Exploring Machine Learning and Decision Trees for Data Classification and Visualization in Big Data

Gulay Tarverdiyeva, Aytaj İsmayilova

Abstract

Everything that surrounds us today automated. Machines have grown to be a need in our daily lives. Every single human activity involves a machine. Without a question, one of the most important and useful technologies now in use is machine learning. By the use of machine learning, a system may infer assumptions from its past output without being explicitly programmed. Machine learning adopted in nearly all industries today. Therefore, it is crucial to understand all machine-learning paces. Based on numerous algorithms, machine learning applied to the appropriate sectors. I analyze machine learning in this work and present the decision tree method as a solution to the issue.

Keywords: Machine learning, big data, decision tree, data classification, data visualization.

Machine learning is the process through which data transformed into knowledge. Over the past fifty years, there has been an explosion in data. This vast volume of data is useless until we examine it and uncover the hidden arrangements. The various arrangements that we would otherwise struggle to uncover inside complex data are automatically found using machine learning techniques. Future events can predicted, and all types of sophisticated decisions can made using the hidden arrangements or patterns and understanding of a situation. Machine Learning is broadly categorized into 4 types. One of them is supervised machine learning that I used its classification category with the help of Decision Tree Algorithm in my statement of the problem.

Using feature values, decision trees categorize data. With training data, decision trees built repetitively using a top-down greedy method that sequentially selects features. [5] The training set of data arranged into a tree-structure plan by decision tree classifiers. A decision tree built by starting with a root node that contains all of the data, selecting splitting criteria recursively, and then expanding leaf nodes with divided data subsets in accordance with the splitting criteria. Based on quality metrics like information gain, splitting criteria selected that handle the complete data set of each expanding node. Application of decision trees to big data applications is therefore challenging.

Building a classification network using pre-existing patterns is an example of learning by doing [2] A new class may declared or added to an existing class because of such an observation. New theories and information that ingrained in the input patterns made easier to understand by this classification. The neural network model's related training improves the classification parameters. In order to classify postgraduate students according to their performance during the admission period, this research took into consideration the two learning algorithms, supervised and unsupervised, and evaluated their properties. Although the error back-propagation supervised learning technique is particularly effective for many non-linear real-time tasks, researchers discovered that the unsupervised model outperforms the supervised learning algorithm in the KSOM student classification challenge.

In this study, a method based on SVM suggested and used to diagnose tuberculosis. [4] The experimental results are promising and demonstrate that SVM techniques deliver a notable performance to reach 94.7% with only a little runtime required for training. The outcomes additionally assessed to those of previous works. This study's categorization accuracy results were more accurate than those of the other investigations were. As a result, it can said that SVM is effective for diagnosing tuberculosis.

In the interest of comprehending present practices in ML4VIS research—that is, using ML approaches to address issues relating to data visualization—we surveyed 85 publications for this article. [10] We used pre-existing visualization models and organized the six visualization processes into an ML4VIS pipeline to map out the function of ML4VIS in general visualizations. To demonstrate how the needs in visualization created as ML tasks, the six visualization processes are also linked with the learning tasks in ML. ML4VIS existing practices and potential areas for future research are highlighted based on our analysis of recent studies. They think that this survey can advance future research in the field of ML4VIS and offer insightful information.

In many spheres of life, including the medical field, commercial domains, etc., the need to extract the important information from raw data has become apparent. [8] As we have already seen, ML and DM have applications in both the medical and commercial fields for making wise business decisions and predicting illnesses. We compared and analyzed ML classifiers for the prediction of heart and hepatitis disorders in this article. Six distinct classifiers used to predict heart and hepatitis problems. Different classifiers perform differently on the same dataset, according to experimental results. Of of six classifiers tested, RF outperformed them all in terms of illness prediction on both datasets, according to the report.

The product databases of Amazon, Flipkart, Snapdeal, and Paytm used for our experiments. [1] Based on our approach, which handles the attribute-wise distribution of phrases to conform to the structured nature of e-lists, approximately 70% accurate results obtained. The best part is that by using normalization, their method may produce superior outcomes without assigning weight to lengthy material.

Machine learning frequently involves doing data analysis tasks using a variety of parameters, however these tasks time-consuming [11]. Reduce overall execution time is the suggested approach for optimizing the job assignment for machine learning. Our approach may be expanded to execute data analytics jobs, execute jobs in memory, integrate jobs, support machine learning, and optimize job assignment depending on execution. To forecast how long these jobs will take to complete on the extended execution, machine-learning approaches have created.

