Problem Set #8

MACS 30100, Dr. Evans
Due Monday, Mar. 12 at 11:30am

1. **Neural network horse race (10 points).** For this problem, you will test the predictive accuracy of three models on classifying wines into one of three possible *cultivars*.The data in the file strongdrink.txt. The data are comprised of 176 observations, each of which is a chemical analysis of an Italian wine. Each wine is from one of three known cultivars (a cultivar is a group of grapes selected for desirable characteristics that can be maintained by propagation). The chemical analysis determined the quantities of the following 13 different constituents (the last 13 variables):

Variable	Name	Variable	Name
Alcohol	`alco`	Nonflavanoid phenols	`nonfl_phen`
Malic acid	`malic`	Proanthocyanins	`proanth`
Ash	`ash`	Color intensity	`color_int`
Alkalinity of ash	`alk`	Hue	`hue`
Magnesium	`magn`	OD280/OD315 of diluted wines	`OD280rat`
Total phenols	`tot_phen`	Proline	`proline`
Flavanoids	`flav`		

   (a) Create a scatterplot of the data where the x-variable is alcohol (*alco*) and the y-variable is color intensity (*color_int*). Make the dot of each of the three possible *cultivar* types a different color. Make sure your plot has a legend.

   (b) Use `sklearn.linear_model.LogisticRegression` to fit a multinomial logistic model of `cultivar` on features alcohol (*alco*), malic acid (*malic*), total phenols (*tot_phen*), and color intensity (*color_int*) with the following linear predictor.

$$Pr(cultivar_i = j|X\beta_j) = \frac{e^{\eta_j}}{1 + \sum_{j=1}^{J-1} e^{\eta_j}} \quad \text{for} \quad j = 1, 2$$

$$\text{where} \quad \eta_j = \beta_{j,0} + \beta_{j,1} alco_i + \beta_{j,2} malic_i + \beta_{j,3} tot\_phen_i + \beta_{j,4} color\_int_i$$

   Use k-fold cross-validation to estimate the MSE of the multinomial logit model.

```
    clf_mlog = KFold(n_splits=4, shuffle=True, random_state=22)
```

   Play with the tuning parameter values `penalty` and `C` to get the lowest possible k-fold MSE. Report your minimized overall MSE along with the tuning parameter values you used for `penalty` and `C`.

(c) Use `sklearn.ensemble.RandomForestClassifier` to fit a random forest model of `cultivar` on the same four features used in part (b). set `bootstrap=True`, set `oob_score=True`, and set `random_state=22`. Use OOB cross-validation to generate the MSE of your random forest classifier. Play with the values of the tuning parameters `n_estimators`, `max_depth`, and `min_samples_leaf` to try and find the lowest possible MSE from the OOB cross validation. Report your minimized overall MSE along with the tuning parameter values you used for `n_estimators`, `max_depth`, and `min_samples_leaf`.

(d) Use `sklearn.svm.SVC` to fit a support vector machines model of `cultivar` with a Gaussian radial basis function kernel `kernel='rbf'` on the four features used in parts (b) and (c). Fit the model using k-fold cross validation with $k = 4$ folds exactly as in part (b).

```
    clf_svm = KFold(n_splits=4, shuffle=True, random_state=22)
```

Play with the penalty parameter `C` and the coefficient on the radial basis function `gamma` to try and find the lowest possible MSE from the k-fold cross validation. Report your minimized overall MSE along with the tuning parameter values you used for `C` and `gamma`.

(e) Use `sklearn.neural_network.MLPClassifier` to fit a single hidden layer neural network model of `cultivar`. Fit the model using k-fold cross validation with $k = 4$ folds exactly as in parts (b) and (d).

```
    clf_mlp = KFold(n_splits=4, shuffle=True, random_state=22)
```

Play with the tuning parameters of the hidden layer sizes `hidden_layer_sizes`, activation function `activation`, and the regularization penalty `alpha` to try and find the lowest possible MSE from the k-fold cross validation. Report your minimized overall MSE along with the tuning parameter values you used for `hidden_layer_sizes`, `activation`, and `alpha`.

(f) Which of the above three models do you think is the best predictor of `cultivar`? Why?