COLLEGE STUDENT MARKS PREDICTION

PANKAJ RAINA (11910326)

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab, India

pankajraina05@gmail.com

Abstract-While performing tests on my topic which is Marks Prediction, I did experiment with a real-world dataset, and to explore how machine learning algorithms can be used to ease our work and to predict the new output with the help of the given data. I was expected to gain experience using a common machine learning library and was expected to submit a report and a research paper about the dataset and the algorithms used. After performing the required tasks on a dataset of my choice, herein lies my research paper.

I Introduction

Machine learning is a sub-domain of computer science which evolved from the study of pattern recognition in data, and from the computational learning theory in artificial intelligence. It is the first-class ticket to most interesting careers in data analytics today. As data sources proliferate along with the computing power to process them, going straight to the data is one of the most straightforward ways to quickly gain insights and make predictions

Machine Learning can be thought of as the study of a list of sub-problems, viz: decision making, clustering, classification, forecasting, deep-learning, inductive logic programming, support vector machines, reinforcement learning, similarity and metric learning, genetic a algorithm, sparse dictionary learning, etc. Supervised learning, or classification is the machine learning task of inferring a function from a labelled data In Supervised learning, we have m training set, and a test set. The training and test set consists of a set of examples consisting of input and output vectors, and the goal of the supervised learning algorithm is to infer a function that maps the input vector to the output vector with minimal error. In an optimal scenario, a model trained on a set of examples will classify an unseen example in a correct fashion, which requires the model to generalize from the training set in a reasonable way. In layman's terms, supervised learning can be termed as the process of concept learning, where a brain is exposed to a set of inputs and result vectors and the brain learns the concept that relates said inputs to output A wide array of supervised machine learning algorithms are available to the machine learning enthusiast, for example Neural Networks Decision Trees, Support Vector Machines Random Forest, Naive Bayes Classifier, Bayes Net, Majority Classifier(4.7.8.9) etc. and they each have their own merits and demerits. There is no single algorithm that works for all cases, as merited by the No free lunch theorem. In this project, I tried and found patterns in a dataset. which is a sample of marks obtained by studying for certain number of hours and attempt to throw various intelligently picked algorithms at the data and see what sticks.

II History of Machine Learning

The name machine learning was coined in 1959 by Arthur Samuel. Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field: "A computer program is said to learn from experience E with respect to some class of tasks 7 and performance measure P if its performance at tasks in 7, as measured by P, improves with experience E." This follows Alan Turing's proposal in his paper "Computing Machinery and Intelligence", in which the question "Can

machines think?" is replaced with the question "Can machines do what we (as thinking entities) can do?". In Turing's proposal the characteristics that could be possessed by a thinking machine and the various implications in constructing one are exposed.

III Types of Machine Learning

The types of machine learning algorithms differ in their approach, the type of data they input and output, and the type of task or problem that they are intended to solve. Broadly Machine Learning can be categorized into four categories.

1. Supervised Learning

2.Unsupervised Learning

3.Reinforcement Learning

4.Semi-supervised Learning

Machine learning enables analysis of massive quantities of data. While it generally delivers faster, more. accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly.

Supervised Learning

Supervised Learning is a type of learning in which we are given a data set and we already know what correct output should look like, having the idea that there is a relationship between the input and output. Basically, it is learning task of learning a function that maps an input to an output based on example input output pairs. It infers a function from labeled training data consisting of a set of training examples. Supervised learning problems are categorized

Unsupervised Learning

Unsupervised Learning is a type of learning that allows us to approach problems with little or no idea what our problem should look like. We can derive the structure by clustering the data based on a relationship among the variables in data. With unsupervised learning there is no feedback based on prediction result. Basically, it is a type of self-organized learning that helps in finding previously unknown patterns in data set without pre-existing label.

Reinforcement Learning

Reinforcement learning is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the lineal behaviour within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best.

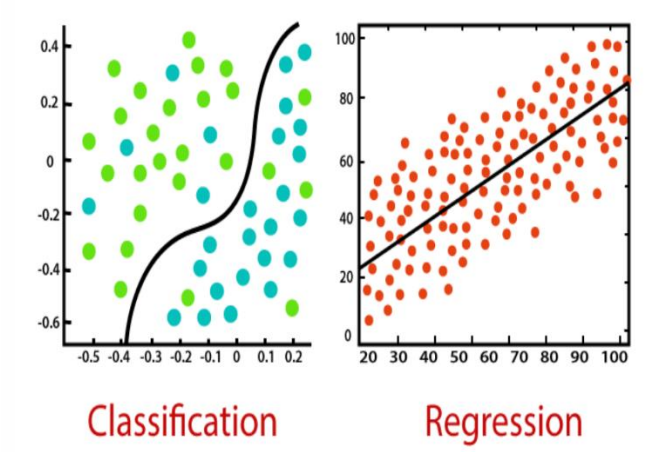
Semi-supervised Learning

Semi-supervised leaming falls somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training-typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method can considerably improve learning accuracy. Usually, sem-supervised Iraming is chosen when the acquired labeled data requires skilled and relevant resources in order to train it/ leam from it. Otherwise, acquiring unlabeled data generally doesn't require additional resources.

Classification is about predicting a label, by identifying which category an object belongs to based on different parameters.

