in the area of AI/ML for diabetes diagnosis tends to trade off either model interpretability or accuracy. Through this project, I sought to reduce the need for this trade-off by introducing a final step in data science pipelines, where models were interpreted post hoc to give clinicians and even patients a better idea of how predictions were generated. This step has the potential to increase trust in models that are typically considered "black boxes", as it computes and visualizes the influence of each input label on the models' output predictions.

Methods, Data and Approach

Data Used

The dataset I used was taken from MIT's GOSSIS (Global Open Source Severity of Illness Score Initiative, and is available publicly on Kaggle. To quote the Competition Data Access and Use Restriction section of the datathon rules: "The Dataset contains clinical measurements of a sample of U.S. patients, and has no dates or geographic locations. Uniquely associated identifiers appear, but these values were randomly assigned to provide consistency in the dataset. These are not encrypted or hashed values." Under Kaggle's terms of data access and use (found under the Competition Data section of the datathon rules), the dataset is restricted to use for the WiDS datathon as well as academic research and education. The connection between this dataset and my problem statement should be self-explanatory; indeed, the problem statement was shaped by the data available in this dataset.

A common consideration around medical data is whether patient details have been sufficiently deidentified. As documented by MIT GOSSIS, "...it is hereby certified that the Dataset, as received by Dr. [Latanya] Sweeney, and dated December 11, 2020, meets the current standards for de-identification under the federal regulations known as the Safe Harbor Provision of the Privacy Rule of the Health Information Portability and Accountability Act (HIPAA) regarding the confidentiality of healthcare data. This certification is based on the standards established for the Safe Harbor Provision of the HIPAA regulation." Another ethical consideration, in this case, had to do with the fact that the dataset only contained measurements of a sample of U.S. patients. Of course, this was not ethically problematic in itself, but could have been if the conclusions drawn from this data were applied to other populations without regard for differences that may have arisen. Since my analysis focused on an academically-oriented exercise that was not intended to be generalizable to other populations, this did not become a concern.

Methodology

1. Basic data preprocessing and analysis

There were six columns with categorical data in the dataset I used. I one-hot encoded these columns, since there were no ordinal relationships between the values in these columns. Following this, I filled in the NaN values in the dataframe with the means of the columns they were in. Finally, I normalized the data so all numerical values were between 0 and 1. This step helped standardize the ranges and variances of each variable, which in turn prevented variables with wider ranges or higher variances from over-influencing the results of models that classified based on the distance between points.

2. Comparison of common classifying and diagnostic models

In this stage, I tested the CatBoost Classifier, Light Gradient Boosting Machine, Extreme Gradient Boosting, Gradient Boosting Classifier, Ada Boost Classifier, Random Forest Classifier, Linear Discriminant Analysis, Logistic Regression, Extra Trees Classifier, Naive Bayes, Decision Tree Classifier, K Neighbours Classifier, Quadratic Discriminant Analysis, SVM - Linear Kernel, and Ridge Classifier. I tested these models with PyCaret, a low-code framework.

3. Feature selection

I conducted feature selection by training an LGBM model on the cleaned data and using the feature importance property of the model to extract the top X most important features. After running analyses on several subsets of features, I decided to use the top 50 features for the rest of my analysis. This step of the process reduced the number of variables used to train each top-performing model. Anecdotally, I noticed that this helped to improve the performance of the models, hence making them more functional for a clinical setting. Feature selection can also occasionally make results more accurate, though this was not something I observed.

4. Tuning and training top performers

I trained the CatBoost and LGBM models, tuning each of them to achieve optimal performance. In both this stage and stage two, the models were evaluated based on the area under the receiver operating curve (AUROC), which shows the trade-off between true positive rate (TPR) and false positive rate (FPR) across different decision thresholds (Glassbox Medicine) and is a typical performance metric in studies working with medical data.

