Machine Learning Feature Selection and Analysis

Austin Dase
Department of Computer and Information Sciences
Towson University
Towson, United States
adase1@students.towson.edu

Dataset

- Wine Quality Dataset from the University of California, Irvine Machine Learning Repository
- Previously used by P. Cortez et. al. [2]
- 11 input features, one class label
- Class labels range from 0-10
  - 0 (poor quality) 10 (excellent quality)
- Features: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol

Experimental Environment

- Windows 10
- Anaconda3, Python 3.6.4
  - TensorFlow 1.8 [8]
  - Scikit-learn 0.19.1 [7]
  - numpy
- Intel® Core™ i5-8600K 6 Core CPU
- NVIDIA GeForce GTX 1060 6GB GPU
- 16GB DDR4 Balistx RAM

Motivations

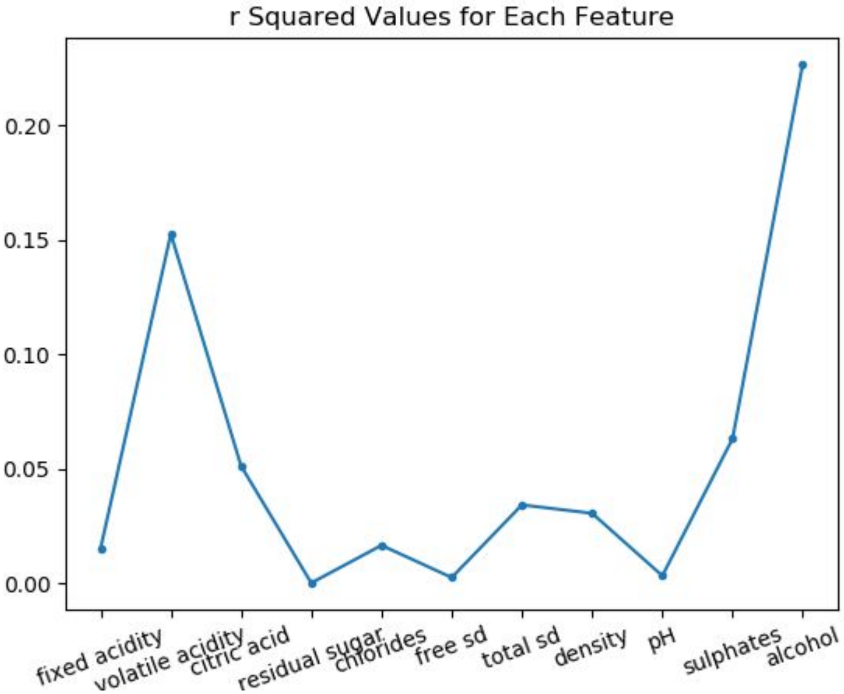
- Learn how quantitative qualities of wine impact subjective quality score
- Understand the sense of tastes better
- Provide insight into ‘black box’ Machine Learning (ML) models
- Research applicable to the wine quality data, but also to anyone looking to draw conclusions about the ‘meaning’ of the model a classifier builds

Approach - Outline

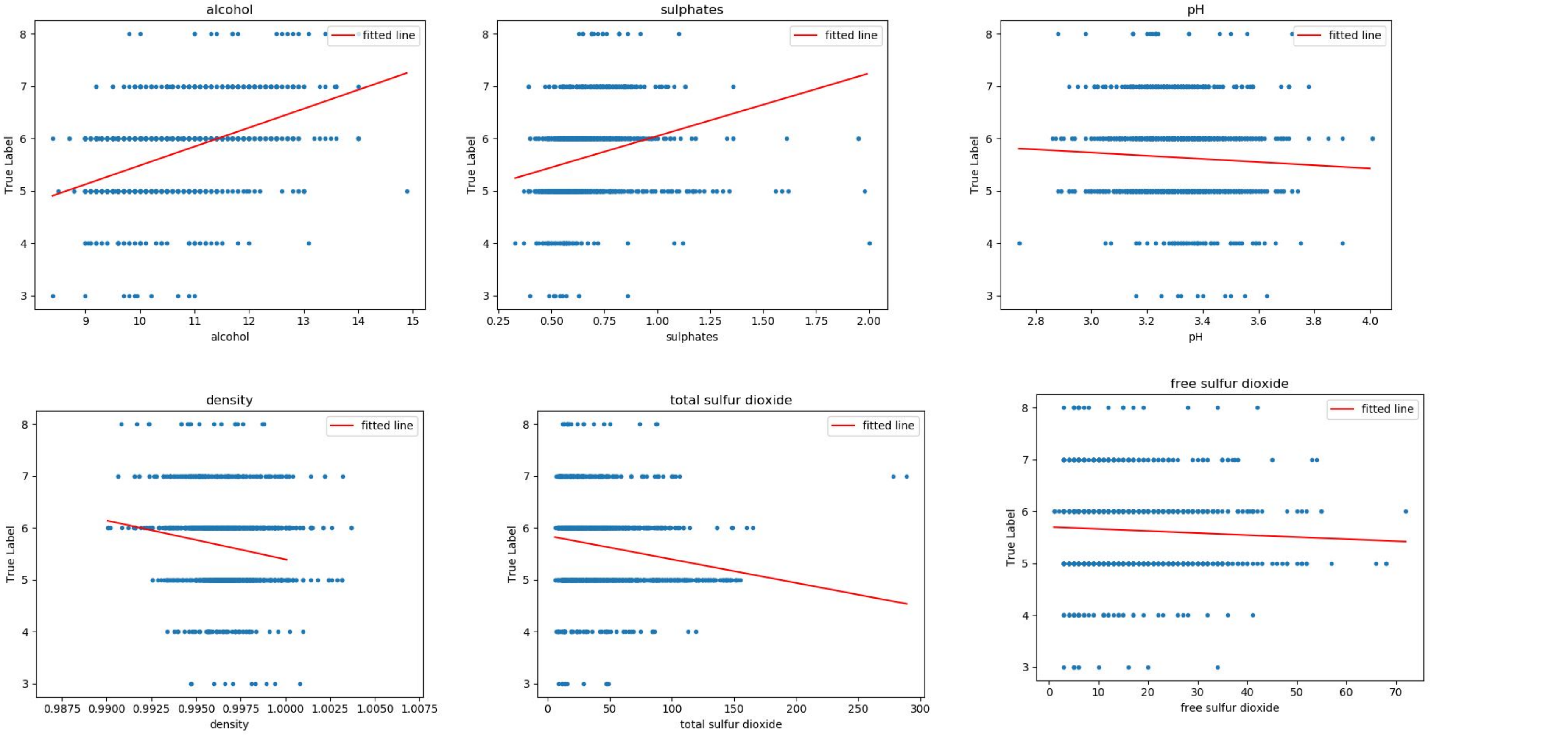
- EDA to gain insight, have a benchmark and make predictions
- Build optimal ML classifiers; Neural Network (NN), Support Vector Machine (SVM), K-Nearest Neighbor (KNN)
- Train and test the classifiers for every possible subset combination of features
  - Re-train for every subset
  - Re-split the data into train and test for each subset
- Record the test accuracy for each subset, for each classifier
- Analyze the test accuracy for each classier
  - As a whole as well as for each subset

