IT Job Analysis and Forecasting Using Naive Bayes Algorithm

M.Kannadasan[1], Dr.K.Venkataramana[2]
[1]*Student, M.Kannadasan, KMM Institute of Post Graduate Studies*
[2]*Associative professor, Dr.K.Venkataramana, KMM Institute of Post Graduate Studies*

***Abstract*- Information Technology (IT) is an example of a general purpose technology that has the potential to play an important role in economic growth, as well as other dimensions of economic and social development. IT may have a special role to play in growth and development simply because of empirical characteristics that apply at the current time. Now days in industry, so many jobs are there based on different platforms like Java, .NET, Oracle, PHP like that. So it is necessary to collect the information about jobs in all platforms and predict the future analysis. Because of its contribution in improving the economy of country and its importance in future. in this paper we are we are going to predict the it jobs in future by using the data mining techniques and predict which platform will have how many jobs based on the previous data collected by us.**

***Index Terms*- Naive Bayes; Decision tree; SQL database**

INTRODUCTION

Today, India is home to a portion of the finest programming organizations on the planet. The Software organizations in India are presumed over the globe for their productive IT and business related arrangements. The Indian Software industry can give the important driving force to the Indian Economy toward development. The product administrations industry in India has made a key stage for the Software items Industry to prosper and underwrite upon. Specifically, the current and proceeding with fast development in IT make it a dynamic segment that is an appealing hopeful as a supporter of development consequently alone. Step by step the IT businesses are expanding significantly. Indian educational modules is behind circumstances as far programming dialects are concerned. They stay with BASIC, FORTRAN, and some "set apart for death" like PERL, Flash, Algal, and Object Pascal; how are these understudies anticipated that would make progress into a universe of Java, C, C++, Python, Ruby on Rails, and so on [1].

The world has entered the new thousand years, which will be the Information Technology Age. PCs and Internet (World-Wide Web) have turned into a basic piece of our own and expert lives; IT Careers have increased monstrous notoriety in the course of recent years. With the appearance of the PC the data innovation industry experienced a quantum change. Today Computers have not just expected vital significance in the corporate world, they are by and large adequately utilized as a part of different fields going from space investigation to sustenance handling and managing an account to correspondence and so forth. The product unrest completely changed the way we work. Accessibility of modest and simple to utilize programming bundles has expanded profitability levels complex. Most likely no segment is untouched by data innovation. Assembling, Finance, Marketing, Entertainment, Education, Mass Media, Environment, Communication and a few different fields are receiving its rewards [2], [3]. In this time of Information Technology, which has changed the entire world, India has confronted the world guidelines and is being respected the world over for its talented IT Professionals [4]. Despite the fact that trusted that there would be a droop in the IT field, it keeps on developing, and offers openings for work to individuals who have the correct aptitudes and preparing. Data Technology occupations touch about each field in all aspects of the nation and by that sheer nearness itself offer countless employment opportunities. The quick advancement of innovations, for example, organizing, multi-media and the Internet/WWW have made absolutely new employment classifications where none existed a couple of years prior. This segment is likewise the one that is seeing the quickest development and

change rate. New programming and strategies turn out each month and experts need to keep pace with the fast progressions [5].

A. SKILLS AND PERSONAL QUALITIES

One should have the following personal attributes if one wishes to enter this field:

- Flexibility and willingness to learn new things, technologies and adopt new methods of work
- Logical thinking
- Creativity
- Ability to focus and concentrate
- Accuracy
- Organizational and administrative abilities
- Confidence
- Ready to work for long hours and ability to work hard
- High intellectual capacity
- Ability to take decisions
- Ability to get well with people and good communication skills
- Academic and technical skills.

Nearly 70% of the IT students in India are not unemployed because they don't have correct idea about which platform will have better necessity in future. No one can predict which course having demand in future. In this case students can face so many problems about their jobs in future.

So it is necessary to collect the information from different websites and it companies and blogs to give the correct guidance for the future employees.

Today, Bangalore is known as the Silicon Plataea of India and contributes 38% of Indian IT Exports. India's second and third largest software companies are headquartered in Bangalore, as are many of the global Companies. Cities like Hyderabad, Chennai, Pune and Gurgaon are also emerging as technology hubs, with many global IT companies establishing headquarters there. Numerous IT companies are also based in Mumbai. So by collecting the efficient information from all these companies and from all these cities we will get correct information about the past data and we can predict the future analysis [2].

II. RELATED WORK

The contribution of software to India's economic development, paying particular attention to the role of software in the absorption of labour and the development of human capital in the economy. The success of the software industry has increased the relative value of professional workers, not only programmers, but also managers and analysts. The growing importance of human capital, in turn, has led to innovative models of entrepreneurship and organization, pioneered by the software sector, and these are slowly taking root and spreading to other sectors of India's industry [1].

