Machine Learning – Exoplanet Model Comparison

By

Amusa S. Adebayo

**Introduction**

Artificial Intelligence (AI) has been defined as a broad scientific discipline with roots in philosophy, mathematics and computer science focused on understand and developing of systems that demonstrate properties of intelligence (Panch, Szolovits & Atun, 2018). Machine Learning is a subset of AI, in which computer programs (algorithms) learn associations of predictive power from input/training data. Machine learning uses a broader set of statistical techniques than those typically used in medicine. Advances in AI has led to development of newer techniques like Deep Learning (based on models with less assumptions about the underlying data and capability for handling more complex data). Deep Learning enables input of large volumes of raw data into a machine and the discovery of the representations necessary for detection or classification. Supervised Learning involves training of computer programs to learn associations between inputs and outputs in data via analysis of outputs of interest which were predefined a supervisor (typically human). Unsupervised Learning involves computer programs that learn associations in data without external definition of associations of interest while in Reinforcement Learning, computer programs learn actions based on their ability to maximize a defined reward (Panch, Szolovits & Atun, 2018). In this project, we examine programs for training and testing computer programs for predictive capability, the outcomes of which were evaluated based on model parameters.

**Models:** Five different models were trained and tested. Models’ performances were evaluated using statistical parameters as shown in the Table below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Parameters | Evaluation criteria | Accuracy |
| SVM Model with GridSearchCV Tunning | {'C': 1.0,  'break\_ties': False,  'cache\_size': 200,  'class\_weight': None,  'coef0': 0.0,  'decision\_function\_shape': 'ovr',  'degree': 3,  'gamma': 'scale',  'kernel': 'linear',  'max\_iter': -1,  'probability': False,  'random\_state': None,  'shrinking': True,  'tol': 0.001,  'verbose': False} | Training Data Score: 0.8455082967766546  Testing Data Score: 0.8415331807780321  Grid.best\_params:  {'C': 50, 'gamma': 0.0001}  Grid\_best\_score:  0.8823155822702828 | Test Acc: **0.879** |
| SVM Model with GridSearchCV Tuning -- reduced features | {'C': 1.0,  'break\_ties': False,  'cache\_size': 200,  'class\_weight': None,  'coef0': 0.0,  'decision\_function\_shape': 'ovr',  'degree': 3,  'gamma': 'scale',  'kernel': 'linear',  'max\_iter': -1,  'probability': False,  'random\_state': None,  'shrinking': True,  'tol': 0.001,  'verbose': False} | Training Data Score: 0.8111768071714667  Testing Data Score: 0.7991990846681922  Grid\_best\_score:  0.8197565474934325  Grid\_best\_parameter:  {'C': 50, 'gamma': 0.0001}  Grid\_best\_estimator:  SVC(C=50, gamma=0.0001, kernel='linear') | Test Acc: 0.806 |
| Model3: Decision Tree | {'ccp\_alpha': 0.0,  'class\_weight': None,  'criterion': 'gini',  'max\_depth': None,  'max\_features': None,  'max\_leaf\_nodes': None,  'min\_impurity\_decrease': 0.0,  'min\_impurity\_split': None,  'min\_samples\_leaf': 1,  'min\_samples\_split': 2,  'min\_weight\_fraction\_leaf': 0.0,  'random\_state': None,  'splitter': 'best'} | Training Data Score: 1.0  Testing Data Score: 0.8524027459954233 |  |
| Model4: KNN Classifier with GridSearchCV Tuning | KNN\_Parameters:  {'algorithm': 'auto',  'leaf\_size': 30,  'metric': 'minkowski',  'metric\_params': None,  'n\_jobs': None,  'n\_neighbors': 19,  'p': 2,  'weights': 'uniform'} | Grid\_best\_score:  0.8104141348721793  Grid\_best\_parameters:  {'metric': 'manhattan', 'n\_neighbors': 25, 'weights': 'distance'}  Grid\_best\_estimator: KNeighborsClassifier(metric='manhattan', n\_neighbors=25, weights='distance') | k=19 Train Acc: 0.825  k=19 Test Acc: 0.795 |
| Model5: Logistic Regression with GridSearchCV Tuning | Parameters: {'C': 1.0,  'class\_weight': None,  'dual': False,  'fit\_intercept': True,  'intercept\_scaling': 1,  'l1\_ratio': None,  'max\_iter': 100,  'multi\_class': 'auto',  'n\_jobs': None,  'penalty': 'l2',  'random\_state': None,  'solver': 'lbfgs',  'tol': 0.0001,  'verbose': 0,  'warm\_start': False} | Training Data Score: 0.8504672897196262  Testing Data Score: 0.8432494279176201  Grid\_best\_score:  0.865722170878845  Grid\_best\_parameters:  {'C': 10}  Grid\_best\_estimator: LogisticRegression(C=10) | k=29 Train Acc: 1.000  k=29 Test Acc: 0.811 |
| Model6: Random Forest | Parameters:  {'bootstrap': True,  'ccp\_alpha': 0.0,  'class\_weight': None,  'criterion': 'gini',  'max\_depth': None,  'max\_features': 'auto',  'max\_leaf\_nodes': None,  'max\_samples': None,  'min\_impurity\_decrease': 0.0,  'min\_impurity\_split': None,  'min\_samples\_leaf': 1,  'min\_samples\_split': 2,  'min\_weight\_fraction\_leaf': 0.0,  'n\_estimators': 200,  'n\_jobs': None,  'oob\_score': False,  'random\_state': None,  'verbose': 0,  'warm\_start': False} | Training Data Score: 1.0  Testing Data Score: 0.9016018306636155 |  |

