Machine Learning Application in Power Generation Prediction

Bishal Thapa
Department of Engineering
Texas State University
San Marcos, TX, U.S.A
b_t220@txstate.edu

***Abstract*—Non-renewable sources of energy are depleting and the only reliable long-term solution to this vital problem is the use of renewable sources of energy. Solar energy is one of the major alternatives to traditional energy sources like fossil fuels, coal, gas, and oil. Solar energy is a clean and free source of energy. Solar power generation is important for any power plant to determine the specific amount of power that it needs to fulfill the demand. In this way, they do not have to rely on generating power from non-renewable sources. This paper deals with the implementation of machine learning algorithms to predict the power generation of a certain power plant. Supervised machine learning algorithms like Support Vector Machine (SVM), Decision Trees, Random Forest, and Logistic Regression (LR) techniques have been applied to predict power generation. The amount of power generated from the plant depends on various environmental factors and this paper takes into account such factors to predict the power generation.**

Keywords—Supervised learning, power generation, support vector machine, decision trees, random Forest, logistic regression.

I. INTRODUCTION

Solar power is the energy from the sunlight converted into electrical or thermal energy[1]. Solar energy is harnessing the heat from the sun by using a variety of technologies, the most common being the Photovoltaic (PV) cells. The technologies used in harnessing such power can be either active solar or passive solar, based on how the solar energy has been harnessed. Solar energy is an important renewable source of energy that can decrease the use of a traditional non-renewable source of energy like fossil fuels, coal, natural gas, nuclear energy, and oil. There is a tremendous amount of solar energy available and if we could harness that energy properly, it would solve the problem of depleting the non-renewable energy sources.

The use of renewable sources of energy is increasing rapidly as it helps to mitigate the use of non-renewable sources and hence decreases the greenhouse gases emission in the atmosphere which we would get by burning fossil fuels and coal. Furthermore, the non-renewable sources are limited in amount and once depleted cannot be restored. It was actually predicted by BP that within 53 years, all of the oil reserves will be consumed at the current rate[2], [3]. The demand for global energy in 2018 was supplied with natural gases, oil, and coal, which resulted in a 1.7% rise in carbon emissions [4]. So, it's wise to save such limited resources and focus on renewable sources of energy. Due to this reason, lately, there has been a considerable amount of focus on renewable sources and has increased a trend to exploit renewable sources of energy.

The output from a solar panel depends highly on the various environmental factors like sky cover, atmospheric temperature, pressure, humidity, wind speed, etcetera. Hence, the output of the photovoltaic (PV) systems for power generation mainly depends on the weather condition. Bad weather and environmental condition could decrease the power output of the plant. This hampers the ability to fulfill the needs of the consumers. The irregularity of the power generation on the plant could negatively affect the grid system and could decrease the stability and reliability of the operation[5]. This could also affect the grid system from an economic aspect. So, power generation prediction is one of the most important parameters for the solar plant and it helps to know how much power is going to be generated on a particular day in advance. Power generation prediction leads to a reliable power supply and it leads to various social and economic benefits. It is essential to know how much power is generated when the environmental conditions are at different conditions. Hence weather condition is the main factor for the supply-and-demand balance in the PV system operation. The Photovoltaics could be arranged and set up correctly depending upon the information of the power prediction to increase their power generation and realize their full potential.

Machine learning technology has provided a means and opportunity to power generation prediction in a precise and accurate manner. Many machine learning techniques can be used to forecast power generation. Supervised machine learning and artificial neural networks are two of the most used techniques in solar power forecasting. In this research paper, the power prediction is done by using supervised machine learning. Different types of supervised machine learning are applied to classify the power generation into correct ranges of power.

The rest of this research paper is organized as follows; the second section of the research paper discusses the background of the work done on power prediction using machine learning. It includes the findings of the various research done in this sector by many scholars. The third part of the paper discusses in detail about the dataset used in this experiment. It contains a description of the features of the dataset and the data preprocessing done to make it ready to be used in the machine learning models. The fourth section discusses the procedures and the machine learning techniques used for predicting power generation. The fourth section contains all of the results of this experiment in detail. It shows how the different machine learning algorithms performed on the dataset based on the testing accuracy and classification plots. The fifth part of the section deals with the conclusion followed by the final section which is the future work discussion.