Regression is about predicting a continuous output, by finding the correlations between dependent and independent variables.

Author: Pankaj Raina
Lovely Professional University



IV Linear Regression

Linear Regression is known as one of the simplest Machine learning algorithms that branch from Supervised Learning and is primarily used to solve regression problems.

The use of Linear Regression is to make predictions on continuous dependent variables with the assistance and knowledge from independent variables. The overall goal of Linear Regression is to find the line of best fit, which can accurately predict the output for continuous dependent variables. Examples of continuous values are house prices, age, and salary.

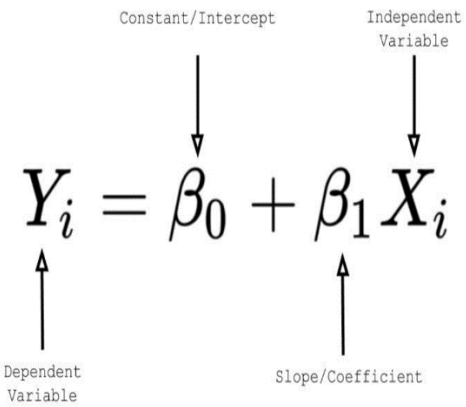
Simple Linear Regression is a regression model that estimates the relationship between one single independent variable and one dependent variable using a straight line. If there are more than two independent variables, we then call this Multiple Linear Regression.

Using the strategy of the line of best fits helps us to understand the relationship between the dependent and independent variable, which should be of linear nature.

The Formula for Linear Regression

If you remember high school Mathematics, you will remember the formula: ***y = mx + b*** and represents the slope-intercept of a straight line. 'y' and 'x' represent variables, 'm' describes the slope of the line and 'b' describe the y-intercept, where the line crosses the y-axis.

For Linear Regression, 'y' represents the dependent variable, 'x' represents the independent variable, $\beta 0$ represents the y-intercept and $\beta 1$ represents the slope, which describes the relationship between the independent variable and the dependent variable.



V Literature Review

A core objective of a leamer is to generalize from its experience. The computational analysis of machine leaming algorithms and their performance is a branch of theoretical computer

Author: Pankaj Raina
Lovely Professional University

science known as computational learning theory. Because training sets are finite and the future uncertain, learning theory usually does not yield guarantees of the performance of algorithms. Instead, probabilistic bounds on the performance are quite common. The bias variance decomposition is one way to quantify generalization error.

For the best performance in the context of generalization, the complexity of the hypothesis should match the complexity of the function underlying the data. If the hypothesis is less complex than the function, then the model has underfit the data. If the complexity of the model is increased in response, then the training error decreases. But if the hypothesis is too complex. the the model is subject to overfitting and generalization will be poorer

In addition to performance bounds, learning theorists study the time complexity and feasibility of learning in computational learning theory, a computation is considered feasible if it can be done in polynomial time. There are two kinds of time complexity results. Positive results a show m that a certain a class n of functions can be learned in polynomial time. Negative results show that certain classes cannot be learned in polynomial time.

VI The Challenges which Machine Learning is facing

While there has been much progress in machine learning, there are also challenges. For example, the mainstream machine learning technologies are black-box approaches, making us concerned about their potential risks. To tackle this challenge, we may want to make machine learning more explainable and controllable as another example, the computational comp of machine learning algorithms is usually very high, and we may want to invent lightweight algorithms or implementations. Furthermore, in many domains such as physics, chemistry, biology, and social sciences, people usually seek elegantly simple t equations (eg, the Schrödinger equation) to uncover the underlying laws behind various phenomena. Machine learning takes much more time. You must gather and prepare data, then train the algorithm. There are much more uncertainties. That is why, while in traditional website or application development an experienced team can estimate the time quite precisely, a machine learning project used for example to provide product recommendations can take much less or much more time than expected. Why? Because even the best machine learning engineers don't know how the deep learning networks will behave when analysing different sets of data. It also means that the machine learning engineers, and data scientists cannot guarantee that the training process of a model can be replicated.

VII Applications of Machine Learning

Machine learning is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that which makes it more similar to humans: The ability to leam. Machine learning is actively being used today, perhaps in many more places than one would expect. We probably use a learning algorithm dozen of time without even knowing it. Applications of Machine Learning include

Web Search Engine: One the reasons why search engines like google, Bing etc work so well is because the system has learnt how to rank pages through a complex learning algorithm.

Photos tagging Applications: Be it Facebook or any other photo tagging application, the ability to tag friends makes it even more happening. It is all possible because of a face recognition algorithm that runs behind the application.

Spam Detector: Our mail agent like Gmail or Hotmail does a lot of hard work for us in classifying the mails and moving the spam mails to spam folder. This is again achieved by a spam classifier running in the back end of mail application.

VIII Technology Implemented

Python - The New Generation Language

Python is a widely used general-purpose, high level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for an emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code. Python is dynamically typed and garbage collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

Features

1.Interpreted
In Python there is no separate compilation and execution steps like C/C++. It directly run the program from the source code. Internally, Python converts the source code into an intermediate form called bytecodes which is then translated into native language of specific computer to run it.

2.Platform Independent

Python programs can be developed and executed on the multiple operating system platform. Python can be used on Linux, Windows, Macintosh, Solaris and many more.