5. Ensemble learning

Ensemble learning is a machine learning approach that combines two or more models, aiming to improve the accuracy of predictions by combining the strengths of the corresponding models. However, model ensembles do not necessarily perform better and are often more costly to train and deploy. They also tend to be less interpretable, which can decrease clinician confidence. I explored two main methods of ensemble learning: simple weighted ensemble learning, which takes the output probabilities of separate models and combines them in a weighted sum; and ensemble learning with models.

6. Interpreting and explaining model results

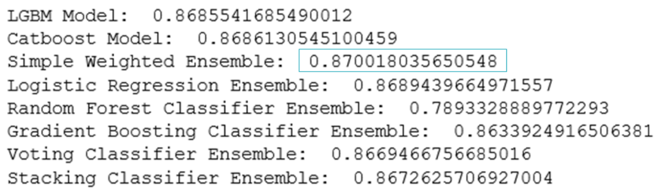
After receiving feedback on my project proposal, I decided to add this phase of the project. Using a Python framework called shap, I analyzed the predictions that both the LGBM and CatBoost models made based on the test set. I was able to obtain beeswarm plots indicating the influence of each input label on the models' outputs, as well as waterfall plots indicating the influence of input labels on individual predictions made by the models.

Discussion

Findings and Implications

In stage two of this analysis, the CatBoost Classifier and Light Gradient Boosting Machine models were selected as final contenders for tuning and training in stage four. The main considerations for model selection were the area under the curve (AUC) and time taken (TT), as this analysis aimed to use models that could be trained in clinical settings — where both time and computing resources are scarce — while still maximising the balance of benefits to costs. The data generated in this stage is available in Appendix 1.

After the selection in stage two, the CatBoost and LGBM models were tuned and trained further. The graphic below lists AUCs of these models, as well as those achieved through ensemble learning methods in stage five. The simple weighted ensemble (which combined the probabilities predicted by each model in a 50/50 weighted sum) performed the best, with an AUC of 0.87002. The CatBoost and LGBM models were individually able to achieve AUCs of 0.86861 and 0.86855 respectively. Model ensembling methods did not generally improve the AUC, with only the logistic regression ensemble performing better than the individual models at an AUC of 0.86894.



```
LGBM Model:  0.8685541685490012
Catboost Model:  0.8686130545100459
Simple Weighted Ensemble:  0.870018035650548
Logistic Regression Ensemble:  0.8689439664971557
Random Forest Classifier Ensemble:  0.7893328889772293
Gradient Boosting Classifier Ensemble:  0.8633924916506381
Voting Classifier Ensemble:  0.8669466756685016
Stacking Classifier Ensemble:  0.8672625706927004
```

The interpretability analysis was conducted on both the CatBoost and LGBM models. Beeswarm plots visualizing the impact of each of the top twenty input labels on the output predictions of the model were mostly similar for both models.