II. BACKGROUND

In the last few years, numerous scholars have worked on the prediction of power generation using Machine learning. Usually, power generation forecasting has been done using

either of two ways: direct forecasting and indirect forecasting[5]. In the case of indirect forecasting, numerical weather prediction, artificial neural networks, and image-based statistical models have been used. In the case of direct forecasting, the power generation prediction is usually done by using the data from the historical samples and using that to predict the power generation[5]. Mitsuru, Akira, Yousuke, and Hisahito found out that the forecasting of the power generation of a PV system is better using the direct method than the indirect method, by implementing both the indirect as well as direct methods to predict the power generation of the next day for a PV system[6]. They demonstrated the forecasting technique for PV systems by taking into account the weather conditionskudo[6].

Ye Ren, P.N. Suganthan, and N.Srikanth researched solar irradiance forecasting using ensemble methods[7]. There has been important research on estimating the PV-array hourly power under cloudy weather conditions by Foad H. Gandoman, Fatira Raeisi, and Abdollah Ahmadi[8]. The authors have presented a review on short tern solar PV power forecasting in modern electrical networks[8].

Wan and his peers also presented a paper on the prediction of photovoltaic and solar power for smart grid energy management[9]. The authors analyzed the various techniques for PV solar prediction and discussed the applications of solar power prediction in smart grid energy management[9].

III. DATASET

The dataset used in this experiment was collected and compiled by Alexandra Constantin. The dataset used in this experiment is described below:

A. *Data Acquisition*

This dataset was collected and compiled by Alexandra Constantin. The dataset contains the information of the solar Photovoltaic(PV) systems at the University of California, Berkeley.

B. *Data Preprocessing*

The dataset contains information on various environmental factors that could affect the power generation of the power plant. Those features are included as columns in the dataset. The total number of columns in the dataset is 16. It includes ‘Day of Year’, ‘Year’, ‘Month’, ‘Day’, ‘First Hour of Period’, ‘Is Daylight’, ‘Distance to Solar Noon’, ‘Average Temperature (Day)’, ‘Average Wind Direction (Day)’, ‘Average Wind Speed (Day)’, ‘Sky Cover’, ‘Visibility’, ‘Relative Humidity, ‘Average Wind Speed (Period)’, ‘Average Barometric Pressure (Period)’ and ‘Power Generated’. There are 2,921 rows in the dataset filled with integers for 10 columns, Boolean for one column, and float values for 5 columns, with one empty value for the Average Wind Speed (Period) column in the 714th row. This empty column was dropped using python’s pandas library.

The output label ‘Power Generated’ has 1529 unique values. Since most of the values of the ‘Power Generated’ column are unique to itself, we divided them among certain ranges as the minimum power generation and maximum power generation in KW. So, the labels were divided to get five unique values which we labeled as 0 to 4. We took this column as the output label or the target data series for the experiment. The first five rows of the dataset are shown in Table 1.

Table 1: Dataset sample

Day of Year	Year	Month	Day	First Hour of Period	Is Daylight	Distance to Solar Noon	Average Temperature (Day)	Average Wind Direction (Day)	Average Wind Speed (Day)	Sky Cover	Visibility	Relative Humidity	Average Wind Speed (Period)	Average Barometric Pressure (Period)	Power Generated
245	2008	9	1	1	FALSE	0.859897	69	28	7.5	0	10	75	8	29.82	0
245	2008	9	1	4	FALSE	0.628535	69	28	7.5	0	10	77	5	29.85	0
245	2008	9	1	7	TRUE	0.397172	69	28	7.5	0	10	70	0	29.89	5418
245	2008	9	1	10	TRUE	0.16581	69	28	7.5	0	10	33	0	29.91	25477
245	2008	9	1	13	TRUE	0.065553	69	28	7.5	0	10	21	3	29.89	30069

There was a data imbalance in the dataset. Data imbalance occurs when the number of instances for some classes in the target column is not the same as other classes. The class imbalance was dealt with by using the resample method of the Sci-kit learn library in python[10]. All of the classes were balanced to the same value of 738 by using this resample method.

Label encoder of the Sci-kit learns library was used to change the string data into numerical data so that it can be processed by the machine learning model. Label Encoder encodes the target values with values between 0 and the number of classes[11]. The data were divided into a training set and testing set in the ratio of 4:1.