Machine learning is a subject of artificial intelligence that offers machines with the capacity autonomously learn from data and past experiences while finding patterns to generate predictions with minimal human interaction. A model built using machine learning algorithms on a training dataset. [6] The created model is used by the trained ML algorithm to make predictions as new input data supplied.

Big data has used machine learning. [9] Big data is a vast collection of both structured and unstructured information that too big processed by conventional database and software methods. Big data technologies have a significant impact on value generation and scientific breakthroughs. Massive Parallel Processing (MPP), Distributed File Systems, Cloud Computing, etc support big Data. In addition to conventional cloud infrastructure services, Big Data supported by technologies like Hadoop, databases/servers SQL, nosql, and MPP databases, among others.

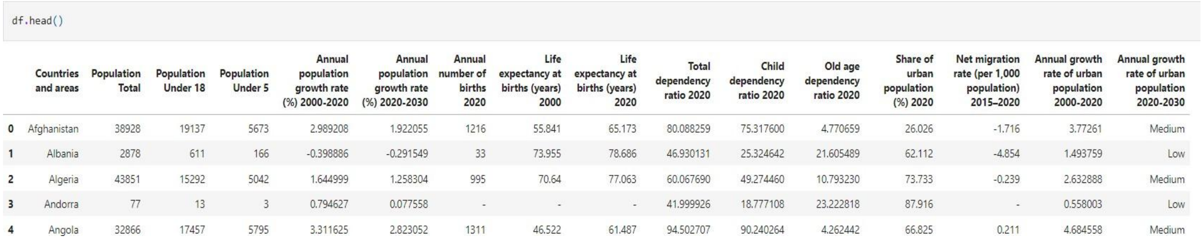
Machine learning has 4 types: Supervised learning, Unsupervised learning, Semi-supervised learning, reinforcement learning. [7] When we talk about supervised learning, we are talking about a particular kind of problem where the input data is a matrix X and we are trying to predict a response y. Where Y = {c1, c2} and X = {x1, x2,..., xn} have n predictors each. Using demographic features as predictors, one possible application would be to estimate the likelihood that a web user will click on advertisements. This frequently used to forecast click-through rates (CTR). Then y = "click, doesn't click," and the predictors may include the user's IP address, the day he first visited the site, his city and nationality, among other potentially relevant information. Finding groups within groups that are similar to one another without access to a class to learn from is a challenge that is addressed by unsupervised learning. Finding groups that share comparable examples in each group while being distinct from one another requires learning a mapping from predictors. Customer segmentation is a use case for unsupervised learning. For instance, segmenting people based on how they use their phones is a regular activity in the telecommunications sector. The marketing division would then be able to target each group with a unique product as a result.

The technique of educating a model by providing it with accurate input and output data is known as supervised learning. "Labeled data" refers to this pair of input and output data. Obtaining the tagged training data is the first step in this process' initiation. The label, which is the algorithm's output, gives feedback. For training, testing, and validation, the tagged data should then be split into three groups. The approach modifies the model and reduces error using the training set. The validation set is separate from the training set and enables one to assess the performance of the learning algorithm on their own. The test set is the final set and should only be used if the model on the validation set has been determined to be the best one. A supervised learning model's objective is to anticipate the proper label for recently supplied input data. Problems with supervised learning can be further broken down into two categories. They are classification and regression. Regression is the term used when the class attribute is continuous as opposed to discrete, which is classification. This article's goal is to demonstrate how to use the classification approach to solve problems. These output variables have a categorical nature and are used to deal with categorization problems like yes or no, true or false, male or female, etc. Two practical applications of this topic are spam detection and email filtering. A decision tree is a type of prediction model that links observations of an item to judgments about its goal values. [5] A decision tree can be used in decision analysis to formally and visually reflect decisions and decision-making. The objective is to learn straightforward decision rules derived from the data features in order to build a model that predicts the value of a target variable. The structure of a decision tree reveals that the branches reflect the conjunction of features that result in classification, non-leaf nodes indicate features, and leaves represent classification.