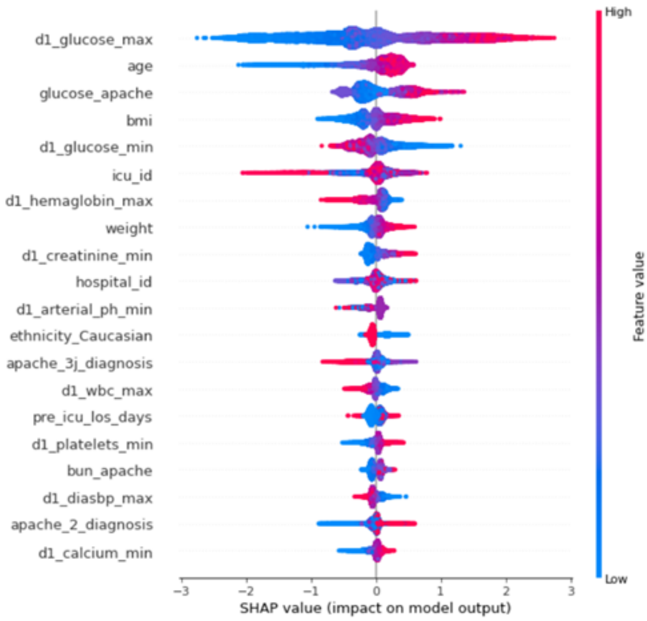


Figure 2: Interpretability analysis of the LGBM model

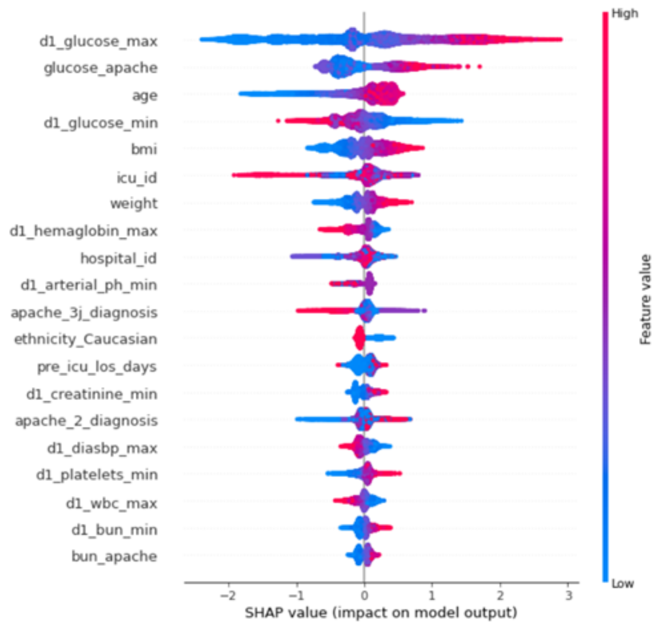


Figure 3: Interpretability analysis of the CatBoost model

Following this, waterfall plots visualizing the impact of each input label on individual predictions were also generated. Though the order of importance of each input label varied from the results indicated in the interpretability analyses of the aggregate predictions generated by each model, the results from these plots were still mostly in line with the aforementioned analyses. Two examples, one with a positive f(x) and one with a negative f(x) (as generated by shap) are shown here:

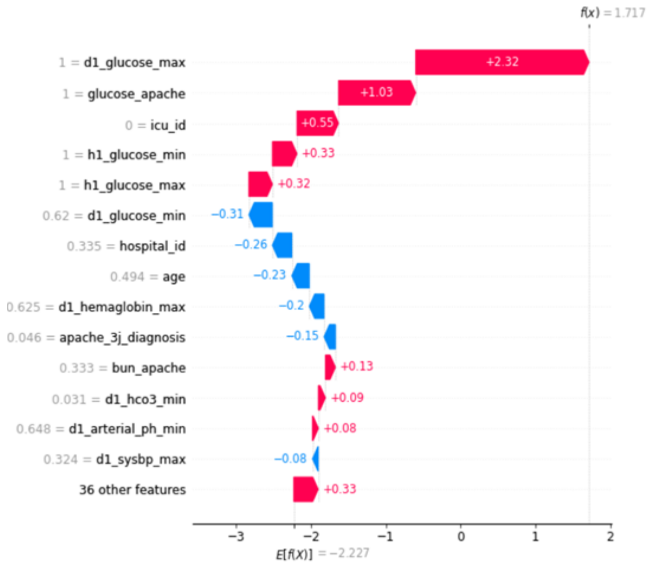


Figure 4: Interpretability analysis of an individual prediction generated by the CatBoost model (positive f(x))