Approach - EDA

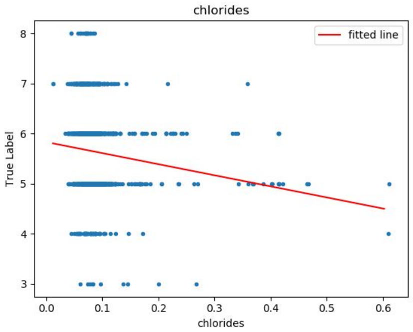
- First examine the feature’s correlation to their class label
  - Make predictions based on this information, also use as a benchmark for comparison

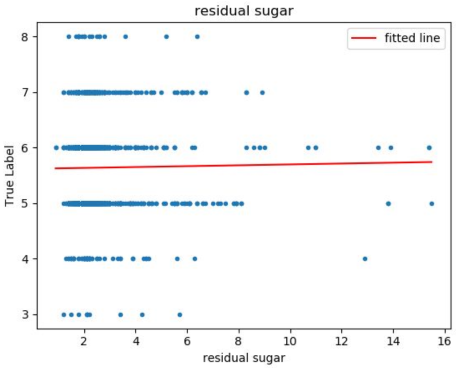


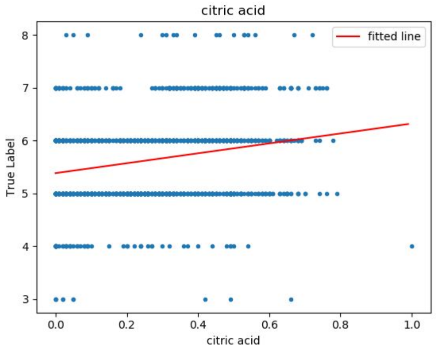
Approach - EDA 2

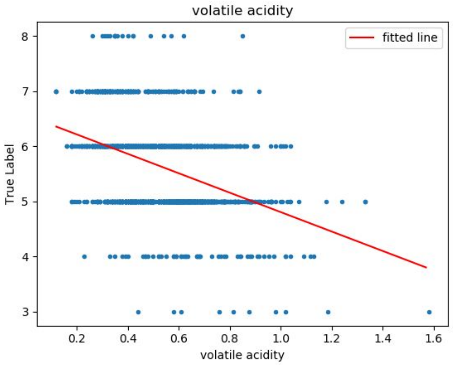


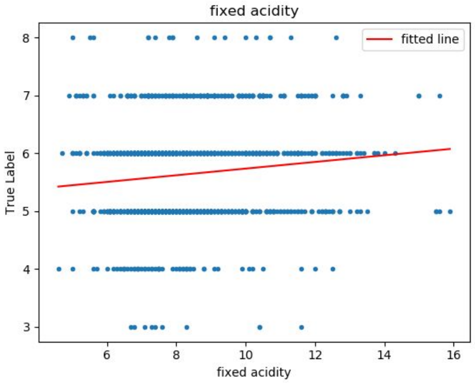
Approach - EDA 2





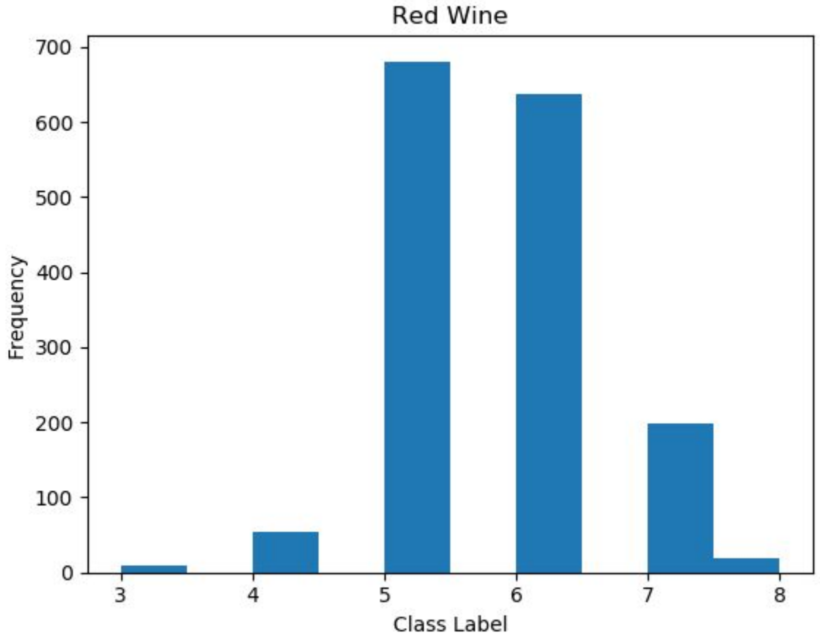






Approach - EDA 3

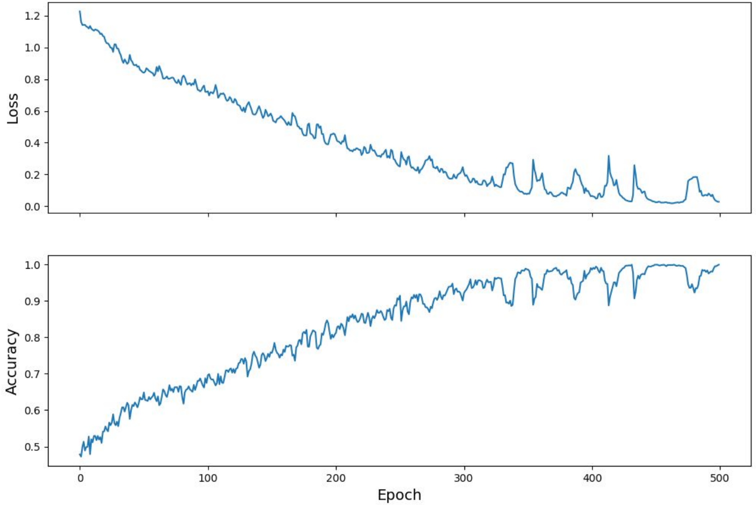
- Histogram of counts



Approach - Build the Model

- NN
  - Tensorflow, 5 Hidden Layers, 200 nodes at each layer, 50 iterations of testing