Linear Discriminant Analysis (LDA) was used to limit the features of the dataset into two. LDA is one of the dimensionality reduction techniques used to maximize the separation between classes[12]. Another solution to the numerous features is the Principal component analysis (PCA). We tried the PCA in the experiment, but found out that the LDA performed better than PCA in this case. The LDA gave better accuracy results for all of the four classifiers used in the experiment.

Some of the features of the dataset (timestamps) were not used for the experiment as they were deemed not important to the output which was the power generation. Hence, the modified final dataset contained 10 features and 1 output feature. The first rows of the input feature dataset are shown in Table 2.

Table 2: Input features to the model

Is Daylight	Distance to Solar Noon	Average Temperature (Day)	Average Wind Direction (Day)	Average Wind Speed (Day)	Sky Cover	Visibility	Relative Humidity	Average Wind Speed (Period)	Average Barometric Pressure (Period)
FALSE	0.859897	69	28	7.5	0	10	75	8	29.82
FALSE	0.628535	69	28	7.5	0	10	77	5	29.85
TRUE	0.397172	69	28	7.5	0	10	70	0	29.89
TRUE	0.16581	69	28	7.5	0	10	33	0	29.91
TRUE	0.065553	69	28	7.5	0	10	21	3	29.89

IV. PROCEDURES

The general procedure flowchart of the experiment is show in figure 1. The various procedure followed during the experiment is briefly described in the following section.

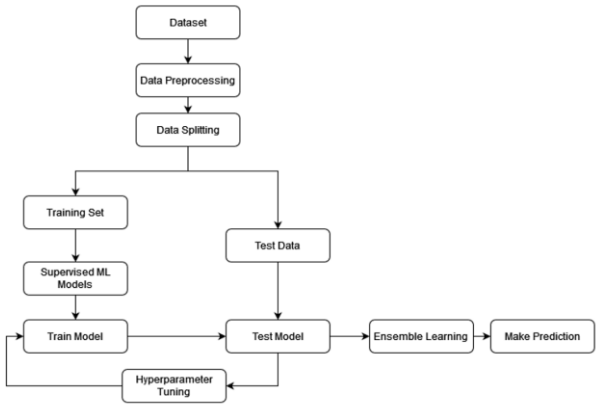

Figure 1: Procedure Flowchart

A. Classification Techniques

We used various supervised machine learning techniques to predict power generation. They are briefly discussed below.

1) Support Vector Machines (SVM): Support Vector Machine is a type of supervised machine learning algorithm that is used to analyze the data for regression analysis, classification analysis, as well as outliers detection*[13]*. SVMs are highly effective in high-dimensional spaces*[13]*. SVM maximizes the margin between the classes, which is the distance between the training samples and the decision boundary. We used python's Sci-kit learn library to import SVM to classify the dataset. SVM depends on various parameters like the 'C' parameter and the 'kernel'. There are many kernels that we can specify. The 'C' parameter is responsible for the penalty for misclassification. Large values of 'C' indicated a large penalty and smaller values indicate a smaller penalty for misclassification.

2) Logistic Regression (LR): Logistic Regression is a type of classification model which is used to separate the linearly separable classes in the dataset*[12]*. It is widely used to describe the relationship between a dependent variable and one or more independent variables.

3) Decision Trees (DT)): Decision trees are a type of supervised machine learning that is used for both classifications as well as regression*[14]*. It works by breaking down the problem by making decisions based on questions. It makes the decision based on the Information Gain (IG). The various parameters of the decision tree are criterion which can be gini or entropy, type of splitter, maximum depth, and minimum samples that is necessary to split an internal node*[14]*.

4) Random Forest: Random forest is an ensemble method for the decision trees as it fits multiple decision trees to improve the accuracy as well as control overfitting*[15]*. There are various parameters upon which the random forest is based upon, and some of them are; the number of trees to be used, criterion which could be gini or entropy, maximum depth of the tree, and the minimum number of samples that is necessary to split an internal node*[15]*.

B. Grid Search

Grid search is responsible for tuning the hyper-parameters of the machine learning model. Hyperparameters are those which control the learning process. Grid search is used to evaluate a machine learning model for various combinations of the hyperparameters that are specified in the grid search. It helps to find the optimal values of the hyperparameter for the given model.

The 'GridSearchCV' object from the sklearn.model_selection module of the Sci-kit learn library was used in the experiment to implement the grid search for the experiment and determine the best hyperparameter for each of the machine learning models. Grid search was implemented on the LR, Random Forest, Decision Trees, and SVM to optimize the hyperparameters of each model. Table It shows the value of the parameters optimized using the grid search.