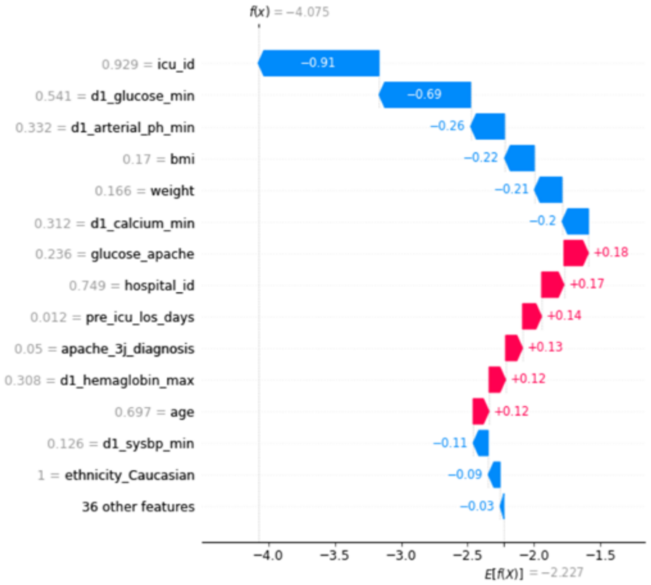


Figure 5: Interpretability analysis of an individual prediction generated by the CatBoost model (negative f(x))

A major practical contribution of this research is that it opens the door for more complex, “black-boxy” models to be used in the field of medical AI research. Previous work in this field tended to trade off either model interpretability or accuracy; this analysis reduces the need for this trade-off. Though model interpretability is difficult to achieve with the most accurate tree ensemble models, this analysis suggests that post hoc explanatory methods such as SHapley Additive exPlanations (SHAP) can mitigate this issue by generating explanations for both aggregate and individual outputs. Accordingly, the analysis conducted on this dataset suggests that AI/ML methods for prediction of diabetes mellitus among ICU patients show promise for deployment in clinical settings.

Another important implication of this analysis is the demonstrated potential of post hoc explainability and/or interpretability methods for use as part of patient-facing AI systems. Since these methods can generate explanations for individual patients as well as aggregate predictions, they might be used by clinicians employing predictive AI systems to both better understand the results generated by these systems and to explain these results to patients. Thus, the methods used in this project show potential for inspiring greater clinician and patient confidence in medical AI.

Limitations and Potential Directions for Future Research

This analysis is subject to several limitations. Firstly, the models trained on the dataset are likely not generalizable to other populations without significant retraining. This is due to the specificity of certain input labels (e.g. those related to ethnicity), as well as potential differences in sociocultural context that might influence the prevalence of diabetes mellitus in differing nations or regions. It also remains unclear whether the dataset itself contained significant bias which might have affected the training of the models.

Secondly, the feature selection process may not have been completely representative of the data. Of the six labels for ethnicity (generated through one-hot encoding), only one — 'ethnicity_Caucasian' was among the top 50 features that eventually formed the final dataset. In light of the fact that ethnic Caucasians might be at significantly higher risk of diabetes mellitus, this label was not dropped; however, since this label was the 12th most important in generating model predictions for both the LGBM and CatBoost classifiers, it begs the question of why others were not considered as important. The feature selection process also selected the 'Unnamed_0' label, which only represented the index of each row in the dataset. This label was dropped prior to model training, and it remains unclear why a label which should have been completely uncorrelated to the predicted outcomes was selected as part of this process.

Finally, the models trained as part of this project were not prospectively tested in an actual clinical setting. However, this remains a metric which is crucial to evaluating the real-world performance of these models. In light of this, future research might focus on training models and/or creating an AI prediction system for clinical deployment. Other gaps remaining in the literature that future research might seek to fill include conducting quantitative or qualitative evaluations of bias that might be present in this and similar datasets, optimizing models for limited training time and computing resources, and improving the generalizability of models such as those used in this analysis without sacrificing the valuable information that is provided by "local" or region-specific data.

Reflection

Although it may seem like human-centered data science does not come into play in this project until stage six (where the results generated by each model are interpreted), HCDS principles informed my decision-making at each stage of this project. While selecting the dataset I would work with for this project, I paid close attention to ethical considerations including the deidentification of the dataset, the data heterogeneity, and whether the collected data respected the personal autonomy of each patient. The dataset I selected was deemed sufficiently deidentified by HIPAA standards and the only identifying information therein was given by randomly generated IDs used to differentiate between patients. However, the dataset only contained measurements from a sample of U.S. patients, and so was not heterogenous enough to be generalized to other populations (as addressed in the Discussion section).

