Muhammad Shahbaz

Project Report for Machine Learning Course

Instructor: gitimoni Saikia

Contents

[Purpose 2](#_Toc521374598)

[Classification: 1](#_Toc521374599)

[Datasets: 1](#_Toc521374600)

[Algorithm Evaluation: 2](#_Toc521374601)

[Regression: 3](#_Toc521374602)

[Datasets: 3](#_Toc521374603)

[Algorithm Evaluation: 4](#_Toc521374604)

[References: 5](#_Toc521374605)

# Purpose

The Purpose of this report is to compare different classification and regression algorithms as part of the term completion project for the Machine Learning course for Advance Diploma in Data Science offered by Metro College Technology, Toronto, Ontario, Canada.

As per the requirement for this project report, I have selected 6 datasets from multiple sources for classification and regression algorithms evaluation (3 each). The metadata for each dataset is given accordingly.

# Classification:

## Datasets:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Description | Attributes | Target | No. of Rows |
| BankNote | Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images. | 1.variance of Wavelet Transformed image 2. skewness of Wavelet Transformed image 3. curtosis of Wavelet Transformed image 4. entropy of image | Class (0,1) | 1372 |
|  |  |  |  |  |
| Occupancy | Ground-truth occupancy was obtained from time stamped pictures that were taken every minute. | 1.date time year-month-day hour:minute:second  2.Temperature, in Celsius  3.Relative Humidity, %  Light, in Lux  CO2, in ppm  4.Humidity Ratio, Derived quantity from temperature and relative humidity, in kgwater-vapor/kg-air | Occupancy(0,1) | 9752 |
|  |  |  |  |  |
| Shuttle | Shuttle dataset contains data from US space shuttle program. The shuttle dataset contains 9 attributes all of which are numerical. The first one being time. The last column is the class | 9 integer | 1 Rad Flow  2 Fpv Close  3 Fpv Open  4 High  5 Bypass  6 Bpv Close  7 Bpv Open | 14500 |

## Algorithm Evaluation:

Logistic Regression, KNN, SVM and Random Forest Classifier were used for classification. Following are the evaluation results. The value K for K-Fold cross validation was taken as 10.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Algorithm | Algorithm Parameters | Accuracy | Confusion Matrix (Test) | K-fold (Mean) |
| BankNote | Logistic Regression | default | 0.979 | [[188 7]  [ 0 148]] | 0.981 |
| SVM | kernel = 'linear', random\_state = 0 | 0.982 | [[189 6]  [ 0 148]] | 0.987 |
| RandomForest | n\_estimators=20,  criterion='entropy',  random\_state=0,  max\_depth=4 | 0.973 | [[188 7]  [ 2 146]] | 0.977 |
| KNN | K=5,  metric = 'minkowski',  p=2 | 1.0 | [[195 0]  [ 0 148]] | 0.998 |
| Findings | KNN reflected best accuracy in terms of both simple and k-fold validation. | | | | |
|  |  |  |  |  |  |
| Occupancy | Logistic Regression | default | 0.993 | [[1903 14]  [ 3 518]] | 0.991 |
| SVM | kernel = 'linear', random\_state = 0 | 0.990 | [[1891 19]  [ 3 525]] | 0.991 |
| RandomForest | n\_estimators=20,  criterion='entropy',  random\_state=0,  max\_depth=3 | 0.991 | [[1894 16]  [ 5 523]] | 0.992 |
| KNN | n\_neighbors=5, metric = 'minkowski', p=2 | 0.990 | [[1896 14]  [ 9 519]] | 0.979 |
| Findings | RandomForest classifier came up with the best accuracy as the data seems to be non-linear. | | | | |
|  |  |  |  |  |  |
| Shuttle | Logistic Regression | default | 0.931 | [[2857 0 0 30 0 0 0]  [ 1 0 0 1 0 0 0]  [ 2 0 0 3 2 0 0]  [ 205 0 0 313 0 0 0]  [ 0 0 1 0 208 0 0]  [ 0 0 0 1 0 0 0]  [ 1 0 0 0 0 0 0]] | 0.928 |
| SVM | kernel = 'linear', random\_state = 0 | 0.984 | [[2843 2 0 42 0 0 0]  [ 1 1 0 0 0 0 0]  [ 3 0 4 0 0 0 0]  [ 8 0 0 510 0 0 0]  [ 0 0 1 0 208 0 0]  [ 0 0 0 0 0 1 0]  [ 1 0 0 0 0 0 0]] | 0.980 |
| RandomForest | n\_estimators=20,  criterion='entropy',  random\_state=0,  max\_depth=4 | 0.997 | [[2887 0 0 0 0 0 0]  [ 0 0 0 2 0 0 0]  [ 2 0 2 3 0 0 0]  [ 0 0 0 518 0 0 0]  [ 0 0 0 0 209 0 0]  [ 0 0 0 1 0 0 0]  [ 1 0 0 0 0 0 0]] | 0.996 |
| KNN | n\_neighbors=5, metric = 'minkowski', p=2 | 0.998 | [[2885 1 1 0 0 0 0]  [ 0 1 0 1 0 0 0]  [ 1 0 6 0 0 0 0]  [ 0 0 0 518 0 0 0]  [ 1 0 0 0 208 0 0]  [ 0 0 0 0 0 1 0]  [ 1 0 0 0 0 0 0]] | 0.997 |
| Findings | Another example to non-linear data. KNN performed better than the others in terms of accuracy and k-fold validation. | | | | |