  - More complex model with higher iterations achieved higher accuracy (99%), but long execution
  - Baseline accuracy: 58.26%
- SVM
  - Scikit-learn, Radial Basis Function (RBF) kernel, C = 3, all others defaults
  - Baseline accuracy: 83.98%
- KNN
  - Scikit-learn, k = 15, all others defaults
  - 57.63% accuracy

NN Training Metrics

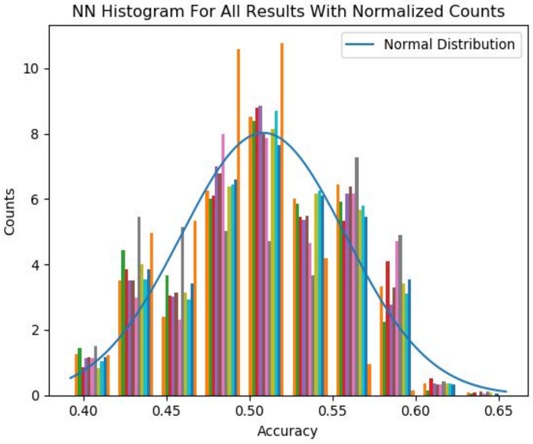


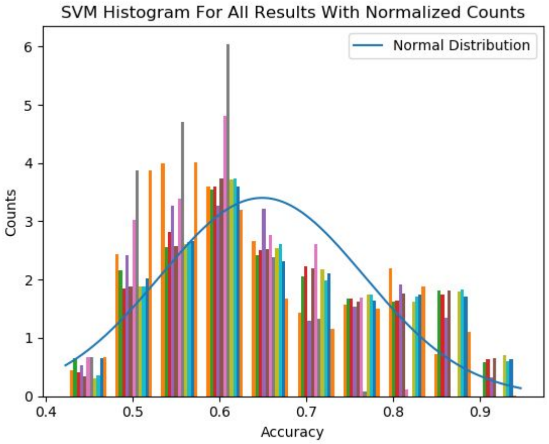
Approach - Storing the Results

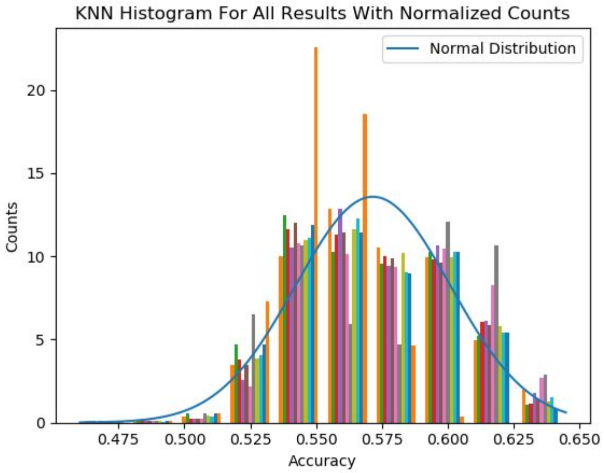
- 1024 x 11 matrix
- Each column is a feature
- 1024 rows for 2^{n-1} possible subset combinations