3.Multi-Paradigm

Python is a multi-paradigm programming language. Object-oriented programming and structured programming are fully supported, and many of its features support functional programming and aspect-oriented programming.

4.Simple

Python is a very simple language. It is a very easy to learn as it is closer to English language. In python more emphasis is on the solution to the problem rather than the syntax.

5.Rich Library Support

Python standard library is very vast. It can help to do various things involving regular expressions. documentation generation, unit testing, threading, databases, web browsers, CGI, email, XML, HTML, WAV files, cryptography, GUI and many more.

Free and Open Source

Firstly, Python is freely available. Secondly, it is open source. This means that its source code is available to the public. We can download it, change it, use it, and distribute it. This is called FLOSS (Free/Libre and Open-Source Software). As the Python community, we're all headed toward one goal - an ever-bettering Python.

IX Why Python Is a Perfect Language for Machine Learning?

1. A great library ecosystem -

A great choice of libraries is one of the main reasons Python is the most popular programming language used for Al. A library is a module or a group of modules published by different sources which include a pre-written piece of code that allows users to reach some functionality or perform different actions. Python libraries provide base level items so developers don't have to code them from the very beginning every time. ML. requires continuous data processing, and Python's libraries. let us access, handle and transform data. These are some of the most widespread libraries you can use for ML and AI:

- Scikit-leam for handling ML algorithms like clustering, linear and logistic regressions, regression, classification, and others.
- Pandas for high-level data structures and analysis. It allows merging and filtering of data, as well as gathering it from other external sources like Excel, for instance. Keras for deep learning. It allows fast calculations and prototyping, as it uses the GPU in addition to the CPU of the computer.
- TensorFlow for working with deep learning by setting up, training, and utilizing artificial neural networks with massive datasets Matplotlib for creating 2D plots, histograms charts, and other forms of visualization.

- NLTK for working with computational linguistics, natural language recognition, processing.
- Scikit-image for image processing
- PyBrain for neural networks, unsupervised and reinforcement learning o Caffe for deep learning that allows switching between the CPU and the GPU and processing 60+ min images a day using a single NVIDIA K40 GPU.
- StatsModels for statistical algorithms and data exploration.

In the PyPI repository, we can discover and compare more python libraries

2. A low entry barrier

Working in the ML and Al industry means dealing with a bunch of data that we need to process in the most convenient and effective way. The low entry barrier allows more data scientists to quickly pick up Python and start using it for Al development without wasting too much effort into learning the language. In addition to this, there's a lot of documentation available, and Python's community is always there to help out and give advice.

3. Flexibility

Python for machine learning is a great choice, as this language is very flexible:

- It offers an option to choose either to use OOPS or scripting.
- There's also no need to recompile the source code, developers can implement any changes and quickly see the results.
- Programmers can combine Python and other languages to reach their goals.

4. Good Visualization Options

For Al developers, it's important to highlight that in artificial intelligence, deep learning, and machine learning, it's vital to be able to represent data in a human-readable format. Libraries like Matplotlib allow data scientists to build charts, histograms, and plots for better data comprehension. effective presentation, and visualization. Different application programming interfaces also simplify the visualization process and make it easier to create clear reports.

5. Community Support

It's always very helpful when there's strong community support built around the programming language. Python is an open-source language which means that there's a bunch of resources open for programmers starting from beginners and ending with pros. A lot of Python documentation is available online as well as in Python communities and forums, where programmers and machine learning developers discuss errors, solve problems, and help each other out. Python programming language is absolutely free as is the variety of useful libranes and tools

6. Growing Popularity

As a result of the advantages discussed above, Python is becoming more and more popular among data scientists According to Stack Overflow, the popularity of Python is predicted to grow until 2020. at least. This means it's easier to search for developers and replace team players if required. Also, the cost of their work maybe not as high as when using a less popular programming language

X Data Preprocessing, Analysis & Visualization

Machine Learning algorithms don't work so well with processing raw data. Before we can feed such data) to an ML algorithm, we must preprocess it. We must apply some transformations on it. With data preprocessing, we convert raw data into a clean data set. To perform data this, there are 7 techniques

1.Rescaling Data

For data with attributes of varying scales, we can rescale attributes to possessi same scale. We rescale attributes into the range 0 to 1 and call it normalization. We use the MinMaxScaler class from scikit learn. This gives us values between 0 and 1.

2. Standardizing Data

With standardizing, we can take attributes with a Gaussian distribution and different means and standard deviations and transform them into a standard Gaussian distribution with a mean of 0 and a standard deviation of 1.

3. Normalizing Data

In this task, we rescale each observation to a length of 1 (a unit norm). For this, we use the Normalizer class.

4. Binarizing Data

Using a binary threshold, it is possible to transform our data by marking the values above it I and those equal to or below it, 0. For this purpose, we use the Binarizer class

5. Mean Removal

We can remove the mean from each feature to center it on zero

6. One Hot Encoding

When dealing with few and scattered numerical values, we may not need to store these. Then, we can perform One Hot Encoding. For k distinct values, we can transform the feature into a k-dimensional vector with one value of 1 and 0 as the rest values

7. Label Encoding -

Some labels can be words or numbers. Usually, training data is labelled with words to make it readable. Label encoding converts word labels into numbers to let algorithms work on them.

