Ayman Ahmed and Mohammed Iqbal
Machine Learning Final Project - Write Up

Introduction:

The data set that was used was the World Happiness Report 2021, which contained the countries within the world and those countries happiness scores and ranking based on data from the Gallup World Poll. The dataset also contains six measurements for each country - the GDP per capita, social support provided by the government, life expectancy of individuals, freedom to make decisions, generosity and perceptions of corruption. According to the dataset these measurements have no impact on the happiness score reported. However, we wanted to know if it was possible to predict and classify if a country was above the average happiness score based on these measurements from the dataset. For our dataset we took the average happiness score which was calculated to be 5.53 and considered countries whose happiness score was above 5.53 to be happy and below to be unhappy.

Supervised Analysis:

When conducting our supervised analysis we split the data into training data and testing data which was decided randomly. Three learning models were used to classify the countries - logistic regression, support vector machines (SVM) and neural networks. For logistic regression, ridge regularization, lasso regularization and polynomial feature transformation were used. Three kernel functions were used in the support vector machine, linear function kernel, polynomial function kernel, and radial-basis function kernel. Lastly for neural networks, the sigmoid activation function, the reLU activation function and ridge regularization was used.

Table of Results:

Learning Models	Training Accuracy	Testing Accuracy
Logistic Regression + Ridge Regularization	C = .0001, Accuracy = 71% C = .001, Accuracy = 87.5% C = .01, Accuracy = 87.5% C = .1, Accuracy = 90% C = 1, Accuracy = 95% C = 10, Accuracy = 100%	C = .0001, Accuracy = 68% C = .001, Accuracy = 78.2% C = .01, Accuracy = 78.2% C = .1, Accuracy = 79% C = 1, Accuracy = 88% C = 10, Accuracy = 91%
Logistic Regression + Lasso Regularization	C = .0001, Accuracy = 52.5% C = .001, Accuracy = 52.5% C = .01, Accuracy = 52.5% C = .1, Accuracy = 47.5% C = 1, Accuracy = 76.25% C = 10, Accuracy = 76.25%	C = .0001, Accuracy = 46.3% C = .001, Accuracy = 46.3% C = .01, Accuracy = 46.3% C = .1, Accuracy = 53.6% C = 1, Accuracy = 78.2% C = 10, Accuracy = 78.2%
Logistic Regression + Polynomial Transformation	C = .0001, Accuracy = 93.8% C = .001, Accuracy = 100% C = .01, Accuracy = 100% C = .1, Accuracy = 100% C = 1, Accuracy = 100% C = 10, Accuracy = 100%	C = .0001, Accuracy = 88% C = .001, Accuracy = 97% C = .01, Accuracy = 95.6% C = .1, Accuracy = 92.7% C = 1, Accuracy = 92.7% C = 10, Accuracy = 92.7%
SVM + Linear Kernel	C = .0001, Accuracy = 52.5% C = .001, Accuracy = 85% C = .01, Accuracy = 87.5% C = .1, Accuracy = 93.7% C = 1, Accuracy = 97.5% C = 10, Accuracy = 98.7%	C = .0001, Accuracy = 45.3% C = .001, Accuracy = 82.6% C = .01, Accuracy = 79.7% C = .1, Accuracy = 86.9% C = 1, Accuracy = 92.7% C = 10, Accuracy = 94.2%
SVM + Polynomial Kernel	C = .0001, Accuracy = 52.5% C = .001, Accuracy = 52.5% C = .01, Accuracy = 52.5% C = .1, Accuracy = 85% C = 1, Accuracy = 86% C = 10, Accuracy = 87.5%	C = .0001, Accuracy = 46.3% C = .001, Accuracy = 46.3% C = .01, Accuracy = 46.3% C = .1, Accuracy = 78.2% C = 1, Accuracy = 76.8% C = 10, Accuracy = 78.2%

SVM + Radial-basis Kernel	C = .0001, Accuracy = 52.5% C = .001, Accuracy = 52.5% C = .01, Accuracy = 52.5% C = .1, Accuracy = 52.5% C = 1, Accuracy = 81.25% C = 10, Accuracy = 85%	C = .0001, Accuracy = 46.3% C = .001, Accuracy = 46.3% C = .01, Accuracy = 46.3% C = .1, Accuracy = 46.3% C = 1, Accuracy = 81.1% C = 10, Accuracy = 81.1%
Neural Network + Sigmoid Function	51.6%	84.26%
Neural Network + Ridge Regularization	76.3%	58.42%
Neural Network + reLU Function	36.666%	53.93%

Why:

Based on our findings we were able to find that it was possible for some learning models to determine if a country was happy or not based on the countries GDP per capita, social support, life expectancy, freedom, generosity, and perceptions of corruption. Due to our test data and training data being selected at random we decided to run our code multiple times to determine if we would retrieve similar results with different training and testing data. From our conclusion, we have determined that when using Logistic Regression and Support Vector Machine there is minimal variance with testing data and training data as the training accuracy and test accuracy is relatively close for each iteration (∓5%). However when using Neural Networks we have determined that there is a higher amount of variance as the difference between each iteration was from ∓10%.

From our experiment we are able to notice that from our Logistic Regression with larger C values the training and testing data accuracy was very accurate (with some exceptions) and were close to each other, representing that there was minimal bias when running Logistic Regression. Furthermore, in our best test with Logistic Regression and Polynomial Transformation with C=.001 our training accuracy was 100% and our testing accuracy was 97%. Overall Logistic Regression helped classify if countries were considered happy or unhappy however there was a small amount of overfitting as the training accuracy was higher than our testing accuracy but overall the variance was low.

Under Support Vector Machines the training and testing data were around each other's percentages but there was low amounts of bias as our training accuracy was higher than our testing accuracy but were relatively close. Furthermore, under low amounts of C our training and testing accuracy was low. However SVM was able to classify our countries into happy and

unhappy at larger values of C with its highest being with a linear kernel at C = 10 and training accuracy was 98.7% and testing accuracy was 94.2%.

Lastly, under Neural Networks with the exception of Neural Networks with Ridge Regularization, there was lots of underfitting and bias with the training data. With low amounts of training and testing accuracy for both. However under Neural Networks with Ridge Regularization was able to find a number of hidden layers with alpha/lambda values to allow the training accuracy to jump a significant amount, to peak values ranging anywhere from 60% to 90% at its best.