Approach - Analyze Accuracy Results 1

- First, plot the results in aggregate

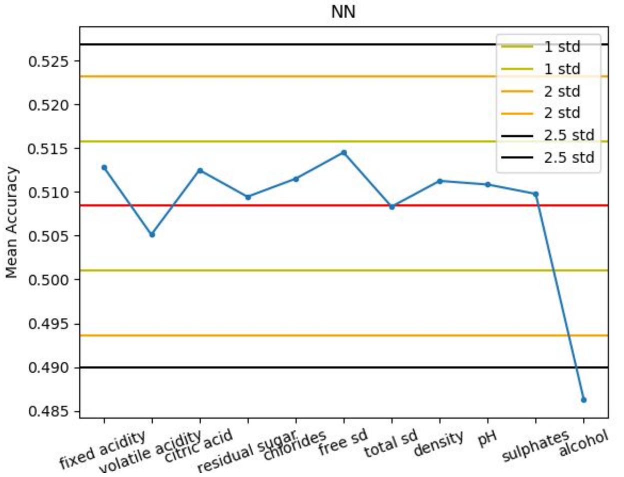

NN Histogram For All Results With Normalized Counts

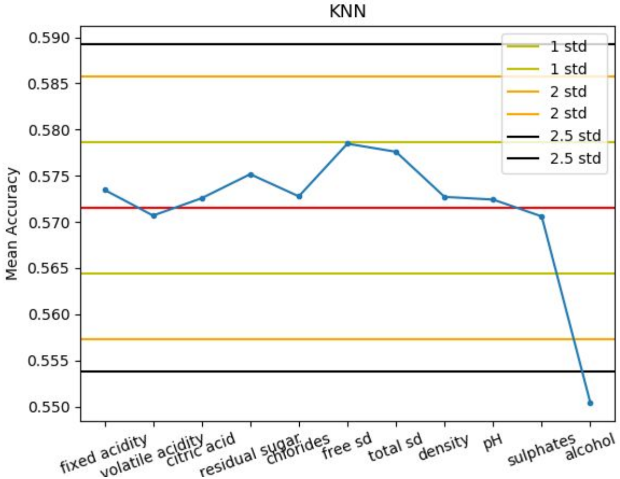

SVM Histogram For All Results With Normalized Counts

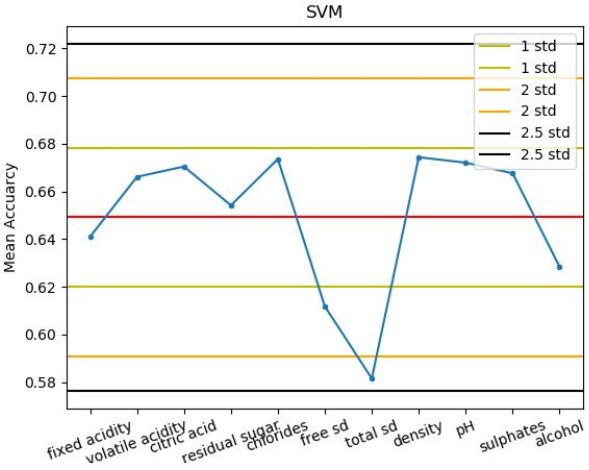

KNN Histogram For All Results With Normalized Counts

Approach - Analyze Accuracy Results 2

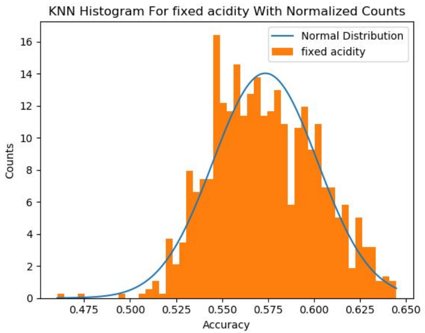
- Second, plot the means of each feature (columns)