When examining the intermediate results from each stage of this analysis, I applied a human-centered lens. An example of this is found in the feature selection process in stage three, where I note that the only ethnicity label included in the top fifty most important features is 'ethnicity_Caucasian'. Though this alone was not enough to indicate sample bias, I believe that it raises a relevant question about how feature selection techniques currently operate and whether the comparatively larger proportion of Caucasians in this dataset may have influenced this result.

Finally and most obviously, I applied human-centered data science principles to the interpretation and explanation of each model's predictions. This allowed me to better understand how each input variable influenced the results generated by the models. Though I was not able to quantify whether the models were biased, this gave me a qualitative look at the inner workings of both models. Besides this, HCDS principles also informed my choice of visualizations. I made efforts to select plots that would be simple and display the relevant relationships clearly so that this section would be accessible to people with less statistical and computing knowledge. Overall, I would say that HCDS principles were central to both the planning and execution of my project.

References

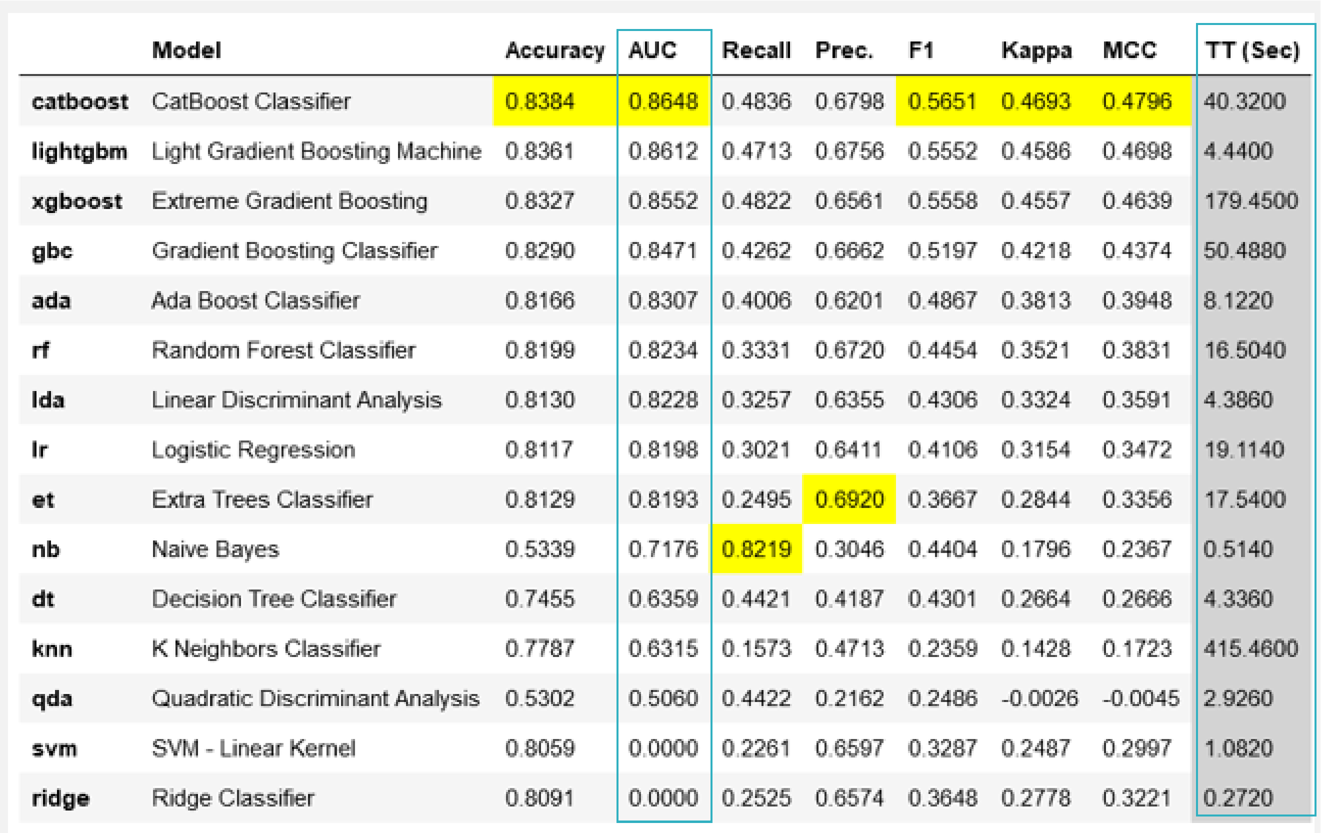
Abhari S, Niakan Kalhori SR, Ebrahimi M, Hasannejadasl H, Garavand A. Artificial intelligence applications in type 2 diabetes mellitus care: focus on machine learning methods. *Healthc Inform Res.* 2019;25(4):248. doi:10.4258/hir.2019.25.4.248