Data innovation (IT) is a case of a universally useful innovation that can possibly assume an essential part in monetary development, and in addition different measurements of financial and social improvement. This paper audits a few interrelated parts of the part of data innovation in the development of India's economy. It considers the surprising accomplishment of India's product send out part and the overflows of this accomplishment into different IT empowered administrations, endeavors to make IT and its advantages accessible to India's provincial masses, web based business for the nation's developing white collar class, the utilization and effects of IT in India's assembling segment, and different types of e-administration, including inward frameworks and additionally subject interfaces [3].

India's manufacturing sector is receiving renewed attention as an underperformer in contributing to the nation's Gross Domestic Product (GDP) and employment growth, with a new National Manufacturing Policy (NMP) stating ambitious goals for increasing the share of manufacturing in GDP. In this context, the role of information technology (IT) as a contributor to manufacturing productivity also needs to be carefully examined. This paper uses five years of panel data for Indian manufacturing plants to examine the relationship of investment in IT to productivity, as measured by gross value added. We find some evidence that plants with higher levels of IT capital stock have higher gross value added, controlling for other inputs [4].

III. METHODOLOGY

The steps Involved in the IT job analysis is:

1. Data Collection
2. Classification
3. Prediction
4. Visualization

A. DATA COLLECTION

In data collection we are going to collect different information from different websites, blogs and from past data records of the companies' etc.The collected data is stored in the database for the further analysis. Here we are going to store our data in the SQL database. In this SQL database we should know what we storing in advance and we should specify its size.

B. CLASSIFICATION

For the classification of data what we have collected we are using one algorithm called naive Bayes. It is a supervised learning method. Naive Bayes classifier is a probabilistic classifier which when given an input gives a probability distribution of set of all classes rather than providing a single output.

Naive Bayes is a direct framework for building classifiers: models that dole out class marks to issue examples, spoke to as vectors of highlight esteems, where the class names are drawn from some limited set.

C. PREDICTION

For prediction we are using the decision tree concept. A decision tree is similar to a graph in which internal node represents test on an attribute, and each branch represents outcome of a test. The main advantage of using decision tree is that it is simple to understand and interpret. The other advantages include its robust nature and also it works well with large data sets. This feature helps the algorithms to make better decisions about variables.

The working of decision tree seems to be little confusing but it's really easy. Consider a variety of plant species. We classify them according to order, genus, species etc. Instead we have to classify them into a common category as shrubs and trees. If a new species is identified then we have to classify this into any of the two categories. Basically we categorize it based on its characteristics i.e. we have a set of questions to check whether it satisfies the conditions. If first condition is satisfied then we check the next case and if the first condition itself is not satisfied then there is no need to check the rest. So the series of questions and their answers can be organized in the form of a decision tree. The tree has three types of nodes:

- A Root node, that has incoming edges and zero or more outgoing edges.
- Internal nodes, each of which has one incoming edge and two or more outgoing edges.
- Leaf node or end node, each of which has exactly one incoming edge and no outgoing edges.

This supervised machine learning technique builds a decision tree from a set of class labeled training samples and by using this tree, tests the new samples. It is a predictive model which uses a set of binary rules to calculate the class value. The tree determines:

- Which variable to split at a node.
- Decision to stop or split.
- Assign terminal nodes.

DECISION TREE ALGORITHM PSEUDO CODE

1. Place the best attribute of the dataset at the **root** of the tree.
2. Split the training set into **subsets**. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

For applying decision tree on the programing language the resulted figure as shown below

Figure shows the Example of decision tree

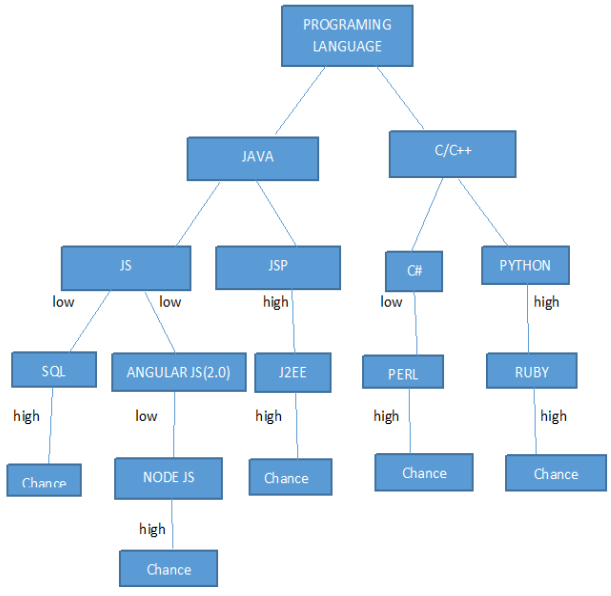

Fig1.Decision tree

D. VISUALIZATION

We can analyze the programing languages like java , python , javascript , c# and c++ ….etc. The predicted jobs for last two years can be graphically represented as follows and it is shown in the below graph.