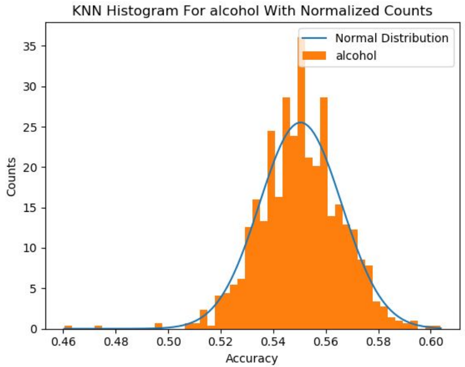


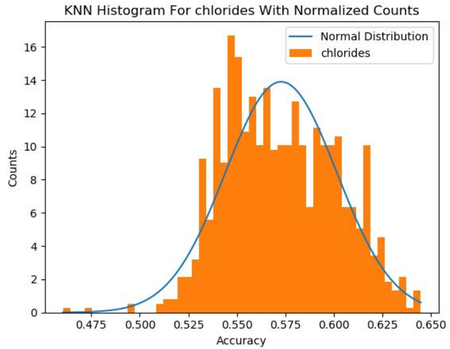


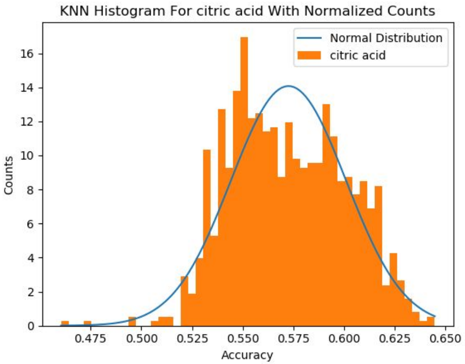


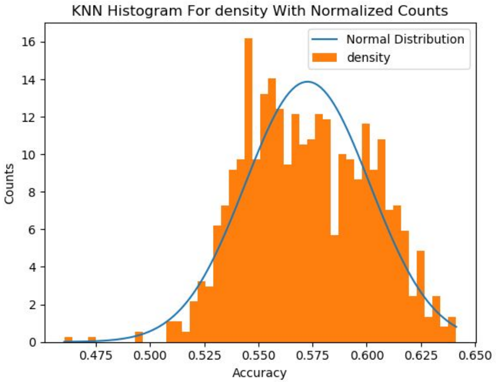
Analyze By Feature - KNN -1

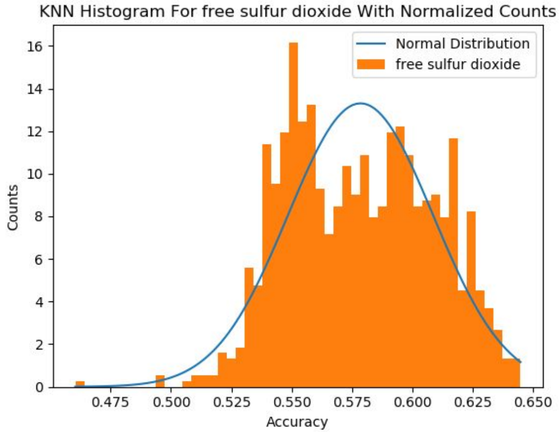

KNN Histogram For fixed acidity With Normalized Counts


KNN Histogram For alcohol With Normalized Counts

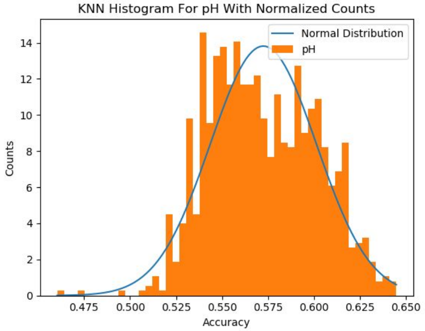

KNN Histogram For chlorides With Normalized Counts

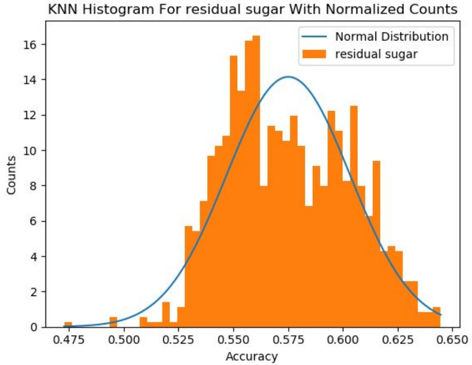

KNN Histogram For citric acid With Normalized Counts

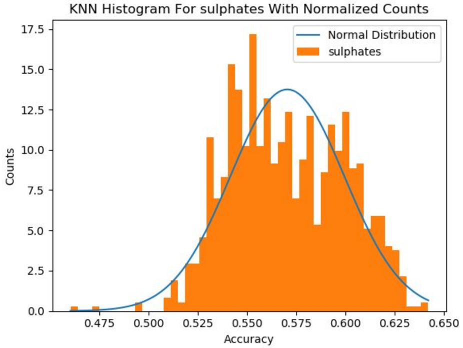

KNN Histogram For density With Normalized Counts

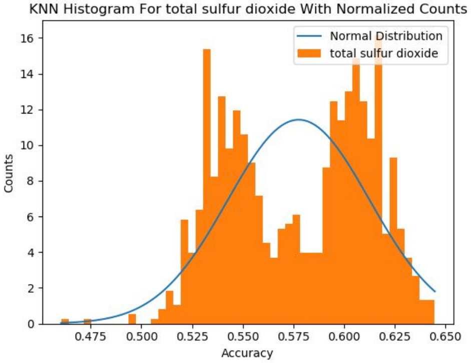

KNN Histogram For free sulfur dioxide With Normalized Counts

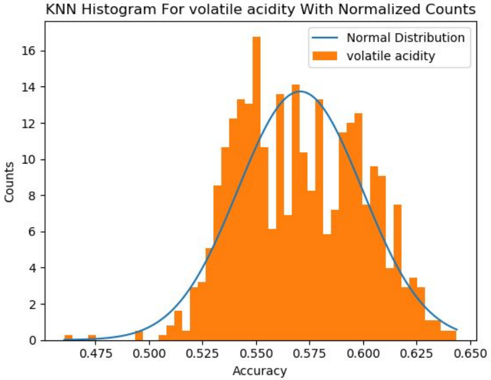
Analyze By Feature - KNN -2



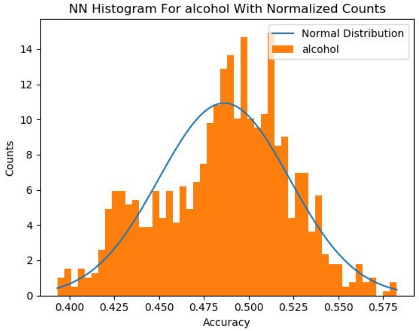


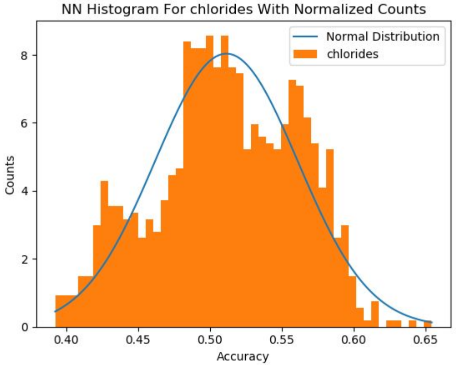


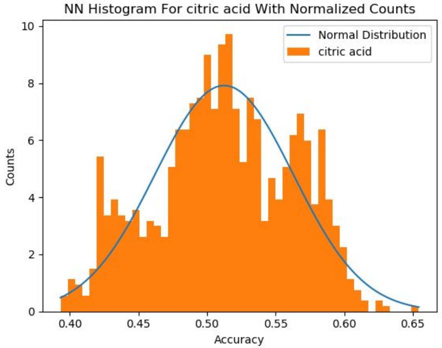


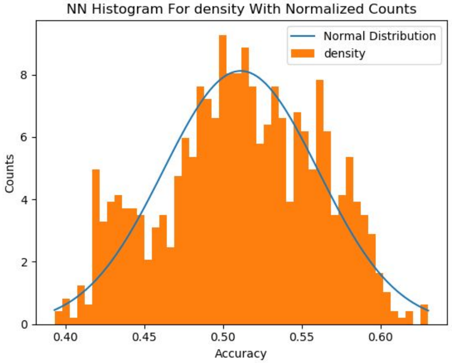


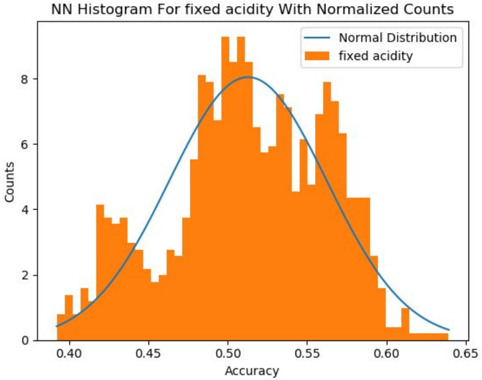
Analyze By Feature - NN - 1