Data Cleaning, Data Transformation, and Feature Engineering are the three processes that make up the Data Preparation stage, which's goal is to get the data into the ideal structure for machine learning. [3] This is a crucial phase that shouldn't be overlooked because utilizing complex algorithms is less significant than using high-quality data. According to the Data Cleansing, data should be detected any missing or incorrect values. Data exploration can begin after you have access to the necessary data. This step is crucial if you are unfamiliar with the data because it allows you to summarize the data in a comprehensible way. All assumptions should be tested at this point. Dealing with missing numbers and outliers, fixing typos, grouping sparse classes, and removing duplicates are just a few of the tasks involved in this stage. A complete dataset with accurate values for each observation would be excellent. In practice, though, you will encounter a lot of "NULL" or "nan" values. The simplest technique to handle missing data is to eliminate all rows with a missing value, but this can result in the loss of important information or bias. As a result, it's critical to investigate whether the missing numbers have a cause or a pattern. For instance, certain demographic groups may not reply to certain survey questions; by eliminating them, learning tendencies within these groups are prevented. Imputing values—replacing missing values with a suitable replacement is an alternative to erasing data. Usually used for continuous variables are the mean, median, and mode. While it is frequently the mode or a new category for categorical data. If a large percentage of the data in a column are missing, you might want completely eliminate the column. An observation that considerably deviates from the majority of previous observations is referred to as a "outlier." When you locate outliers, you should also look into potential causes. Outliers may be a sign of faulty data, such as information that was improperly gathered. If so, you might want to replace or eliminate these data items. As an alternative, your machine -learning model might find these values interesting and helpful. Outliers may be detected by some machine learning methods, including linear regression. [8] As a result, you might only utilize methods that are resistant to outliers, like gradient boosted trees or random forests.

Statement of the problem and the solution

For this problem I chose a raw data which is about demographics information calculated by UNICEF dataset. These demographics consists of data of 202 countries. To classify data I used python programming language and for visualization both orange visualization tool and python. If python is used to visualize data then we have to write code but orange visualization tool does not need any code. It is easier than python. First of all I have to import numpy and pandas libraries for data analysis, matplotlib and seaborn libraries for visualization. We will now read the data from the dataset that downloaded. We can see here top 5 rows of 202 counties with the help of head () function. (Figure 1)



```
df.head()
```

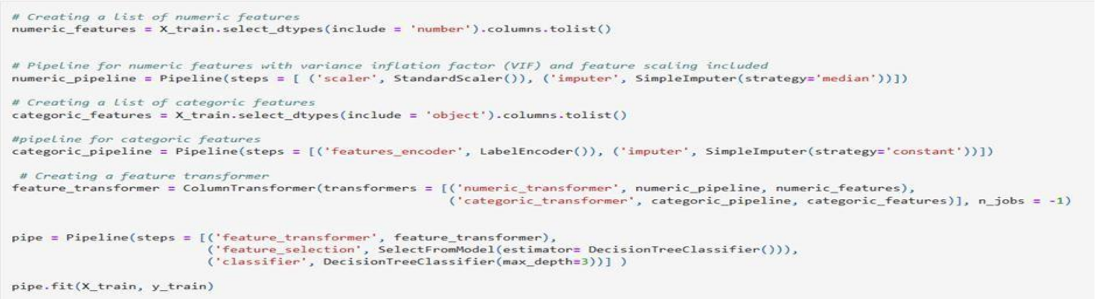
	Countries and areas	Population Total	Population Under 18	Population Under 5	Annual population growth rate (%) 2000-2020	Annual population growth rate (%) 2020-2030	Annual number of births 2020	Life expectancy at births (years) 2000	Life expectancy at births (years) 2020	Total dependency ratio 2020	Child dependency ratio 2020	Old age dependency ratio 2020	Share of urban population (%) 2020	Net migration rate (per 1,000 population) 2015–2020	Annual growth rate of urban population 2000-2020	Annual growth rate of urban population 2020-2030
0	Afghanistan	38928	19137	5673	2.989208	1.922055	1216	55.841	65.173	80.088259	75.317600	4.770659	26.026	-1.716	3.77261	Medium
1	Albania	2878	611	166	-0.398886	-0.291549	33	73.955	78.686	46.930131	25.324642	21.605489	62.112	-4.854	1.493759	Low
2	Algeria	43851	15292	5042	1.644999	1.258304	995	70.64	77.063	60.067690	49.274460	10.793230	73.733	-0.239	2.632888	Medium
3	Andorra	77	13	3	0.794627	0.077558	-	-	-	41.999926	18.777108	23.222818	87.916	-	0.558003	Low
4	Angola	32866	17457	5795	3.311625	2.823052	1311	46.522	61.487	94.502707	90.240264	4.262442	66.825	0.211	4.684558	Medium

Figure 1. Top 5 rows of data frame

We are able to produce descriptive statistics that include Nan values and describe central tendency. The output will include count, mean, standard deviation, maximum, minimum, and 25%, 50%, and 75% values if the data type is numeric. Count, Unique, Top, and Frequency will be displayed if the datatype is an object. In a machine learning model, this code establishes a pipeline for processing and transforming both numerical and categorical information. The pipeline has two primary steps: feature transformation and feature selection, and then decision tree classifier for the final classification. (Figure 2). The first step involves separating the numeric and categorical features using the include parameter in the columntransformer function.