C. Cross-Validation

Cross-validation is a type of resampling method. It resamples the different portions of the data for training and the testing of the machine learning model on different iterations. The K-fold cross-validation technique is used for the experiment. For the training of the model, k-1 number of folds are used and for evaluation of the performance, a single fold is used[12]. The process is performed for k number of times [12].

V. RESULTS

Grid search was used to optimize the hyperparameters of the model. The best training and testing accuracy from the grid search was recorded for each of the models. Table II shows the value of the best training and the testing accuracy for each model along with the optimized parameters. The accuracies were then compared to determine the best model for the power prediction.

Table 3: Optimized Parameters and Accuracies

Classifiers	**Optimized Parameters**	**Training Accuracy (%)**	**Testing Accuracy (%)**
Logistic Regression	C = 10.0, solver = 'lbfgs'	71.230	71.003
Decision Trees	max_depth =20.0 criterion = entropy	99.221	86.450
Random Forest	Criterion = gini n_estimators = 24.0	99.864	92.412
Support Vector Machines	C = 100.0, kernel = rbf, gamma = 1000.0	99.932	87.669

Precision, recall and F1 score were also recorded for all of the models. These are also performance metrics that helps to analyze how good a machine learning model really is. Precision should ideally be 1 for the perfect classifier. Precision is the performance metric that determines the quality of the positive prediction from the model[16]. It is the ratio of the true positive to the sum of true positive plus false negative. The best precision value was provided by random forest classifier with precision of 0.79788. Recall determines the actual positive values that our model captured[17]. Recall

is the ratio of the true positive to the total actual positive values, which is the sum of true positive and false negative. The best recall value was also provided by the random forest classifier with recall of 0.80006. F1 score is another performance metric of a classification problem. F1 score is a harmonic mean of the recall and the precision. The ideal value of precision is also 1, which occurs when both precision and recall are 1. The best F1 score value was provided by random forest which was 0.79848.

After getting the best parameters, training accuracy, and testing accuracy using the optimized parameters, the classification plots were generated. The classification plots were saved for each of the models. Figure 1 shows the classification plot for the best model which was found to be the decision tree ensemble model as the test accuracy of the decision tree model was found to be the highest. Figure 1 classification plot shows that the random forest with the optimized hyperparameter can classify the five different samples.

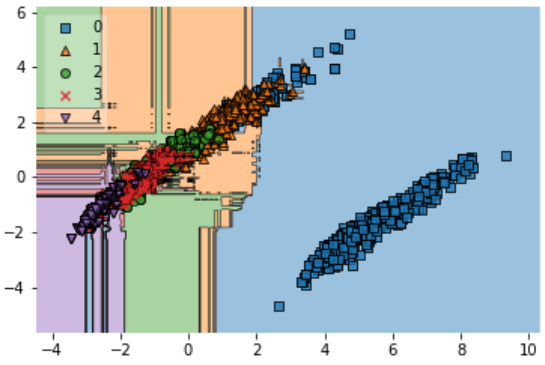

Figure 2: Random Forest Classification Plot showing the five different labels

After getting the classification plot, a confusion matrix was also generated for each of the models. The confusion matrix shows the results of the prediction in a clear fashion. It provides the 'True Positive', 'True Negative', 'False Positive' and 'False Negative' values of the model. A confusion matrix was generated to summarize the true and false predictions made by our machine learning model. Fig 3 shows the confusion matrix generated for the random forest model that provided the best testing accuracy. The number of correct and incorrect classifications can be easily analyzed from the confusion matrix.

Area under the Curve (AUC) Receiver operating characteristic (ROC) is a type of graph that enables us to analyze the performance of the classification by the model [07]. The measure of performance is based on False Positive Rate (FPR) and True Positive Rate (TPR)[book]. There is a diagonal line which is considered as the curve for random guessing. ROC is a type of probability curve whereas the AUC provides the degree of separability [07].

Figure 4 shows the ROC AUC curve for the model with the best test accuracy. It shows how the model is classifying the various labels of the dataset.