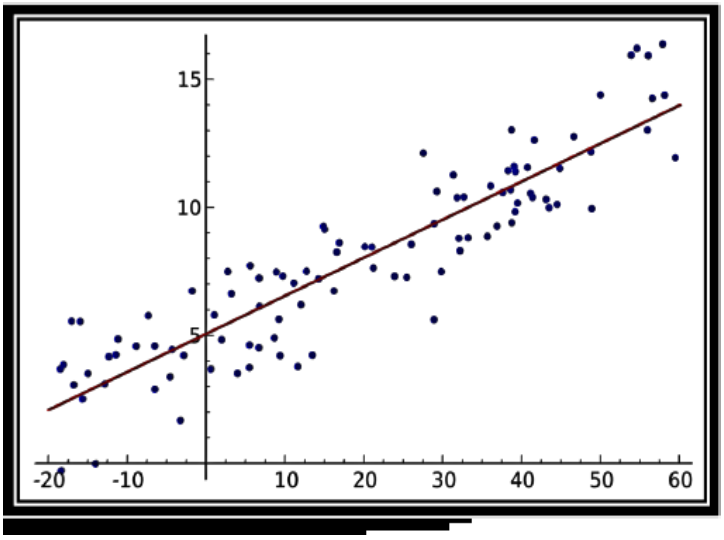
Machine Learning Algorithms

There are many types of Machine Learning Algorithms specific to different use cases. As we work with datasets, a machine learning algorithm works in two stages. We usually split the data around 20%-80% between testing and training stages. Under supervised learning, we split a dataset into a training data and test data in Python M Followings are the Algorithms of Python Machine Learning -

1. Linear Regression

Linear regression is one of the supervised Machine learning algorithms in Python that observes continuous features and predicts an outcome. Depending on whether it runs on a single variable or on many features, we can call it simple linear regression or multiple linear regression. This is one of the most popular Python ML algorithms and often under-appreciated. It assigns optimali weights to variables to create a line ax+b to predict the output We often u linear regression to estimate real values like a number of calls and costs of houses based on continuous variables. The regression line is

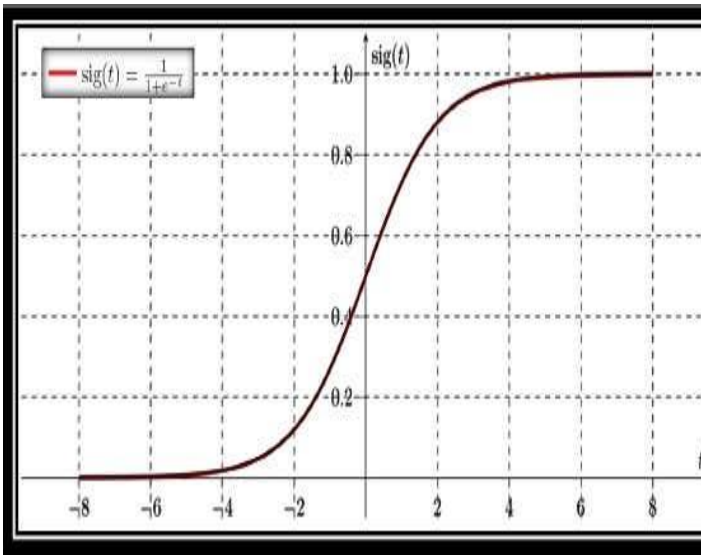
the best line that fits Y-a*X+b to denote a relationship between independent and dependent variables.



2. Logistic Regression -

Logistic regression is a supervised classification is unique Machine Learning algorithms in Python that finds its use in estimating discrete values like 0/1, yes/no, and true false. This is based on a given set of independent variables. We use a logistic function to predict the probability of an event, and this gives us an output between 0 and 1. Although it says 'regression', this is actually a classification algorithm. Logistic regression fits data into a logit function and is also called logit regression.

Author: Pankaj Raina
Lovely Professional University

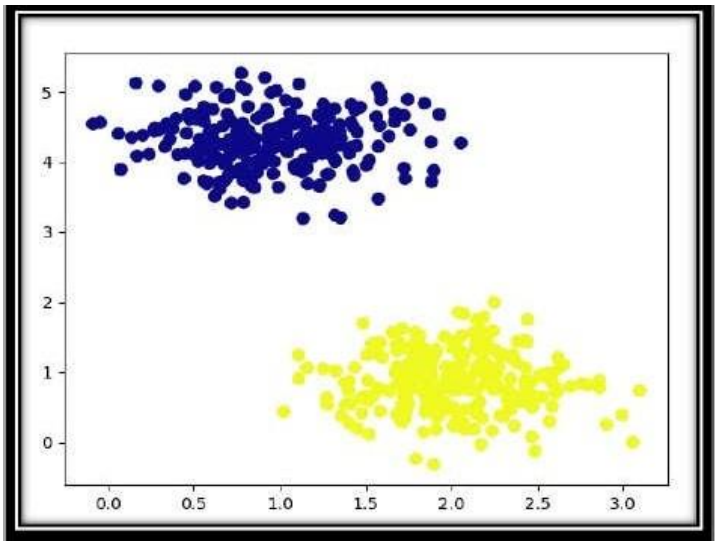


3. Decision Tree

A decision tree falls under supervised Machine Leaming Algorithms in Python and comes of use for both classification and regression- although mostly for classification. This model takes an instance, traverses the tree, and compares important features with a determined conditional statement. Whether it descends to the left child branch, or the right depends on the result. Usually, more important features are closer to the root. Decision Tree, a Machine Learning algorithm in Python can work on both categorical and continuous dependent variables. Here, we split a population into two or more homogeneous sets. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