Dankwa-Mullan I, Rivo M, Sepulveda M, Park Y, Snowdon J, Rhee K. Transforming diabetes care through artificial intelligence: the future is here. Population Health Management. 2019;22(3):229-242. doi:10.1089/pop.2018.0129

Ellahham S. Artificial intelligence: the future for diabetes care. The American Journal of Medicine. 2020;133(8):895-900. doi:10.1016/j.amjmed.2020.03.033

Rigla M, García-Sáez G, Pons B, Hernando ME. Artificial Intelligence Methodologies and Their Application to Diabetes. Journal of Diabetes Science and Technology. 2018;12(2):303-310. doi:10.1177/1932296817710475

Appendix 1



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.8384	0.8648	0.4836	0.6798	0.5651	0.4693	0.4796	40.3200
lightgbm	Light Gradient Boosting Machine	0.8361	0.8612	0.4713	0.6756	0.5552	0.4586	0.4698	4.4400
xgboost	Extreme Gradient Boosting	0.8327	0.8552	0.4822	0.6561	0.5558	0.4557	0.4639	179.4500
gbc	Gradient Boosting Classifier	0.8290	0.8471	0.4262	0.6662	0.5197	0.4218	0.4374	50.4880
ada	Ada Boost Classifier	0.8166	0.8307	0.4006	0.6201	0.4867	0.3813	0.3948	8.1220
rf	Random Forest Classifier	0.8199	0.8234	0.3331	0.6720	0.4454	0.3521	0.3831	16.5040
lda	Linear Discriminant Analysis	0.8130	0.8228	0.3257	0.6355	0.4306	0.3324	0.3591	4.3860
lr	Logistic Regression	0.8117	0.8198	0.3021	0.6411	0.4106	0.3154	0.3472	19.1140
et	Extra Trees Classifier	0.8129	0.8193	0.2495	0.6920	0.3667	0.2844	0.3356	17.5400
nb	Naive Bayes	0.5339	0.7176	0.8219	0.3046	0.4404	0.1796	0.2367	0.5140
dt	Decision Tree Classifier	0.7455	0.6359	0.4421	0.4187	0.4301	0.2664	0.2666	4.3360
knn	K Neighbors Classifier	0.7787	0.6315	0.1573	0.4713	0.2359	0.1428	0.1723	415.4600
qda	Quadratic Discriminant Analysis	0.5302	0.5060	0.4422	0.2162	0.2486	-0.0026	-0.0045	2.9260
svm	SVM - Linear Kernel	0.8059	0.0000	0.2261	0.6597	0.3287	0.2487	0.2997	1.0820
ridge	Ridge Classifier	0.8091	0.0000	0.2525	0.6574	0.3648	0.2778	0.3221	0.2720