Name: Izzy Hargrav, Grant Yap

Emails: ihargrav@u.rochester.edu, gyap@u.rochester.edu

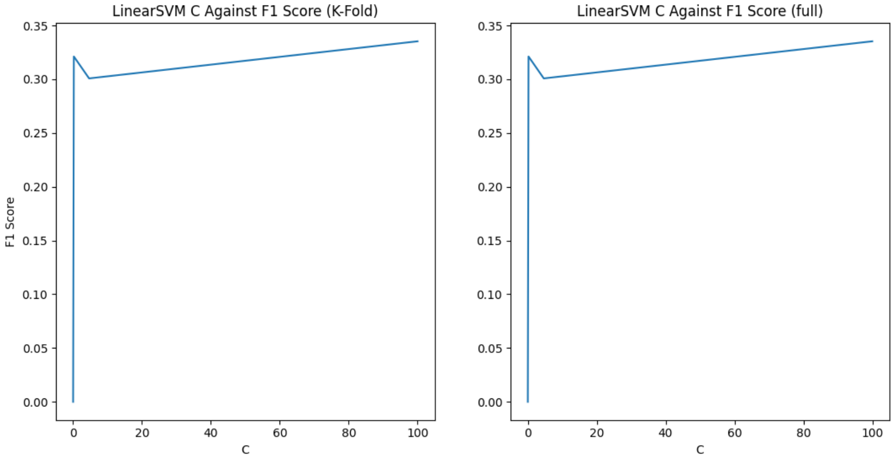
svm-mlp
Project 2 for Machine Learning

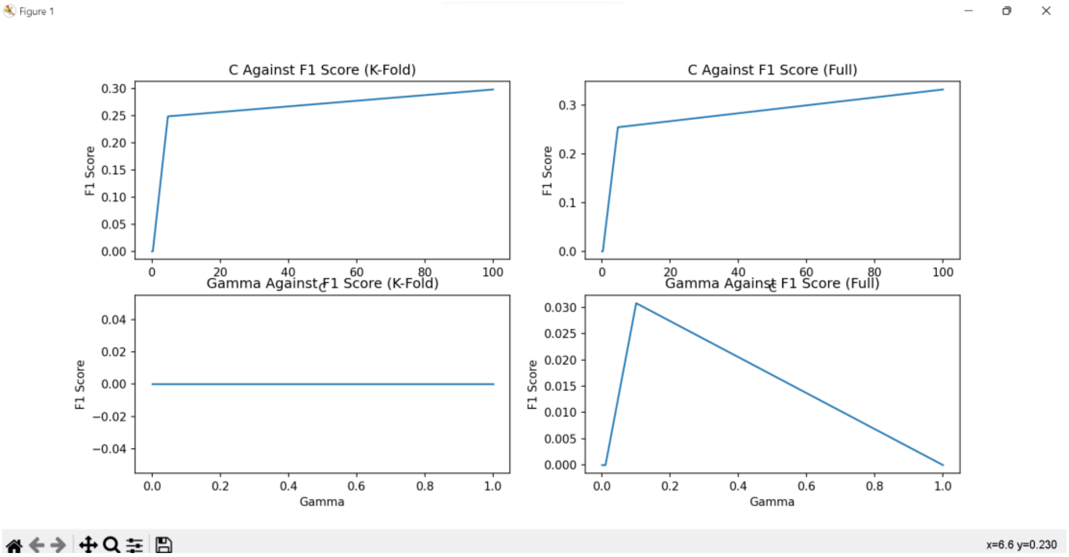
Table of Contents
- load_newts.py - professor given code to process data, modified by us to become usable.
- svm_frogs.py - script with functions for linear svm and kernel svm on newt datset
- svm_grid_search.py - parameter search for linear and kernel SVM, also generates graphs.
- mlp_frogs.py - script with functions for mlp with cross validation and sweeping.
- plot_sweep_results.py - contains the data and prints the graph for the unit sweep for the mlp

SVMs
For the SVMs section of this project, we used sklearn's LinearSVC and SVC with the RBF kernel. We played with all of the parameters to find the best combinations and swept through values of C and gamma to find the best performance.

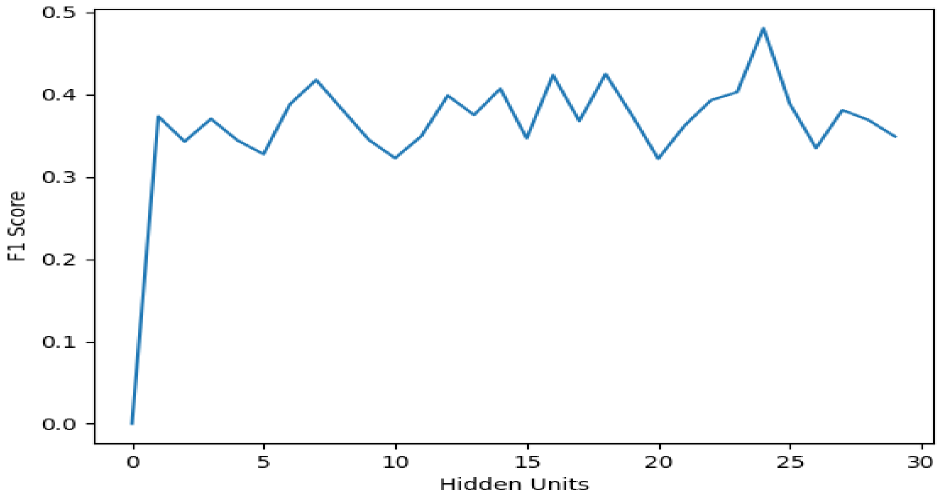
LINEAR SVM
k-fold f1_score: 0.37584371184371185
full f1_score: 0.6526315789473685

KERNEL SVM
k-fold f1_score: 0.28233938736785036
full f1_score: 1.0





```
## MLPs
For the MLPs section of this project, we used sklearns' model section to
use Kfold and metrics to use f1 score as well as keras to emulate a
multilayer perceptron. We played with all of the parameters between
hidden layers, batch size, epochs, and different optimizers and we
settled on using Adam, as well as a hidden unit amount of 30 to get more
consistent results. We also found with this amount, we always got 100%
accuracy on the training set with our model.
Cross-validated F1 Score: 0.469888827912788
```



Best Recommendation
We would recommend an MLP for this type of problem because we got an F1 Score of 0.469888827912788 when using k-fold cross validation in comparison to a value of 0.28233938736785036 when we tried cross validation on kernel SVMs. Due to the results that we found, we would recommend the MLP due to its better performance. As a caveat, the SVMs were easier to get to work, train, and get more consistent results in general.

Our chosen set (stroke_detection)
Our data set we choice takes a variety of factors that are known to be related to strokes such as life styles, genetics, predisposition, and the task is to determine whether someone will have a stroke in their life time based on the factors mentioned. We chose to use a MLP as we got great performance on our validated dataset using cross validation with an F1 score of 0.9512720108032227. Investigating the state of the art from the citation, we saw that we met their state of the art at 0.95.

Citation:

Fedesoriano. (2021, January 26). Stroke prediction dataset. Kaggle. Retrieved November 11, 2022, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Bhuvanchennoju. (2021, June 15). 📣 💡 🎨 data storytelling 🎯AUC focus on 🩸 strokes. Kaggle. Retrieved November 11, 2022, from https://www.kaggle.com/code/bhuvanchennoju/data-storytelling-auc-focus-on-strokes