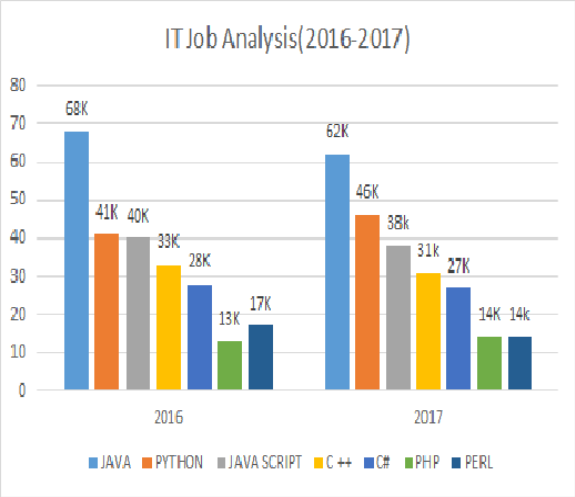

Fig2.Graphical representation of IT jobs in India

IV. ALGORITHM

A. NAIVE BAYES ALGORITHM

It is a supervised learning method. Naive Bayes classifier is a probabilistic classifier which when given an input gives a probability distribution of set of all classes rather than providing a single output.

The advantage of using Naive Bayes Classifier is that it is simple, and converges quicker than logistic regression. Compared to other algorithms like SVM (Support Vector Machine) which takes lot of memory the easiness for implementation and high performance makes it different from other algorithms. Also in case of SVM as size of training set increases the speed of execution decreases.

Naive Bayes is a direct framework for building classifiers: models that dole out class marks to issue examples, spoke to as vectors of highlight esteems, where the class names are drawn from some limited set. It isn't a solitary calculation for preparing such classifiers, yet a group of calculations in view of a typical guideline: all Naive Bayes classifiers expect that the estimation of a specific element is free of the estimation of some other element, given the class variable. For instance, a natural product might be thought to be an apple on the off chance that it is red, round, and around 10 cm in distance across. An Naïve Bayes classifier considers every one of these highlights to contribute freely to the likelihood that this natural product is an apple, paying little mind to any conceivable relationships between's the color, roundness, and measurement highlights.

B. NAIVE BAYES ALGORITHM PSEUDO CODE

Algorithm 1 Pseudocode

1. Given training data set D which consists of documents belonging to different class say class A and B.
2. Calculate the prior probability of class A=number of objects of class A / total number of objects
   Calculate the prior probability of class B=number of objects of class B / total number of objects
3. Find ni, the total number of word frequency of each class.
   na= the total number of word frequency of class A.
   nb= the total number of word frequency of class B.
4. Find conditional probability of keyword occurrence given a class.
   P(word1 / class A) = wordcount / ni(A)
   P(word1 / class B) =wordcount / ni(B)
   P(word2 / class A) = wordcount / ni(A)
   P(word2 / class B) =wordcount / ni(B)
   ……
   ……
   P(wordn / class B) =wordcount / ni(B)
5. Avoid zero frequency problems by applying uniform distribution.
6. Classify a new document C based on the probability P(C /W).
   a) Find P(A / W) = P(A) * P(word1/ class A) * P(word2/ class A) ……* P(wordn / class A).
   b) Find P(B / W) = P(B) * P(word1 / class B) * P(word2/ class B) ……* P(wordn / class B).
7. Assign document to class that has higher probability.

V. CONCLUSION

In this paper we have tested the efficiency of classification and prediction based on different data sets. Classification is done based on the Bayes theorem. Which shows more than 90% efficiency. Finally in this paper we are going to estimate the IT jobs in India by using the data collected from the previous records. And we represented it in the form of graph. We can easily predict which platform having high demand in future. In this case people don't face any problems about their jobs in future.

REFERENCES

[1] Arora, Ashish and Suma Athreye (2002), The Software Industry and India's Economic Development, Information Economics and Policy, 14, 253-273.

[2] Economist (2005), the Real Digital Divide, Technology and Development Survey, The Economist, March 10th.
[3] Eggleston, Karen, Robert T. Jensen, and Richard Zeckhauser (2002), Information and Communication Technologies, Markets, and Economic Development, in The Global Information Technology Report 2001-2002 Readiness for the Networked World, ed., Geoffrey Kirkman et al, Oxford: Oxford University Press, 62-74.
[4] Gangopadhyay, Shubhashis, Manisha G. Singh and Nirvikar Singh (2008), Waiting To Connect: Indian IT Revolution Bypasses The Domestic Industry, New Delhi: Lexis-Nexis-Butterworth.
[5] Helpman, Elhanan (1998), General Purpose Technologies and Economic Growth, ed., Cambridge, MA: MIT Press.