International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 07 | July - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Combining Results of Machine Learning Algorithms to Predict Accuracy

Manohar E.R, MD Khaja Mulge, DR Vibha MB
PG Scholar, PG Scholar, Assistant Professor
Department of Master of Computer Application
Dayananda Sagar College of Engineering

ABSTRACT

In this modern era given the vast volumes of data produced today, data science is one of the most hotly contested topics in the IT world. Its popularity has risen over time, and businesses have begun to use data science approaches to expand their operations and improve consumer happiness. It has became one of important part of modern industries as we know modern industries has huge amounts of to be produced. During this survey we'll discover stacking of machine learning algorithms using ensemble.

As we saw that until today data has grown exponentially, it was a big problem to store this data Until 2010, there were only a few limited options for solving data processing problems. However, in that year, new tools and technologies were developed that made it easier to work with data. solve this problem **[2]**. IT started thinking on how to use this data for better experience for their clients. This is how data science came into existence. As on today data has become one of the most precious asset. Many fields like cyber-security , data security have come into existence to protect this data.

INTRODUCTION

As we know data science the field that uses methods and programming algorithms to extract information from irregular, structured data as well as unstructured data[1]. It is a concept in which statistics is implemented in the form of programming algorithms and analysis is performed on data. Even though nowadays data science is implemented in computer science it draws many features from mathematics, information science and statistics. A person who implements or makes new algorithms for data science it's known as data scientist. This is a field where we are extremely focused on extracting information from datasets that are typically large which is otherwise called enormous information and utilize this data to handle challenges across an expansive range of utilization spaces. This includes preparing data for analysis, Analyzing the data visualizing the data, making data driven solutions**[1]**.

Machine learning

Machine learning is part of Artificial intelligence which is used to predict future outcomes using data from the past. Machine learning has grew extremely today as it is used everywhere, for example from e commerce sites to automobile manufacturing companies all use machine learning. E commence sites use past data from users like what category of products do a particular customer buy, what type of products they buy, what time of the month do customers make more number of purchases, how frequently does the customer purchase anything from the store etc.

Ensemble learning

Ensemble learning is a machine learning technique that can give better performance in predictions by combining the results of several different models. There are many ways to develop an ensemble model, but there are 3 dominant methods in ensemble learning. Bagging , boosting and Stacking are the most popular and used methods of ensemble learning.

Different types of Ensemble Learning:

1. **Boosting:** Weighted averages are used to make weak learners into stronger learners using algorithms , this process is called Boosting . Boosting is all about the

teamwork. one is learning from other which in turn improves the learning this is called Boosting.

2. **Bootstrap Aggregation (Bagging):** General procedure that can be used reduce the variance for those algorithms that have high variance, typically decision trees. Bagging makes each model run independently .The outputs are aggregated without any preference for any model.

Here we are going to use stacking ensemble method on iris data-set. Stacking involves applying different models on a similar dataset another model is applied on the outcomes of these models to make prediction more accurate.

DATA SET

The Fisher's Iris data set / Iris flower data set is a multivariate data set which is an example of linear discriminant analysis by British statistician and biologist Ronald Fisher in his 1936 paper The use of numerous measurements in taxonomic issues**[1]**. Since Edgar Anderson gathered the information to evaluate the morphologic variety of Iris blossoms of three related specie, it is frequently referred to as Anderson's Iris data set. Two of the three species were taken in the Gaspé Peninsula "all from the same pasture, on the same day, and measured by the same individual using the same apparatus."

There are 150 exceptional examples in the information assortment, with 50 examples from every one of the three Iris species (Iris setosa, Iris virginica, and Iris versicolor). Each example had four attributes estimated: the length and width of the sepals and petals in centimeters**[1]**. Fisher devised a linear discriminant model to distinguish the species based on the combination of these four characteristics.
Fisher's study was published in the Annals of Eugenics, and it contains a discussion of how the approaches might be used in the field of phrenology. As a result of this history, some have suggested that the "iris" dataset be phased out of statistical teaching methodologies in favour of less contentious alternatives.

MODELS

Random forest: Random forest is a Supervised Learning calculation which utilizes group learning technique for characterization and regression, not a boosting technique. The trees in random forests are run in parallel. These trees are built independently without interaction while building the trees**[5]**. The algorithm uses a training set of data to build decision trees, and then uses the classifications or regression predictions of the individual trees to generate a class.

```
print(classification_report(random_pred,y_test))
```

	precision	recall	f1-score	support
setosa	1.00	1.00	1.00	19
versicolor	1.00	0.94	0.97	16
virginica	0.94	1.00	0.97	15
accuracy			0.98	50
macro avg	0.98	0.98	0.98	50
weighted avg	0.98	0.98	0.98	50

Gaussian Naive Bayes: The Naive Bayes algorithm is a probabilistic machine learning technique that can be used to solve a variety of classification problems. Document classification, spam filtering, prediction, and other applications of Naive Bayes are common. Because of its name, this method is based on Thomas Bayes' discoveries.**[3]**
It is built on a probabilistic model in which the algorithm may be readily implemented and real-time predictions can be made Subsequently, this calculation is oftentimes used to deal with certifiable issues since it tends to be adjusted to answer client requests rapidly . But, before we go into Nave Bayes and Gaussian Nave Bayes, it's important to understand what conditional probability is.**[3]**

	precision	recall	f1-score	support
setosa	1.00	1.00	1.00	19
versicolor	0.93	0.93	0.93	15
virginica	0.94	0.94	0.94	16
accuracy			0.96	50
macro avg	0.96	0.96	0.96	50
weighted avg	0.96	0.96	0.96	50

	precision	recall	f1-score	support
setosa	1.00	1.00	1.00	19
versicolor	1.00	0.94	0.97	16
virginica	0.94	1.00	0.97	15
accuracy			0.98	50
macro avg	0.98	0.98	0.98	50
weighted avg	0.98	0.98	0.98	50

K-Nearest Neighbours: K-Nearest Neighbours is one of Machine Learning's most basic but crucial categorization algorithms. Supervised learning is used for a variety of applications such as pattern recognition, data mining, and intrusion detection. Non-parametric statistics are commonly used in real-world settings because they do not require assumptions about the data as to conveyance (rather than different calculations, for example, GMM, which expect a Gaussian circulation of the given information). Prior data (also known as training data) is provided, which divides coordinates into groups based on an attribute.[4]

CONCLUSION

In this paper we can across different machine learning data model like K-nearest neighbours, Random forest and Gaussian Naïve Bayes. On using all the models outcomes and picking the most occurring outcome of these models the accuracy is seen to be the highest.

REFRENCES

1) Iris flower data set- Wikipedia
2) what is machine learning- Edureka
3)Gaussian Naïve Bayes- Upgrad
4)K-Nearest Neighbours -GeeksforGeeks
5)Random forest model-Kaggle