4. Support Vector Machine (SVM)

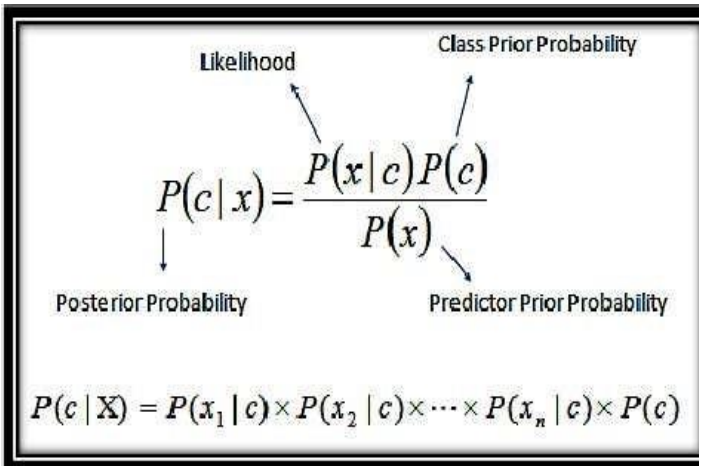
SVM is a supervised classification is one of the most important Machines Learning algorithms in Python, that plots a line that divides different categories of your data. In this ML. algorithm, we calculate the vector to optimize the line. This is to ensure that the closest point in each group lies farthest from each other. While you will almost always find this to be a linear vector, it can be other than that. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. When data are unlabeled, supervised learning is not possible. and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups.



5. Naive Bayes Algorithm -

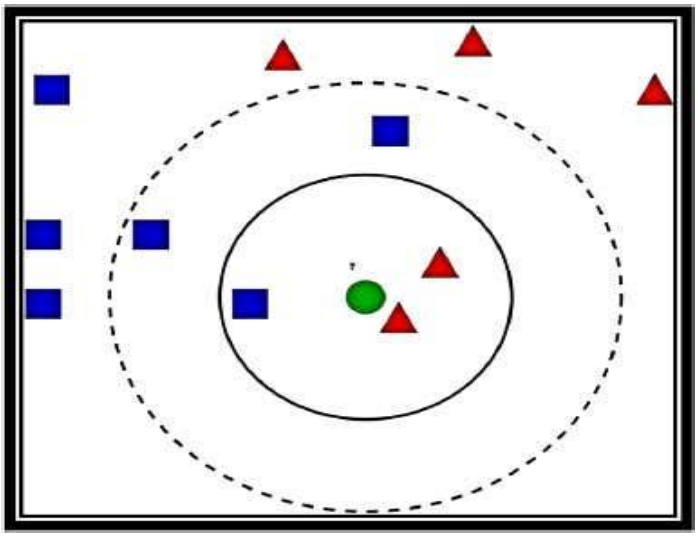
Naive Bayes is a classification method which is based on Bayes theorem. This assumes independence between predictors. A Naive Bayes classifier will assume that a feature in a class is unrelated to any other. Consider a fruit. This is an apple if it is round, red, and 25 inches in diameter. A Naive Bayes classifier will say these characteristics independently contribute to the probability of the fruit being an apple. This is even if features depend on each other. For very large data sets, it is easy to build a Naive Bayesian model. Not only is this model very simple, it performs better than many highly sophisticated classification methods. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

Author: Pankaj Raina
Lovely Professional University



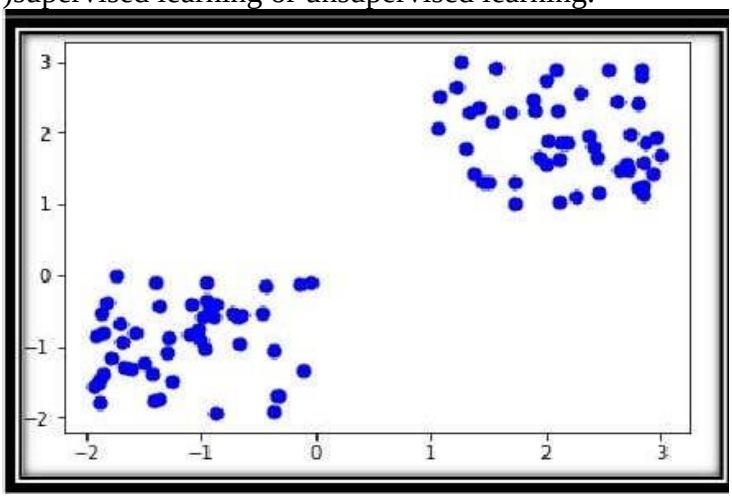
6. KNN Algorithm.

This is a Python Machine Learning algorithm for classification and regression mostly for classification. This is a supervised learning algorithm that considers different centroids and uses a usually Euclidean function to compare distance. Then, it analyses the results and classifies each point to the group to optimize it to place with all closest points to it. It classifies new cases using a majority vote of k of its neighbours. The case it assigns to a class is the one most common among its K nearest neighbours. For this, it uses a distance function. ANN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. ANN is a special case of a variable bandwidth, kernel density "balloon" estimator with a uniform kernel.