The categorical characteristics then converted using label encoding and imputation, while the numeric features are handled using a pipeline that incorporates scaling and imputation. A columntransformer is used to integrate these pipelines and perform the proper changes to each feature.

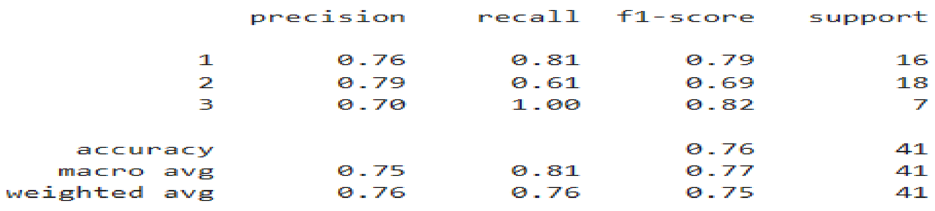
The following stage includes choosing features with a decision tree classifier. Based on the decision tree feature importances of the model, the selectfrommodel method chooses the most crucial features. Lastly, the target variable is predicted using a decision tree classifier with a maximum depth of three.



```
# Creating a list of numeric features
numeric_features = X_train.select_dtypes(include = 'number').columns.tolist()


# Pipeline for numeric features with variance inflation factor (VIF) and feature scaling included
numeric_pipeline = Pipeline(steps = [ ('scaler', StandardScaler()), ('imputer', SimpleImputer(strategy='median'))])

# Creating a list of categoric features
categoric_features = X_train.select_dtypes(include = 'object').columns.tolist()

#pipeline for categoric features
categoric_pipeline = Pipeline(steps = [('features_encoder', LabelEncoder()), ('imputer', SimpleImputer(strategy='constant'))])

 # Creating a feature transformer
feature_transformer = ColumnTransformer(transformers = [('numeric_transformer', numeric_pipeline, numeric_features),
                                                        ('categoric_transformer', categoric_pipeline, categoric_features)], n_jobs = -1)


pipe = Pipeline(steps = [('feature_transformer', feature_transformer),
                         ('feature_selection', SelectFromModel(estimator= DecisionTreeClassifier())),
                         ('classifier', DecisionTreeClassifier(max_depth=3))] )

pipe.fit(X_train, y_train)
```

Figure 2. Pipeline for processing and transforming both numeric and categorical features of dataset

Overall, this pipeline is a helpful tool for managing and choosing pertinent features in a machine learning model quickly and effectively, increasing the model's precision and interpretability. The pipeline can enhance model performance by choosing just the most crucial features, decreasing over fitting, and enhancing the model's interpretability by applying feature selection approaches like Select From Model. Lastly, the decision tree classifier offers a straightforward yet efficient classification technique that is easy enough for stakeholders to understand. Finally, the result of this model is shown in Figure 3.

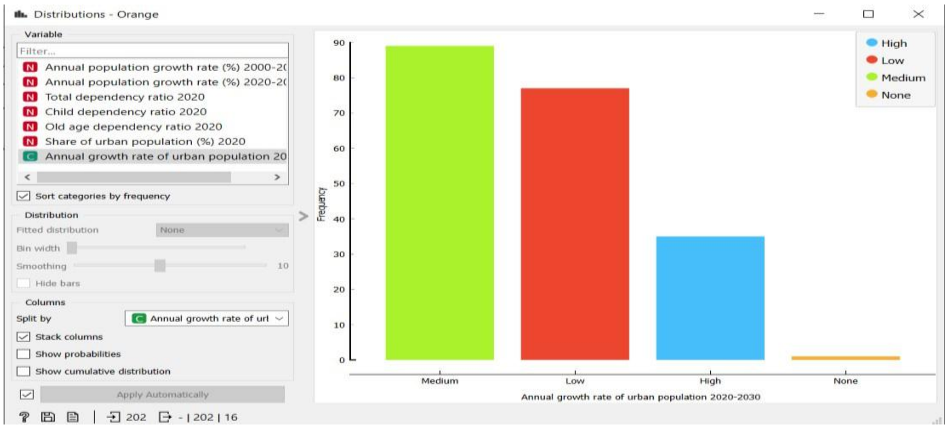


```
              precision    recall  f1-score   support

           1       0.76      0.81      0.79        16
           2       0.79      0.61      0.69        18
           3       0.70      1.00      0.82         7

    accuracy                           0.76        41
   macro avg       0.75      0.81      0.77        41
weighted avg       0.76      0.76      0.75        41
```