**Models Classifications:** Precisions and accuracy of model training and testing are presented below for different models

**Model1:**

**precision recall f1-score support**

CANDIDATE 0.81 0.67 0.73 411

CONFIRMED 0.76 0.85 0.80 484

FALSE POSITIVE 0.98 1.00 0.99 853

**accuracy 0.88 1748**

macro avg 0.85 0.84 0.84 1748

weighted avg 0.88 0.88 0.88 1748

**Model2:**

**precision recall f1-score support**

CANDIDATE 0.61 0.58 0.59 411

CONFIRMED 0.66 0.66 0.66 484

FALSE POSITIVE 0.98 1.00 0.99 853

**accuracy 0.81 1748**

macro avg 0.75 0.75 0.75 1748

weighted avg 0.80 0.81 0.80 1748

**Model3:**

**precision recall f1-score support**

CANDIDATE 0.70 0.73 0.72 411

CONFIRMED 0.75 0.74 0.75 484

FALSE POSITIVE 0.98 0.98 0.98 853

**accuracy 0.85 1748**

macro avg 0.81 0.82 0.81 1748

weighted avg 0.85 0.85 0.85 1748

**Model4:**

precision recall f1-score support

CANDIDATE 0.64 0.54 0.59 411

CONFIRMED 0.66 0.73 0.69 484

FALSE POSITIVE 0.98 1.00 0.99 853

***accuracy 0.81 1748***

macro avg 0.76 0.75 0.75 1748

weighted avg 0.81 0.81 0.81 1748

**Model5:**

Precision recall f1-score support

CANDIDATE 0.70 0.63 0.66 411

CONFIRMED 0.71 0.75 0.73 484

FALSE POSITIVE 0.98 1.00 0.99 853

***accuracy 0.84 1748***

macro avg 0.80 0.79 0.79 1748

weighted avg 0.84 0.84 0.84 1748

**Model 6:**

Precision recall f1-score support

CANDIDATE 0.84 0.75 0.79 411

CONFIRMED 0.84 0.86 0.85 484

FALSE POSITIVE 0.97 1.00 0.98 853

***accuracy 0.90 1748***

macro avg 0.88 0.87 0.87 1748

weighted avg 0.90 0.90 0.90 1748

**Conclusion**

SVM Model with GridSearchCV Tunning has the highest precision and accuracy of 0.81 and 0.88 respectively. Its training score of 0.845 was also very close to its testing score of 0.841 with test accuracy of 0.879. The order of model fitness generally was Model1 > Model3 > Model5 > Model4 > Model2. With training and test scores of 0.845 and 0.841 respectively, the SVM Model with GridSearchCV Tunning has strong capability of predicting the existence of new exoplanet. Further training of the model with more data will make it even more powerful in its predictive capability.

**Reference**

Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. *Journal of Global Health*, *8*(2). https://doi.org/10.7189/jogh.08.020303