7. K-Means Algorithm

k-Means is an unsupervised algorithm that solves the problem of clustering It classifies data using a number of clusters. The data points inside a class are homogeneous and heterogeneous to peer groups k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining A-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. A-means clostering is rather easy to apply to even large data sets, particularly when using heuristics such as Lloyd's algorithm. It often is used as a preprocessing step for other algorithms, for example to find a starting configuration. The problem is computationally difficult (NP-hard), k-means originates from signal processing, and still finds use in this domain. In cluster analysis, the k-means algorithm can be used to partition the input data set into k partitions (clusters), k-means clustering has been used as a feature learning (or dictionary learning) step, in either (semi-)supervised learning or unsupervised learning.

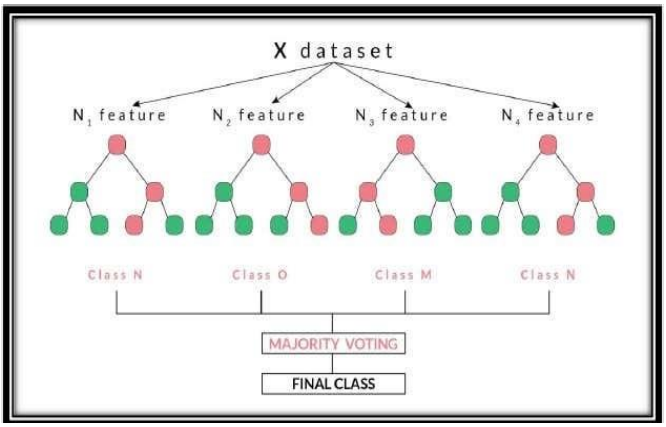


8. Random Forest

A random forest is an ensemble of decision trees. In order to classify every new object based on its attributes, trees vote for class-cach tree provides a classification. The classification with the most votes wins in the forest. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification)

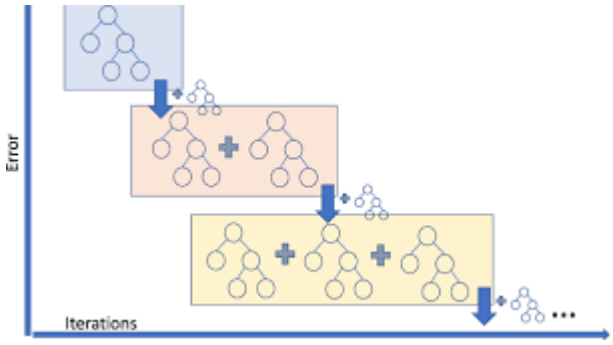
Author: Pankaj Raina
Lovely Professional University

or mean prediction (regression) of the individual trees.



9. Gradient Boosting Regressor

Gradient Boosting algorithm is used to generate an ensemble model by combining the weak learners or weak predictive models.
Gradient boosting algorithm can be used to train models for both regression and classification problem. Gradient Boosting
Regression algorithm is used to fit the model which predicts the continuous value. Gradient boosting builds an additive mode
by using multiple decision trees of fixed size as weak learners or weak predictive models. The parameter, n_estimators, decides
the number of decision trees which will be used in the boosting stages. Gradient boosting differs from AdaBoost in the manner that
decision stumps (one node & two leaves) are used in AdaBoost whereas decision trees of fixed size are used in Gradient Boosting.



XI Dataset

My dataset is a collection of result of marks obtained by studying for certain number of hours. I collected this dataset by my own, for this I asked students about their marks and their study hours.

Study_hours	Student marks
6.83	78.5
6.56	76.74
	78.68
5.67	71.82
8.67	84.19
7.55	81.18
6.67	76.99
8.99	85.46
5.19	70.66
6.75	77.82
6.59	75.37
8.56	83.88
7.75	79.5
7.9	80.76
8.19	83.08
6.55	76.03
6.36	76.04

By doing so I was able to create a appropriate dataset consisting of samples of students marks and study hours. Given above is how my dataset looks like a matrix.

Working: - Linear regression is used to study the linear relationship between a dependent variable Y (students marks) and one or more independent variables X (study hour).

The dependent variable Y must be continuous, while the independent variables may be either continuous, binary, or categorical. The initial judgment of a possible relationship between two continuous variables should always be made on the basis of a scatter plot (scatter graph). This type of plot will show whether the relationship is linear (figure 1) or nonlinear (figure 2).

Author: Pankaj Raina
Lovely Professional University

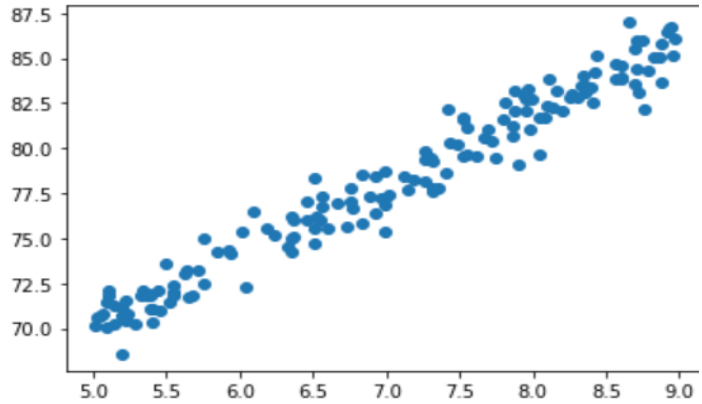


Figure 1
A scatter plot showing linear relationship.