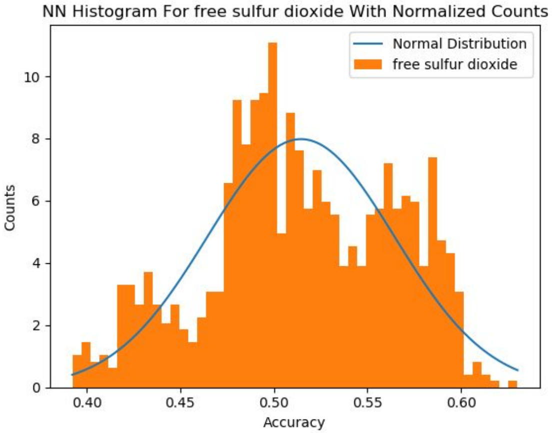

NN Histogram For alcohol With Normalized Counts


NN Histogram For chlorides With Normalized Counts

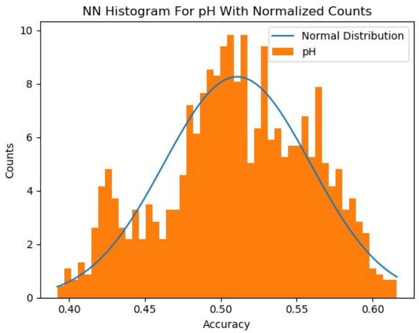

NN Histogram For citric acid With Normalized Counts

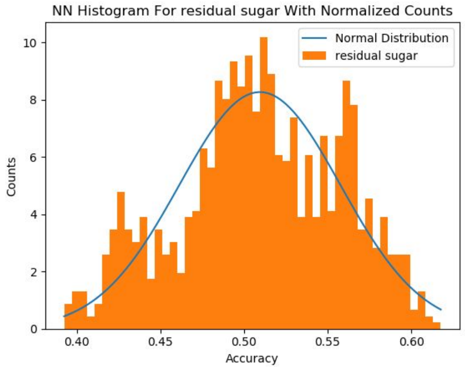

NN Histogram For density With Normalized Counts

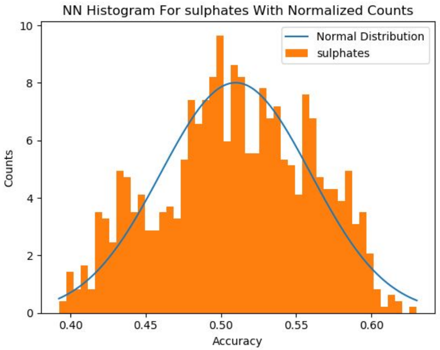

NN Histogram For fixed acidity With Normalized Counts

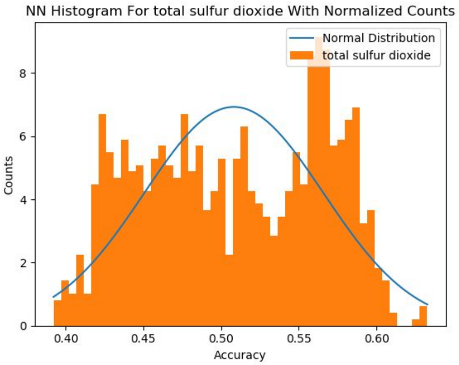

NN Histogram For free sulfur dioxide With Normalized Counts

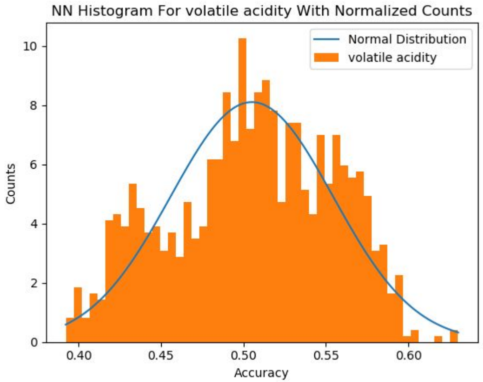
Analyze By Feature - NN - 2



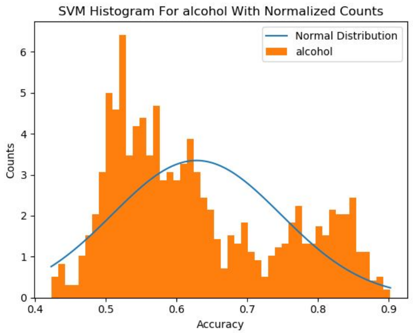


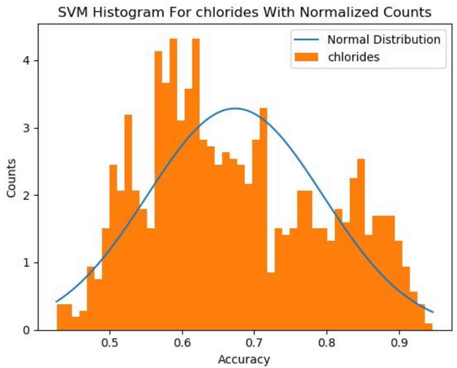


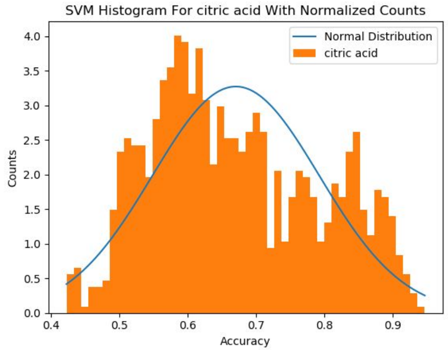


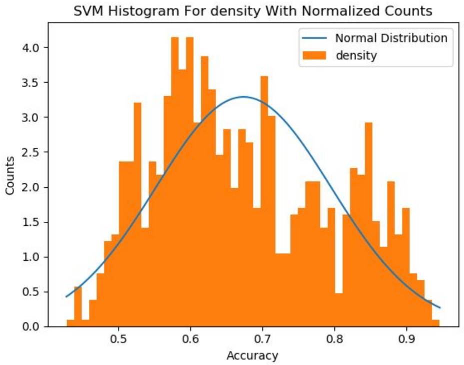


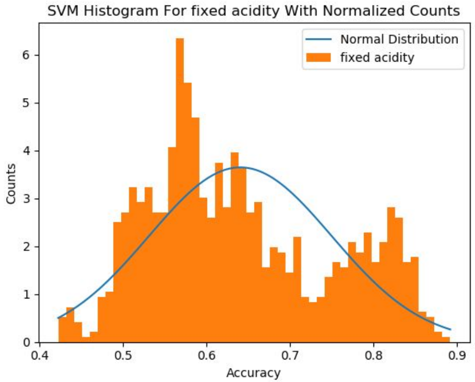
Analyze By Feature - SVM - 1

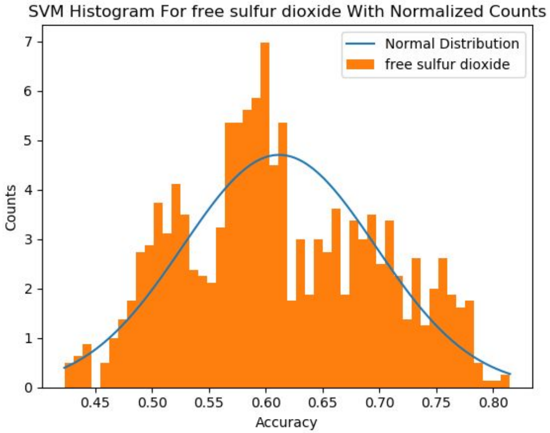




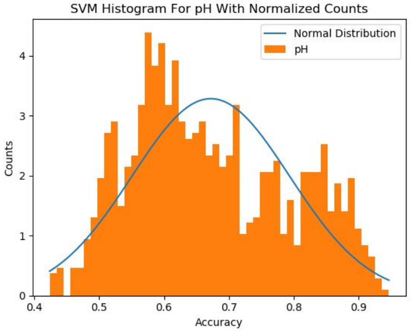


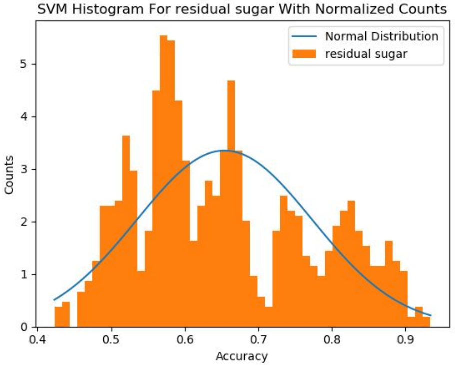


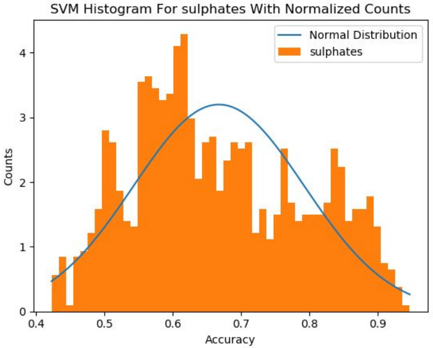