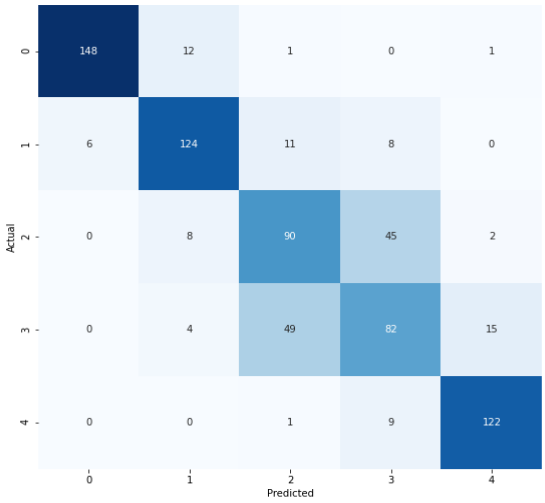

Figure 3: Confusion matrix using Random Forest classifier

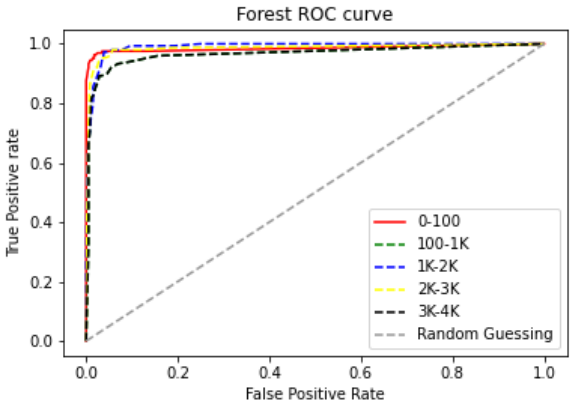

Figure 4: ROC AUC Curve for Random Forest

Furthermore, the accuracies of all the machine learning models were compared and analyzed. The table 5 shows the results of multiple iterations. It shows the average value of precision, recall and f1 score of the classifiers for multiple iterations. It also shows the best accuracy and average accuracy of the classifiers when experiment is done for multiple iterations. The last column in the table shows the accuracy value as the average accuracy value plus the standard deviation and average accuracy minus the standard deviation. It shows the dispersion measure of the test accuracy.

Table 4: Average of multiple iterations of the models

				Test Accuracy	
Classifier	Precision	Recall	F1-Score	Average Test Accuracy	Average ± standard deviation
LR	0.722973	0.725315	0.720622	72.6964	71.615 - 73.777
SVM	0.72127	0.723326	0.718006	87.6964	86.463 - 88.929
Decision Trees	0.746942	0.747007	0.746495	89.756	88.692 - 90.819
Random Forest	0.781103	0.781163	0.78146	89.759	89.688 - 90.829

The three of the machine learning models; logistic regression, decision trees, and random forest were used together along with a majority voting classifier. The majority voting classifier uses multiple classifiers to classify the data.

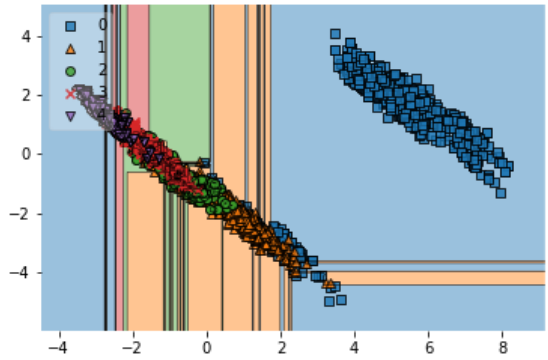

Figure 5: Majority Voting Classification

The results of the majority voting classification were:

ROC AUC: 0.96 (+/- 0.01) [Logistic regression]
ROC AUC: 0.96 (+/- 0.01) [SVM]
ROC AUC: 0.93 (+/- 0.01) [Decision tree]
ROC AUC: 0.96 (+/- 0.01) [Majority voting]

From the results above we can see that the majority voting classifier has the highest ROC AUC value. It increase the model's performance and achieved better performance than single model.

VI. CONCLUSIONS

In this paper, we explored the application of machine learning techniques to forecast the power generation of a solar system. The first part of the project involved the collection and pre-processing of the data. The number of features to be used was decided and the data was prepared to be used for the various machine learning models.

The second part of the project consisted of an experiment with various models and their parameters. The preprocessed data were used with various supervised machine learning models. The best parameter from all of the models was recorded based on the best accuracy of the model.