Figure 3. Model evaluation for train dataset results

Data visualization allows us to view how the data is organized and what sort of relationships the qualities of the data hold. It is the quickest method for checking if the features and output match. Correlation is the possibility that changing one variable may also affect another. If we want to use python, we have to write this code to implement visualization: sns. Countplot (x='Annual growth rate of urban population 2020-2030', data=df, palette='Dark2')

But it is easier to use orange tool because we do not need any code here. We have just choose any column name as a variable then this tool will automatically implement visualization. (Figure 4)



Distributions - Orange

Variable

Filter...

- N Annual population growth rate (%) 2000-20
- N Annual population growth rate (%) 2020-20
- N Total dependency ratio 2020
- N Child dependency ratio 2020
- N Old age dependency ratio 2020
- N Share of urban population (%) 2020
- C Annual growth rate of urban population 20

☑ Sort categories by frequency

Distribution

Fitted distribution: None

Bin width

Smoothing: 10

□ Hide bars

Columns

Split by: C Annual growth rate of url

☑ Stack columns

□ Show probabilities

□ Show cumulative distribution

☑ Apply Automatically

202 | - | 202 | 16

Figure 4. Data visualization in Orange

Conclusion

The use of machine learning by virtually every industry to enhance their procedures and operations has led to the development of machine learning as an essential component of our daily life. Machine learning algorithms are an effective tool for data classification and prediction because they may draw conclusions from prior data without explicitly programmed. Decision trees, one of the many machine-learning algorithms available, have shown to be successful in resolving classification issues, particularly when the data is well structured and simple to understand. Moreover, data visualization is essential to the success of machine learning initiatives because it helps stakeholders comprehend the data and the model-derived insights. It is critical to keep up with the most recent advances in machine learning as big data continues to grow and to use the right algorithms in the right industries.

References

[1] Abhilash Shrivastava, Shekhar Pandey, Supriya Muthuraman. “Data Classification Using Machine Learning Approach”. *The International Symposium on Intelligent Systems Technologies and Applications*. Pp. 112-121. January 2018.

[2] Annamma Abraham, R. Sathya. “Comparison of Supervised and Unsupervised Learning Algorithms

for Pattern Classification". *International Journal of Advanced Research in Artificial Intelligence*. Pp. 34-37. February 2013.

[3] Antonio J. Tallón-Ballesteros, José C Riquelme. "Data Cleansing Meets Feature Selection: A Supervised Machine Learning Approach". *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Pp. 370-377, June 2015.

[4] Amani Yahyaoui, Imene Yahyaoui. "Machine Learning Techniques for Data Classification". *Advances in Renewable Energies and Power Technologies Volume 2*. Pp 1-8. February 2018.

[5] Aparna Ramdoss. "Analysis of Machine Learning and its Algorithms". *National Conference on Rising Trends in Mathematical Analysis and Computing Technologies*. Pp. 22-26. February 2020.

[6] Batta Mahesh. "Machine Learning Algorithms -A Review". Pp. 381-385. January 2019.

[7] Ch Anwar Ul Hassan, Muhammad Sufyan Khan, Munam Ali Shah. "Comparison of Machine Learning Algorithms in Data classification". *24th International Conference on Automation and Computing (ICAC).* Pp. 1-7. September 2018

[8] Cheryl Ann Alexander, Lidong Wang. "Machine Learning in Big Data". *International Journal of Mathematical, Engineering and Management Sciences*. Pp. 52-56. September 2016.

[9] Huamin Qu, Qianwen Wang, Yong Wang, Zhutian Chen. "Applying Machine Learning Advances to Data Visualization: A Survey on ML4VIS". Pp. 1-18. December 2020.

[10] Lakshmi Jupudi. "Machine learning techniques using python for data analysis in performance evaluation". *International Journal of Intelligent Systems Technologies and Applications*. Pp. 3-18. January 2018.