# Regression:

## Datasets:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Description | Attributes | Target | No. of Rows |
| Concrete | Concrete is the most important material in civil engineering. The  concrete compressive strength is a highly nonlinear function of age and  ingredients. These ingredients include cement, blast furnace slag, fly ash,  water, superplasticizer, coarse aggregate, and fine aggregate. | 1.Cement -- kg in a m3 mixture  2.Blast Furnace Slag -- kg in a m3 mixture  3.Fly Ash (component 3) -- kg in a m3 mixture  4.Water -- kg in a m3 mixture  5.Superplasticizer -- kg in a m3 mixture  6.Coarse Aggregate -- kg in a m3 mixture  7.Fine Aggregate -- kg in a m3 mixture  8.Age -- Day (1~365) | Concrete compressive strength -- MPa | 1030 |
|  |  |  |  |  |
| CCPP(Combined Cycle Power Plant) | The dataset contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the power plant was set to work with full load. A combined cycle power plant (CCPP) is composed of gas turbines (GT), steam turbines (ST) and heat recovery steam generators. In a CCPP, the electricity is generated by gas and steam turbines, which are combined in one cycle, and is transferred from one turbine to another. While the Vacuum is collected from and has effect on the Steam Turbine, the other three of the ambient variables effect the GT performance. | 1.Temperature (T) 2.Ambient Pressure (AP) 3.Relative Humidity (RH) 4.Exhaust Vacuum (V) | Net hourly electrical energy output (EP) of the plant. | 9568 |
|  |  |  |  |  |
| CarsMPG | CarsMPG to test the ability to predict car fuel economy  (in mpg) given various specifications: {cylinders, displacement, horsepower,  weight, acceleration, year, origin}.  The dataset is for cars  built in 1970-1982 | 1.Cylinders  2.Displacement  3.Power  4.Weight  5.Displacement  6.Acceleration  7.Year  8.Origin | Mpg | 393 |

## Algorithm Evaluation:

Linear regression, polynomial regression, SVR and Random Forest Regressor were used for Regression. Following are the evaluation results. The value K for K-Fold cross validation was taken as 10.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Algorithm | Algorithm Parameters | Accuracy (RSquare) | Root Mean Squared Error (RMSE) | K-fold (Mean) |
| Concrete | Linear Regression | default | 0.618 | 9.948 | 0.278 |
| SVM(R) | kernel = 'poly', degree =1 | 0.615 | 9.990 | -0.0184 |
| RandomForest | Default | 0.884 | 5.430 | 0.685 |
| Polynomial | Degree=3 | 0.806 | 7.09 | -5.450 |
| Findings | RandomForest gave stable scores for both r-square and k-fold. Others are overfitting due to inconsistency in r-square and k-fold. | | | | |
|  |  |  |  |  |  |
| CCPP(Combined Cycle Power Plant) | Linear Regression | Default | 0.931 | 4.485 | 0.928 |
| SVM(R) | kernel = 'poly', degree =1 | 0.930 | 4.50 | 0.927 |
| RandomForest | Default | 0.960 | 3.403 | 0.959 |
| Polynomial | Degree=5 | 0.945 | 3.986 | 0.940 |
| Findings | Almost all models scored above 90 with both r-square and k-fold validations. RandomForest seems to be a better choice with better r-square and k-fold and least root mean squared error. | | | | |
|  |  |  |  |  |  |
| CarMPG | Linear Regression | Default | 0.825 | 3.303 | 0.642 |
| SVM(R) | kernel = 'poly', degree =1 | 0.816 | 3.392 | 0.629 |
| RandomForest | Default | 0.832 | 3.156 | 0.742 |
| Polynomial | Degree=3 | 0.644 | 4.721 | -0.946 |
| Findings | RandomForest regressor came up as a better option with better stability in r-square and k-fold. It also has the leas mean squared error. | | | | |

# References:

|  |  |
| --- | --- |
| BankNote | <https://archive.ics.uci.edu/ml/datasets/banknote+authentication> |
| Occupancy | <https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection> |
| Shuttle | <https://archive.ics.uci.edu/ml/datasets/Statlog+(Shuttle)> |
| CCPP | <https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant> |
| Concrete | <http://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength> |
| CarsMPG | <https://archive.ics.uci.edu/ml/datasets/auto+mpg> |