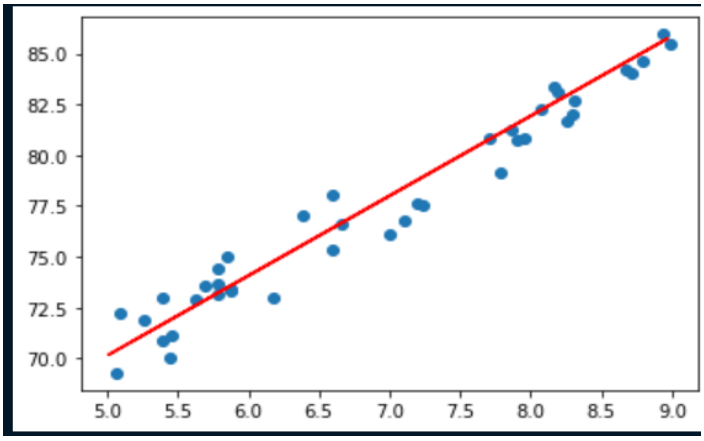


Figure 2
A scatter graph showing exponential relationship. In this case, it would not be appropriate to compute a coefficient of determination or a regression line.

Performing a linear regression makes sense only if the relationship is linear. Other methods must be used to study nonlinear relationships. The variable transformations and other, more complex techniques that can be used for this purpose will not be discussed in this article.

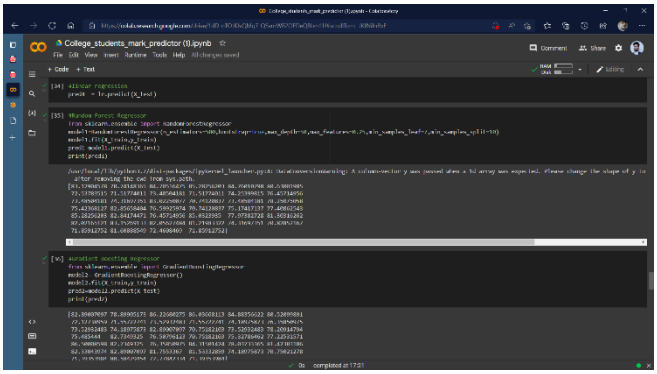
As shown below the predicted data is in simpler format. It is showing how our model is predicting the students marks according to the given marks obtained with respect to the study time.

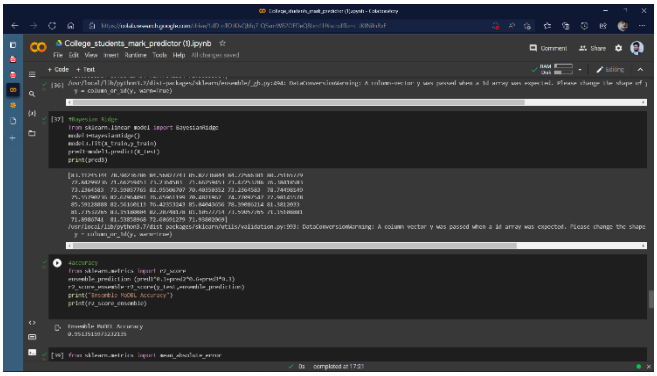
	study_hours	student_marks_original	student_marks_predicted
0	8.300000	82.02	83.113815
1	7.230000	77.55	78.902596
2	8.670000	84.19	84.570030
3	8.990000	85.46	85.829460
4	8.710000	84.03	84.727459
5	7.700000	80.81	80.752384
6	5.690000	73.61	72.841591
7	5.390000	70.90	71.660875
8	5.790000	73.14	73.235162
9	5.390000	73.02	71.660875
10	5.850000	75.02	73.471305
11	6.590000	75.37	76.383737
12	5.790000	74.44	73.235162
13	5.880000	73.40	73.589377
14	8.260000	81.70	82.956386
15	5.070000	69.27	70.401445
16	5.790000	73.64	73.235162
17	7.190000	77.63	78.745168
18	6.380000	77.01	75.557236
19	8.190000	83.08	82.680886
20	6.660000	76.63	76.659237