The results of the experiment showed that the random forest classifier provides the best testing accuracy for the model. The random forest, being an ensemble of trees provided the best accuracy. We observed that the testing accuracy of SVM and decision trees were nearly identical, with SVM just edging out with slightly more testing accuracy. The results of the accuracy and plots from the various models also showed that solar power generation can be predicted with reliable accuracy using the supervised machine learning models. Hence, it is fair to conclude that the machine learning algorithms can be used to reliably predict how much power can be generated from solar plants by taking into account the various environmental factors.

VII. FUTURE WORK

The field of machine learning is always expanding and improving. The machine learning model used in this experiment can also be adjusted differently to get better results. We can improve the various machine learning models used in this experiment and increase the accuracy of the prediction. The precision of the model could also be increased for future work. Furthermore, the model can be retrained by using more features than what was used in this experiment. The number of features used in this experiment could be increased as there may be other factors affecting the power generation prediction. We can also improve the accuracy of the prediction by collecting more amount of reliable data and using those for our machine learning models. The concepts of Artificial Neural Networks could be used on the dataset for the prediction after collecting large data, which could improve the results. The sector of power generation prediction could benefit highly because of accurate machine learning models in the future.

ACKNOWLEDGMENT

I would like to express my gratitude to my Machine Learning teacher Dr. Valles Molina, Damian, who provided me the opportunity to take on this project. He pushed me to my limit and made me realize I could do better. I learned a lot of essential skills during his class. I truly am thankful to Dr. Valles.

REFERENCES

[1] "About Solar Energy," *SEIA*. https://www.seia.org/initiatives/about-solar-energy

[2] "When Fossil Fuels Run Out, What Then?," *MAHB*, May 23, 2019. https://mahb.stanford.edu/library-item/fossil-fuels-run/

[3] "How long will world's oil reserves last? 53 years, says BP," *Christian Science Monitor*, Jul. 14, 2014. [Online]. Available: https://www.csmonitor.com/Environment/Energy-Voices/2014/0714/How-long-will-world-s-oil-reserves-last-53-years-says-BP

[4] J. Howarth, "When will fossil fuels run out?," *Octopus Energy*. https://octopus.energy/blog/when-will-fossil-fuels-run-out/

[5] U. K. Das *et al.*, "Forecasting of photovoltaic power generation and model optimization: A review," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 912–928, Jan. 2018, doi: 10.1016/j.rser.2017.08.017.

[6] M. Kudo, A. Takeuchi, Y. Nozaki, H. Endo, and J. Sumita, "Forecasting electric power generation in a photovoltaic power system for an energy network," *Electr. Eng. Jpn.*, vol. 167, no. 4, pp. 16–23, 2009, doi: 10.1002/eej.20755.

[7] Y. Ren, P. N. Suganthan, and N. Srikanth, "Ensemble methods for wind and solar power forecasting—A state-of-the-art review," *Renew. Sustain. Energy Rev.*, vol. 50, pp. 82–91, Oct. 2015, doi: 10.1016/j.rser.2015.04.081.

[8] F. H. Gandoman, F. Raeisi, and A. Ahmadi, "A literature review on estimating of PV-array hourly power under cloudy weather conditions," *Renew. Sustain. Energy Rev.*, vol. 63, pp. 579–592, Sep. 2016, doi: 10.1016/j.rser.2016.05.027.

[9] C. Wan, J. Zhao, Y. Song, Z. Xu, J. Lin, and Z. Hu, "Photovoltaic and solar power forecasting for smart grid energy management," *CSEE J. Power Energy Syst.*, vol. 1, no. 4, pp. 38–46, Dec. 2015, doi: 10.17775/CSEEJPES.2015.00046.

[10] "sklearn.utils.resample," *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.utils.resample.html

[11] "sklearn.preprocessing.LabelEncoder," *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

[12] S. Raschka and V. Mirjalili, *Python Machine Learning*, Third. Packt Publishing Ltd.

[13] "1.4. Support Vector Machines," *scikit-learn*. https://scikit-learn/stable/modules/svm.html (accessed Dec. 05, 2021).

[14] "1.10. Decision Trees," *scikit-learn*. https://scikit-learn/stable/modules/tree.html

[15] “sklearn.ensemble.RandomForestClassifier,” *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[16] “Precision,” *C3 AI*. https://c3.ai/glossary/machine-learning/precision/

[17] K. P. Shung, “Accuracy, Precision, Recall or F1?,” *Medium*, Apr. 10, 2020. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9