Use of Machine Learning for an Automated Approach to Human Capabilities Screening

Sunil Bhutada, Saylee Morey

Abstract— Machine Learning (ML) is paradigm that constantly learns by means of collecting past knowledge that makes use of historical data to know and trouble shooting. ML is a key enabler of Artificial Intelligence (AI). It is the contemporary method to virtual transformation making our computing method more efficient, cost powerful and dependable. ML algorithms are grouped based on their purpose a) Supervised learning b) Unsupervised learning c) Reinforcement learning. This project consists of assessment of human skills based on historic statistics. 'Capabilities' is the conceptualization for interpersonal comparisons of the capabilities to carry out a mission. Here the emphasis is on bridging the distance between qualitative and quantitative methods to classify talent assessment of human beings and produce report in statistical shape. This research applies machine-learning algorithms to information accrued after talent exams are taken by using applicants. A getting to know module might be built as a way to help subject matter experts (SMEs) to attain the computed assessment end result. These rankings could be used to improve the device and over a period of time the device gets better at assessing the evaluations. This approach will help pre-screening to be extra effective when evaluating applicants previous to induction into the device.

Keywords: PostgreSQL; Kafka; PySpark; ElasticSearch; Kibana; Machine Learning; Docker.

1. INTRODUCTION

Skill Assessment has a secret key to learning practice. Skill represent program of study aspiration, classify what is value expressive, force classroom practices. Have an vital to build up organizations for assessment , reproduce our central part educational goals, return apprentices for enlarging skills and characteristics, will provide benefit to the organisations and apprentices.In the course of recent decades Machine Learning has turned out to be one of the pillars of data innovation and with that, a fairly crucial, but generally concealed, some portion of our life. With the consistently expanding measures of information getting to be accessible there is valid justification to trust that perceptive information investigation will turn out to be considerably increasingly inescapable as a fundamental element for mechanical advancement. Machine learning is the way toward educating machines to learn. It exists at the convergence of software engineering, measurements, and straight variable based math, with experiences from neuroscience and different fields too. Be that as it may, in contrast to customary programming improvement, machine learning includes programming machines to show themselves from information as opposed to teaching them to play out specific assignments in certain ways. Machine learning is generally centered on forecast and making structure out of unstructured information.

PostgreSQL -PostgreSQL is an amazing, open source object-social database framework that utilizes and expands the SQL language joined with numerous highlights that securely store and scale the most confused information outstanding tasks at hand. PostgreSQL accompanies numerous highlights expected to enable engineers to manufacture applications, directors to secure information respectability and construct fault tolerant conditions,and help you deal with your information regardless of how huge or little the dataset. Although being free and open source, PostgreSQL is exceedingly extensible. For instance, you can characterize your very own information types, work out custom capacities, and even compose code from various programming dialects without recompiling your database! PostgreSQL attempts to accommodate with the SQL standard where such conformance does not repudiate conventional highlights or could prompt poor structural choices. Huge numbers of the highlights required by the SQL standard are upheld, however in some cases with marginally contrasting grammar or capacity.

Koa Client -Koa is another web structure planned by the group behind Express, which means to be a littler, progressively expressive, and increasingly establishment for web applications and APIs. By utilizing asynchronous capacities, Koa enables you to dump callbacks and enormously increment blunder dealing with. Koa does not package any middleware inside its center, and it gives a rich suite of techniques that make composing servers quick and pleasant. Here koa client is used with the postman tool inorder to show front end job.A Koa application is an article containing a variety of middleware capacities which are formed and executed in a stack-like way upon demand. Koa is like numerous other middleware frameworks that you may have experienced and anyway a key plan choice was made to give abnormal state at the generally low-level middleware layer. This improves interoperability, strength, and makes composing middleware considerably more pleasant. This incorporates strategies for regular undertakings like substance arrangement, reserve novelty, intermediary backing, and redirection among others. In spite of providing a sensibly extensive number of accommodating strategies Koa keeps up a little impression, as no middleware are packaged Hassan Nazeer and Waheed ,(2017).

Revised Manuscript Received on April 12, 2019.

Dr. Sunil Bhutada, IT Department Sreenidhi Institute of Science & Technology Yamnampet, Ghatkesar Hyderabad - 501 301,Telangana, India. (E-mail : sunilb@sreenidhi.edu.in).

