ISyE 6740 – Spring 2021
Final Project Report

Team Member Name: Daniel Bae (903564139, dbae32)

Project Title: Applications of SVM Beyond Classification Demonstrated via Clinical Data on Dementia

Background:

Dementia is a general term describing a wide group of symptoms related to cognitive impairment and/or cognitive decline. Early detection and planning are crucial as in some cases there exists methods of prevention, treatment, and even reversal of symptoms. There has been success in using analytics for preclinical detection of cognitive change using the popular "clock drawing test", which tracks the movement of a digital pen as participants draw a clock and analyzes the data to determine the risk of cognitive decline. This method has made great strides in systematizing a technique for early detection. Since the process is digital, there is potential for it to be administered remotely, it captures minute details imperceptible to the human eye, and data can be collected from participants over repeated trials among other advantages over traditional methods of diagnosis. Advances in the set of analytics tools available in the effort to minimize the impacts of these diseases have significant potential.

Problem Statement:

This project proposes the application of kernel support-vector machines on data collected from participants that exhibit various levels of dementia in order to discover insights not limited to classification.

Support-vector machines (SVM) are a relatively robust and efficient linear classification tool. With the use of kernels, such as the popular radial basis function (rbf), SVM can also perform non-linear classification. By attempting to maximize the distance between classes, SVM focus on the points nearest or on the boundaries between the classes. These points are known as the support vectors. These support vectors are the "relevant" points that are used to model the boundaries. All other points are, typically, not required by the model and their weights are zero ergo SVM are often described as being memory efficient. A component of this project is to build upon the existing applications of and research in SVM for purposes not limited to classification by using SVM to understand more about the geometry of high dimensional data and observing relationships between features and the label. Specifically, leverage the ratio of support vectors to non-support vectors to determine the relative "roughness" of data and observe change in distance from support vectors given a small change in the value of a feature of a non-support vector to ascertain relative feature importance and relationships between features and the label. By applying these techniques to the data collected from participants demonstrating various levels of dementia, extract information in addition to classification.

Data Source:

The dataset leveraged to test the utility of SVM, in addition to classification, is the "OASIS-1: Cross-sectional MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults" dataset from the "Open Access Series of Imaging Studies" (OASIS) project. Demographic, clinical, and derived data was collected from 436 participants. There are 11 features in total, however, not all of the features were included in the experiment. For example, all of the participants were right-handed therefore the feature on dominant handedness was excluded. The dataset also exhibits missing data for roughly half of the participants. Participants with missing data were not included in this experiment. The total final population was 216 participants each with 8 features. Below is a figure representing the 8 features and the class labels.

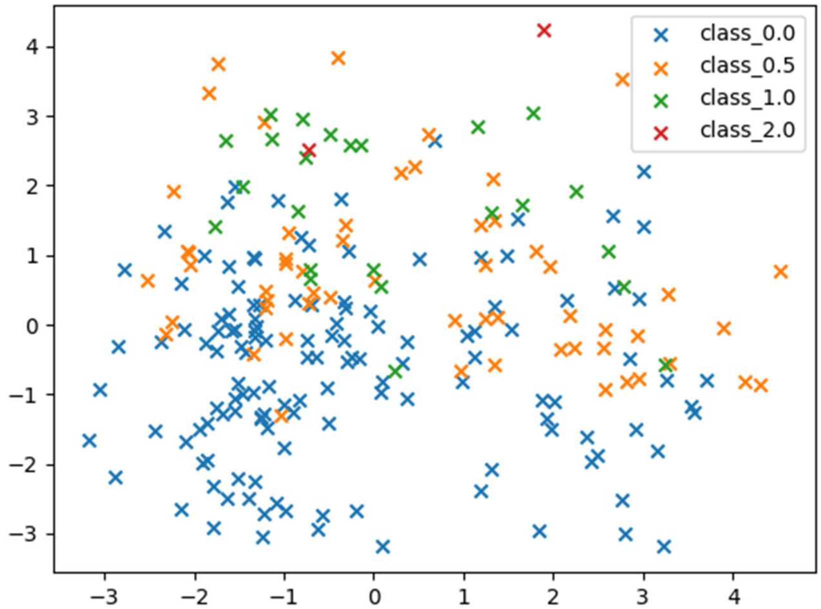
figure 1:

	Field Name	**Full Name**	**Values**
Demographics	Gender	Gender	M/F
	Age	Age	Int
	Educ	Education	Int
	SES	Socioeconomic Status	Int
Clinical/Derived	MMSE	Mini-Mental State Examination	Int
	eTIV	Estimated Total Intracranial Volume	Int
	ASF	Atlas Scaling Factor	Float
	nWBV	Normalized Whole Brain Volume	Float
Class Labels	CDR	Clinical Dementia Rating	0.0 = No Dementia 0.5 = Very Mild Dementia 1.0 = Mild Dementia 2.0 = Moderate Dementia

Preprocessing:

Initially, the intention was to conduct a multiclass experiment. However, the number of individuals diagnosed with mild or moderate dementia were extremely low compared to the other groups. For instance, the number of individuals (without missing data) diagnosed with moderate dementia was only 2. In addition, separating individuals based on the severity of their dementia proved difficult given the data available. This made it incredibly challenging to develop a model that could classify any more accurately than a coin toss. It's possible that some combination of regression and classification may have addressed this issue. Instead, the labels were reassigned into two classes resulting in a binary classification problem. If an individual demonstrated no dementia, they remained in class 0. All other individuals were reassigned to class 1. *Figure 2* was generated by leveraging PCA in order to reduce the dimensionality of the data and to plot it in two dimensions. It shows the challenge of separating classes 0.5, 1.0, and 2.0 without overfitting. The lessons learned from this experiment should be generalizable, therefore overfitting was carefully avoided.

figure 2:



One of the features indicates the gender of the participant with the letter "M" for males and "F" for females. This feature was transformed into two separate binary columns, "Male" and "Female" respectively, using one-hot encoding. Then the features were standardized, column-wise, by subtracting the mean and dividing by the standard deviation.

The data was then split into a training set comprising 70% of the data with the remaining 30% being reserved for testing. Since the dataset is relatively small, proper randomization during the shuffling process was critical. Changes in the random seed value, in some cases, resulted in very different results. Various random seeds were tested using k-fold cross-validation in order to deduce a good seed to use and such that the results of the experiments could be reproduced.

Methodology:

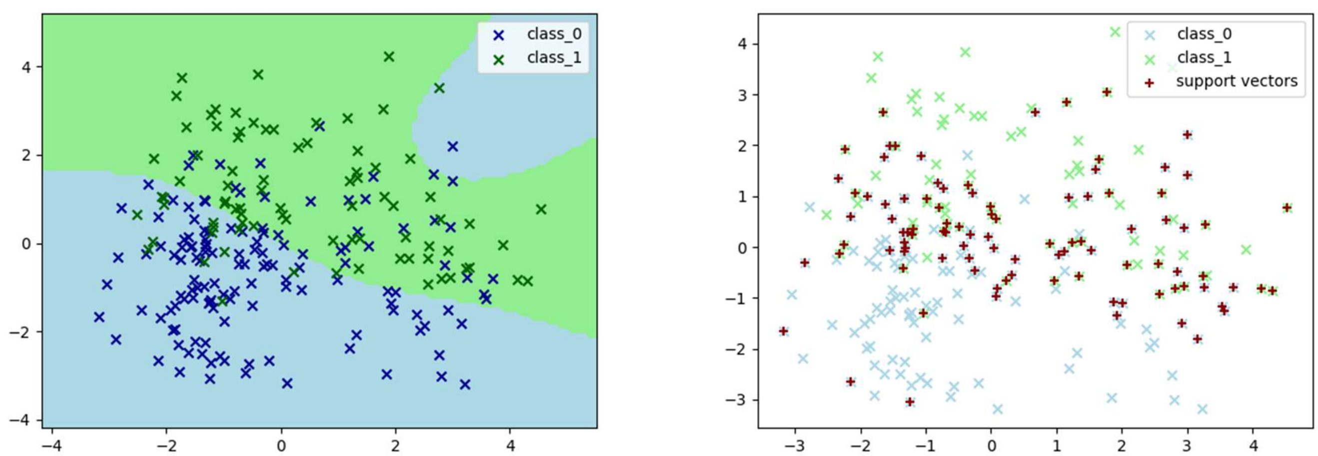
Once preprocessing had been completed, it was critical to train a "good" SVM classifier. The model had to balance bias on the training set with the variance on the testing set in order for experiments on its ability to describe relationships between classes and features to be accurate and generalizable. This was achieved using k-fold cross-validation on the training data to tune the "C" and "gamma" parameters. The most optimal SVM model using the rbf kernel in this scenario was determined to have the following parameter values.