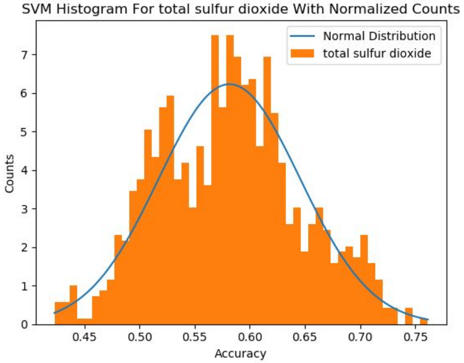


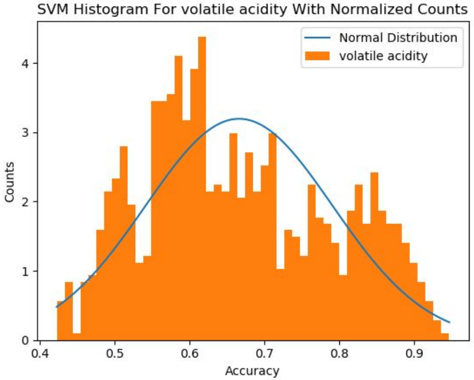
Analyze By Feature - SVM - 2











Analysis

Were any of the distributions normal ?

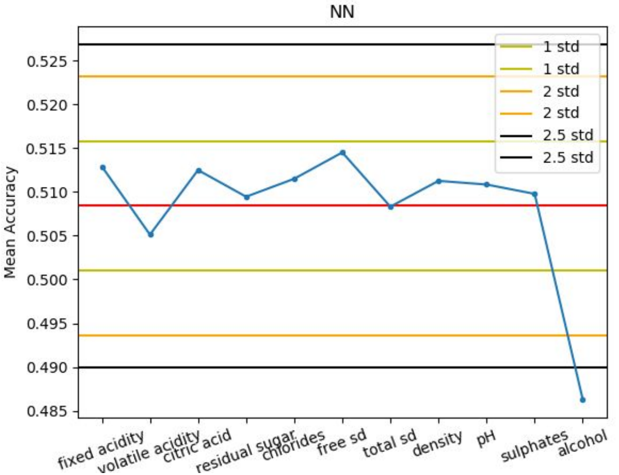
- Shapiro–Wilk test of normality
- In short - no

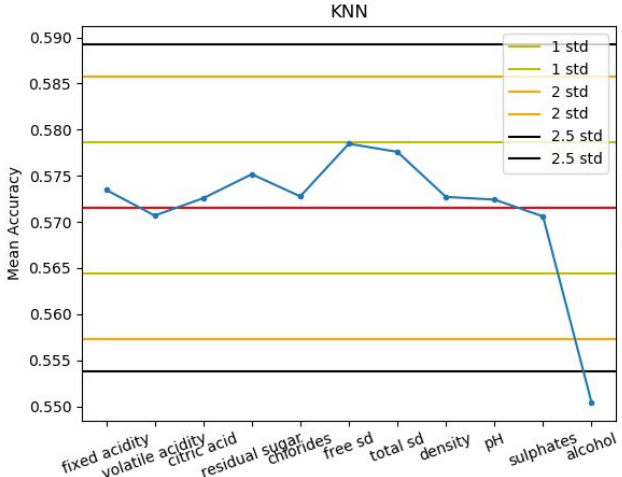
Test of Normality of Results Distribution - NN

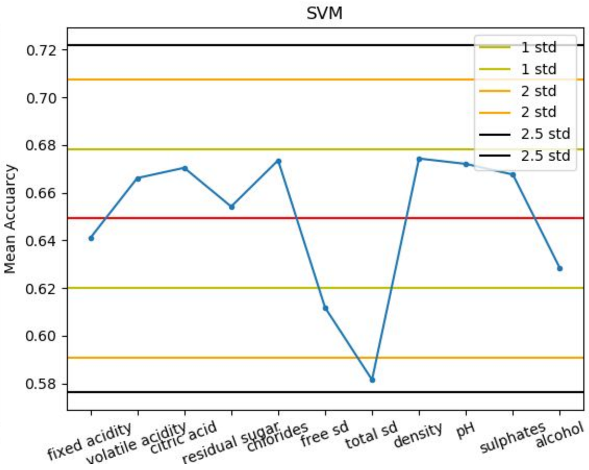
Feature	Test Stat	p-value
Fixed Acidity	.99	2.235e-6
Volatile Acidity	.98	8.928e-10
Citric Acid	.98	9.959e-10
Residual Sugar	.98	3.444e-10
Chlorides	.98	2.511e-9
Free Sulfur Dioxide	.97	1.485e-11
Total Sulfur Dioxide	.94	2.619e-19
Density	.98	3.072e-8
pH	.98	5.543e-9
Sulfates	.98	1.584e-10
Alcohol	.98	2.69e-7
Aggregrate (All Features)	.97	3.622e-37

Discussion and Conclusion

- Regression predictions; Alcohol and Volatile Acidity will have greatest impact
- Alcohol had the largest impact on accuracy for NN and KNN
  - This confirmed the prediction based on regression
- Total Sulfur Dioxide had the largest impact on accuracy for SVM
  - Alcohol also had a less significant impact than for NN and KNN
- Interesting to note the similarity between NN and KNN and not SVM
  - Does this mean we gain insight into the model, or to the real-world ?







References

[1] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386-408.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Modeling wine preferences by data mining from physicochemical properties,” Decision Support Systems, vol. 47, no. 4, pp. 547–553, 2009.

[3] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” Artificial Intelligence, vol. 97, no. 1-2, pp. 273–324, 1997.

[4] R. Setiono and H. Liu, “Neural-network feature selector,” IEEE Transactions on Neural Networks, vol. 8, no. 3, pp. 654–662, 1997.

[5] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” Computers and Electrical Engineering, no. 40, pp. 16–28, 2014.

[6] M. Hall, “Correlation-based Feature Selection for Machine Learning.” Unpublished.

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

[8] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.