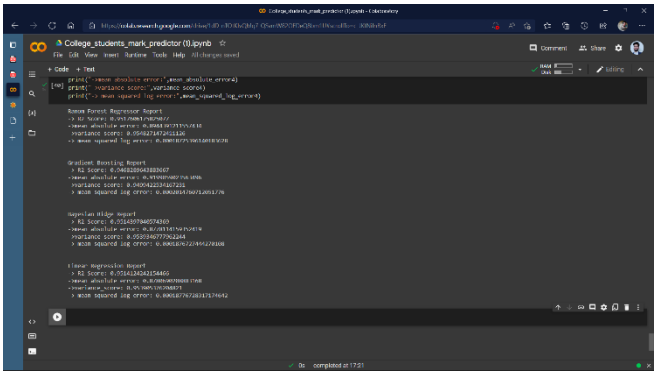
XII COMPARATIVE STUDY

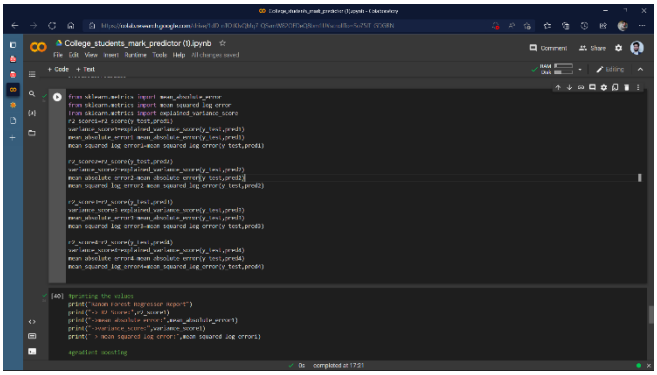
After using different algorithms for continuous data prediction, we got a slight difference in accuracy, r2 score, variance score, mean squared log error and mean absolute error. The Following given images below show the difference in values after using various supervised learning algorithms. For model to be good it is estimated that r2 score should be 0.9 or above and for the variance score it should not be less than 60%. The lesser the mean squared error the higher the accuracy. Random forest regressor uses ensemble learning method that is a technique which combines predictions from multiple machine learning algorithms. We got the overall accuracy around 95% that is a good sign for our model.

Author: Pankaj Raina
Lovely Professional University









```
Ranom Forest Regressor Report
-> R2 Score: 0.9517606175825077
->mean absolute error:
0.8944391211557434
->variance_score: 0.9548271472411126
-> mean squared log error:
0.0001872539614018362 8
```

```
Gradient Boosting Report
-> R2 Score: 0.9468289643883667
->mean absolute error:
0.9199059023563496
->variance_score: 0.9499422534167231
-> mean squared log error:
0.0002014760712051776
```

```
Bayesian Ridge Report
-> R2 Score: 0.9514397040574369
->mean absolute error:
0.8778114159352419
->variance_score: 0.9539346777962244
-> mean squared log error:
0.0001876727444270108
```

```
Linear Regression Report
-> R2 Score: 0.9514124242154466
->mean absolute error:
0.8780690208883168
->variance_score: 0.953905376204821
-> mean squared log error:
0.0001877672831717464 2
```

XIII FUTURE SCOPE

Machine Learning ahs already established it as a modern giant in todays automated world. Though its various advantages make it easy to resolve the computational time as well space complexity. Various algorithms serving different purposes and falling in different categories to fulfil the requirements of end user has set up a high standard for data analysis and in field of research. Likewise in the similar fashion this model is set up on the pillars of supervised learning algorithms because of the training data given for testing and validating. Techniques used for practical implementation include linear regression, random forest regression, Bayesian ridge and Gradient Boosting to predict the accurate output in less time. Classification and Regression can be fulfilled using these algorithms of supervised learning. Model is generating the output of form percentage which takes input as study hours and pre saved data that is study percentage and then predicting the output. This made it easy for end user who doesn't has any prior knowledge of algorithms to practically understand the output in layman language but for research it will generate curiosity to know and get practical knowledge in order to design such machine learning models for automation and minimum human effort to be involved.

Author: Pankaj Raina
Lovely Professional University

XIIV Advantages and Disadvantages

Advantages of Gradient Boosting in our model are it provides predictive accuracy that cannot be trumped, and it can optimize on different loss functions and provides several hyper parameter tuning options that make the function fit very flexible. And no data pre-processing required - often works great with categorical and numerical values as is. Let's also take advantages of logistic regression its advantages are it is easier to implement, interpret, and very efficient to train it makes no assumption about distribution of glasses in feature space it can easily excel to multiple classes and a natural prognostic view of glass prediction it not only provides a pleasure of how appropriate a predictor but also its direction of association whether positive or negative it is very fast at classifying unknown record accuracy for many simple dataset and it performs well when the dataset linearly separable it can interpret model coefficients as indicators of future important logistic regression is less land overheating but it can over hi dimensional data cells one may consider regularisation L1 and L2 techniques to avoid overfitting these scenarios. Now let's see of the disadvantages of the algo and techniques used to build this model. Gradient Boosting Models will continue improving to minimize all errors. This can overemphasize outliers and cause overfitting. computationally expensive - often require many trees (>1000) which can be time and memory exhaustive, the high flexibility results in many parameters that interact and influence heavily the behaviour of the approach (number of iterations, tree depth, regularization parameters, etc.). This requires a large grid search during tuning. Let's see some of the disadvantages of using linear regression, If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting. The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. Non-linear problems can't be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios. It is tough to obtain complex relationships using logistic regression. More powerful and compact algorithms such as Neural Networks can easily outperform this algorithm.

Author: Pankaj Raina
Lovely Professional University

References

KD Nuggets
https://www.kdnuggets.com/2022/03/linear-logistic-regression-succinct-explanation.html

Academia
https://www.academia.edu/74815512/Measures_of_influence_for_the_functional_linear_model_with_scalar_response

National Library of Medicine
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2992018/

ResearchGate
https://www.researchgate.net/publication/324944461_Linear_regression_analysis_study