figure 3:

Parameters	Values
C (regularization parameter)	2
gamma (kernel coefficient)	"scale" (1 / (n_features * X.var()))

Given the somewhat limited data, the model still produced decent results on the new binary classification problem, achieving accuracy of approximately 86% on the testing data without overfitting.

figure 4:



The mutual information (MI) between each of the feature variables and the label variable was calculated and the values are represented in *figure 5*. The MI between each of the variables "Male" and "Female" and the label variable is less than half that of the next variable with the lowest MI, "SES" (socioeconomic status). A participant's age and their mini-mental state examination scores (MMSE) seem to contain the most "information" about the likelihood of exhibiting or developing some level of dementia.

figure 5:

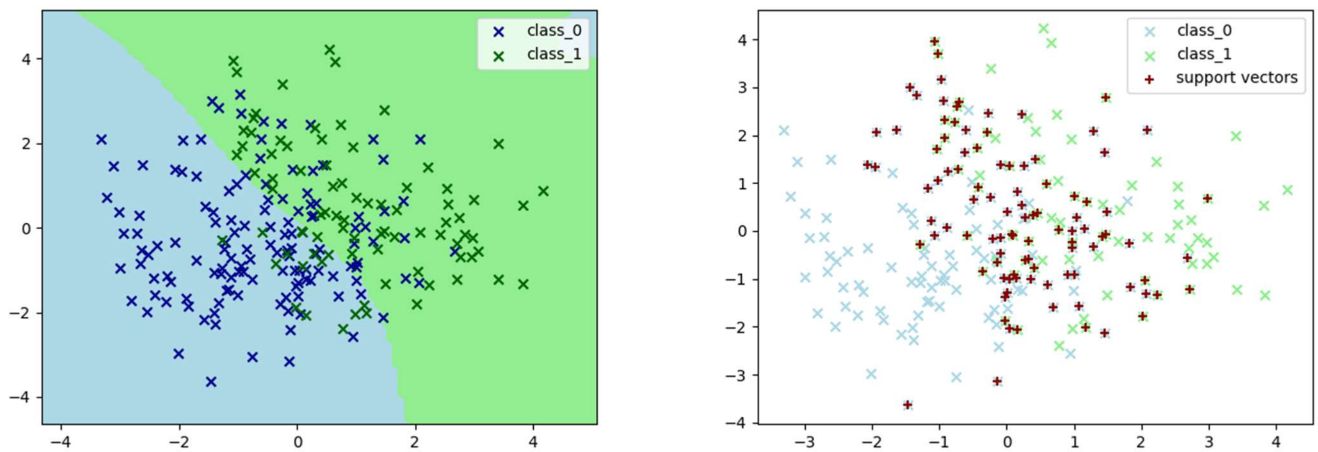
	Male	**Female**	**Age**	**Educ**	**SES**	**MMSE**	**eTIV**	**nWBV**	**ASF**
MI	0.009167	0.009167	0.124343	0.024994	0.020131	0.326894	0.030062	0.152706	0.033987

Based on the MI, a new model was trained on the feature data excluding the "Male" and "Female" datapoints. The optimal parameter values obtained using k-fold cross-validation for the new model is given by *figure 6*. The new model exhibited noticeable improvements over the model that used all of the features. The new model achieved accuracy of around 92% on the testing data, still without overfitting.

figure 6:

Parameters	**Values**
C (regularization parameter)	1
gamma (kernel coefficient)	~0.1329 (using the "median trick")

figure 7:

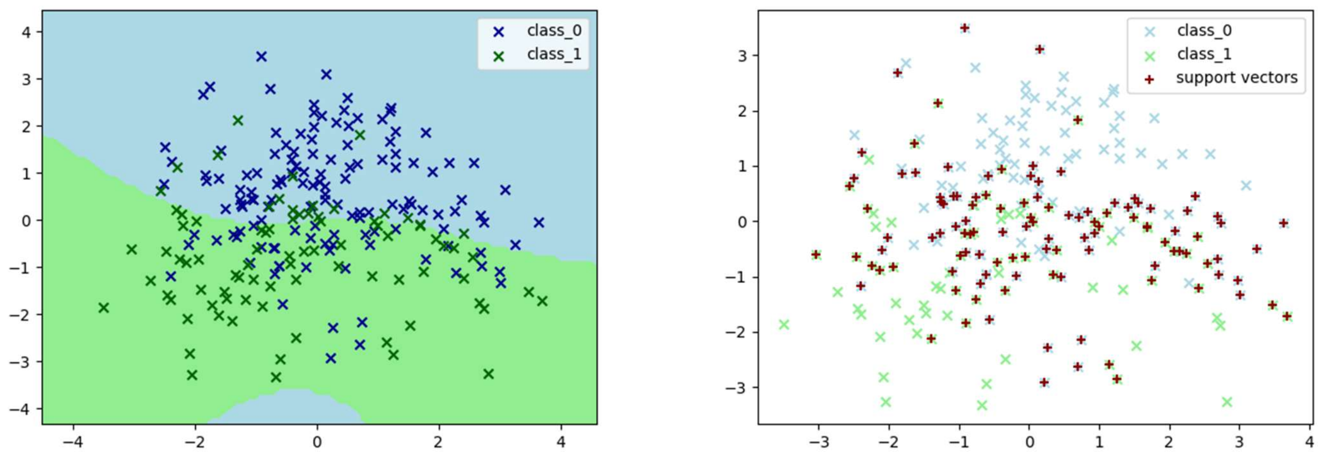


In order to determine whether the same or similar conclusions could be drawn by observing changes in feature values and their influence on distances from support vectors, samples from the original model (using all features) were taken. Thirty samples were taken from each class for a total of sixty samples. All sixty samples were datapoints that were not used as support vectors by the model. For each of these non-support vectors, the KNN algorithm was leveraged to find each of their nearest support vectors and their Euclidean distances. For each sample non-support vector, their features were individually incremented by some percentage, p, keeping all other features unchanged. The absolute value of the change in distance to the nearest support vector was recorded. Since all of the features had been standardized, the same patterns emerged regardless of p's value but for this experiment a p value of 100% was used. Once all of the features of every sample had been tested, the results were aggregated into each feature's average effect on the distance from the nearest support vector. For each class, a feature's average effect was converted into a percentage of the class average. Then the max percentage value for each feature from either of the classes was taken.

figure 8:

	Class 0		Class 1		
	Avg change	**As % of avg**	**Avg change**	**As % of avg**	**max**
Male	3.209111	-1.97%	3.878869	0.97%	0.97%
Female	3.209111	-1.97%	3.878869	0.97%	0.97%
Age	3.470096	6.01%	3.612093	-5.98%	6.01%
Educ	3.329584	1.71%	3.912346	1.84%	1.84%
SES	3.285406	0.36%	3.914493	1.90%	1.90%
MMSE	3.028923	-7.47%	4.054968	5.55%	5.55%
eTIV	3.318904	1.39%	3.803229	-1.00%	1.39%
nWBV	3.256597	-0.52%	3.699018	-3.71%	-0.52%
ASF	3.353774	2.45%	3.820877	-0.54%	2.45%
Avg	3.273501		3.841640		

figure 9:



For the final part of the experiment, the ratio of support vectors to non-support vectors was calculated for both the model using all features as well as the model using the subset of features found using MI.

figure 10:

	Model	
	All Features	**Feature Subset**
Total # of training datapoints (ttl_dat)	151	151
Total # of support vectors (ttl_sv)	91	99
ttl_sv / ttl_dat	60.26%	65.56%
# of training datapoints from class 0 (class_0_dat)	91	91
# of support vectors from class 0 (class_0_sv)	54	58
class_0_sv / class_0_dat	59.34%	63.74%
# of training datapoints from class 1 (class_1_dat)	60	60
# of support vectors from class 1 (class_1_sv)	37	41
class_1_sv / class_1_dat	61.67%	68.33%

Evaluation and Final Results:

The method of evaluating features by observing their influence on distances between non-support vectors and their nearest support vectors produced interesting results (see *figure 8*). They aligned with results generated using MI. Both suggest that “Age” and “MMSE” are potentially more informative and “Male” and “Female” are potentially less informative. Where the two methods did not align was on the feature “nWBV” (normalized whole brain volume). The method being tested can be interpreted as implying that “nWBV” may be less important as well. Something that MI did not reflect. A new model was trained, now excluding the features “Male”, “Female”, and “nWBV”. The results were nearly identical to the results obtained using all of the features, with an accuracy of around 86% on the testing data. It seemed that there was no improvement but at the same time, removing these features did not appear to reduce the efficacy of the model or the informativeness of the feature set suggesting that there was some potential truth to the outcome.

Based on the ratios of support vectors to non-support vectors (*figure 10*), both models used over 50% of the datapoints as support vectors. This suggests that a majority of points lie on or within the margins between the classes. This is confirmed by the 2d plots produced using PCA (*figure 4*, *figure 7*). The plots show that the two classes are not neatly separable and that there is mixing between the classes giving some validity to the findings. The ratios also show that the model using a subset of features found through MI uses a larger portion of the datapoints as support vectors. This is not as clear just by observing the plots but makes logical sense in terms of SVM as typically, reducing the number of features will result in less variance and therefore datapoints being more densely concentrated increasing the likelihood of finding a datapoint on or within the margins between the classes. It is possible to see that the support vectors in *figure 7* are more densely clustered than the support vectors in *figure 4*.

For more conclusive results, repetition and thorough experimentation and testing on more data is required, however, both experiments point to potential for promising results.

Conclusion:

There is often no single best method to achieve a goal in every analytics scenario. A combined narrative often outperforms any single method. Having multiple methods to tackle a problem can improve the amount learned overall. SVM as classification tools can bolster the effort to detect cognitive decline early and accurately. Furthermore, by observing the relationship between the support vectors and the unweighted points, this experiment suggests that it may be possible to learn even more about the underlying data.

SVM can sometimes be characterized as memory efficient due to the fact that non-support vectors are unweighted but it was possible to build on the body of work that already exists on using SVM beyond classification and show that the distance between non-support vectors and support vectors given changes in their features could produce results similar to a generally accepted method of using MI. The ratios of weighted and unweighted points were also able to describe aspects of the data that could also be observed by plotting the data in 2d using PCA. PCA, ISOMAP, and other dimensionality reduction techniques often require careful consideration of metadata such as whether the data exhibits nonlinearity or contains binary/categorical features for instance. Some information is often lost when reducing data from higher dimensions to one that can be visually interpreted by human beings. Therefore, it is important to have a diverse range of tools available for describing data.

Continued and thorough testing and experimentation must be conducted before any conclusions are considered remotely definite or even generally accepted. However, the results of these experiments demonstrated via clinical data on dementia further suggest, in addition to existing works, that there is in fact potential for SVM to provide value beyond simple classification.

Works Cited:

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

OASIS brains. (n.d.).
Retrieved from https://www.oasis-brains.org/#data

Hyunseokc. (2018, March 06). DETECTING EARLY ALZHEIMER'S.
Retrieved from https://www.kaggle.com/hyunseokc/detecting-early-alzheimer-s

Informs. (n.d.). Detecting Preclinical Cognitive Change.
Retrieved from https://www.informs.org/Impact/O.R.-Analytics-Success-Stories/Detecting-Preclinical-Cognitive-Change

Bron, E. E., Smits, M., Niessen, W. J., & Klein, S. (2015). Feature Selection Based on the SVM Weight Vector for Classification of Dementia. *IEEE journal of biomedical and health informatics*, *19*(5), 1617–1626. https://doi.org/10.1109/JBHI.2015.2432832

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2