Saylee Morey, IT Department Sreenidhi Institute of Science & Technology Yamnampet, Ghatkesar Hyderabad - 501 301,Telangana, India. (E-Mail : doremonsaylee@gmail.com)

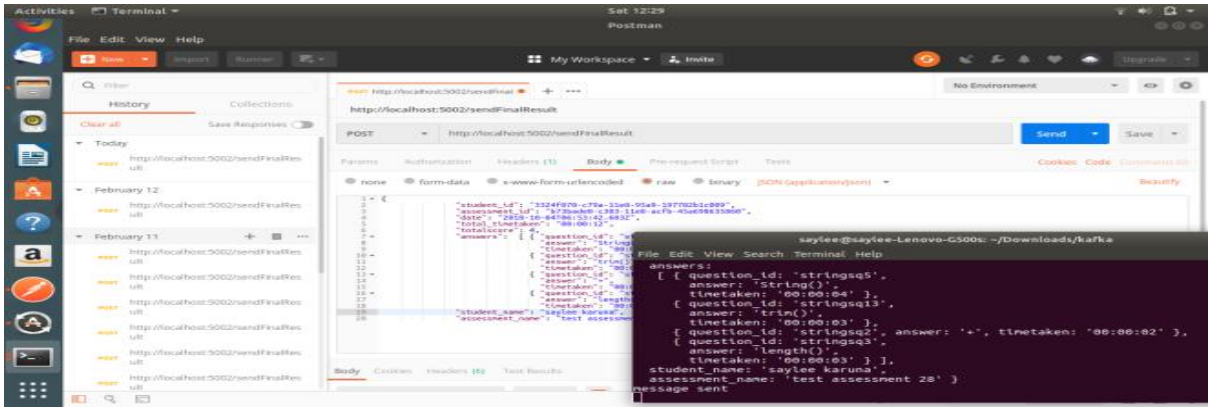

Fig1. postman with koa client

Docker -Docker is a computer program that achieve OS level visualization, It was first released in 2013.which is used to run container. Were the space will be saved with the use of it as it creates images which are present in the docker hub . the main advantage of the docker is it can be fetched on any kind of system in order to implement the setup.

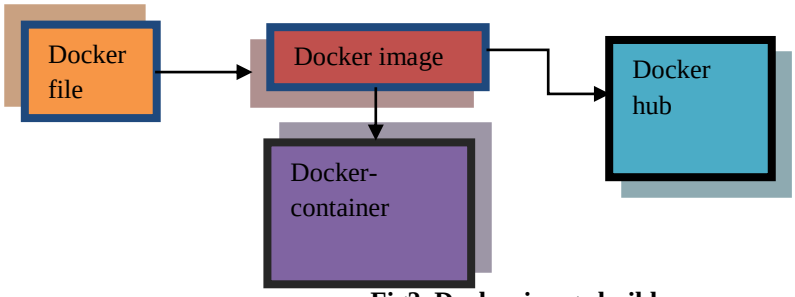

Fig2. Docker image builder

2. METHODS

Machine Learning Classifiers -ML is one the the most important part of Artificial Intelligence which is worried about advancement strategies, techniques and empower machine to gain knowledge. Basically terms are used to improve the calculations which empower the machine to study and executeresponsibilities and exercises. ML covers with insights from multiple points of view Irina Pak and Phoey Lee Teh(2016).

Decision Tree -A decision tree is one of the classifier corresponded with recursive division of occurrence space. DT consist of hubs that structure an conventional tree, which is a synchronized tree called "root" which has no advancelimits. Each hub have accurately one advancelimits. A hub with energetic limits is called an inner or leaves (otherwise called terminal or decision hubs). Adecision tree, has interior hub divisions example space has minimum to nodes as per a specific detachedcompetence of in rankvalue. In complex and everyday case, apiece test thinks about singleattribute with the end goal that the occasion space is divided by the property's estimation. On account of numeric attributes, the conditions allude to a range. leaves are appointed to the divisionlanguage to the suitable objectiverate. On the other hand, the leaves would hold a likelihood vector showing the likelihood of idea quality having a precise rate. Examples are ordered by exploring from the foundation of the tree behind a leaf whichhas specifyby the outcome of tests along the way.

Support Vector Machine (SVM) -Support vector machines (SVMs) are a lot of related administered learning techniques utilized for order and decline. They have a place with a group of sumup straight classifiers. In other terms, Support Vector Machine (SVM) is an order and relapse foretelltackle that utilizes machine learning hypothesis to expand prescient precision while consequently it keepsabsent from over-fit to the information. Support Vector machines can be differentiated as frameworks which use speculation space of direct capacities in a high dimensional component space, prepared with a taking in calculation from advancement hypothesis that actualizes a taking in leaning got from accurate learning hypothesis. Support vector machine was at first famous with the NIPS people group and now it ia a functioning piece of the machine learning research the world over. SVM is well known when, utilizing pixel maps as information; it gives precision equivalent to complex neural systems with expounded highlights in a skill acknowledgment undertaking S. Choudhury and A. Bhowal (2015); R Xiao , J. C Wang, Z X Sun(2002); C. D Guo, S. Z Li. Control (2003, 14); N. Chand et al.(2016).

k-NN Classifier -K-Nearest Neighbors (KNN) is one of the easiest calculations utilized in Machine Learning for degeneration and order issue. KNN calculations utilize information and characterize new information focuses dependent on a similitude measures (for example eliminateeffort). The information is allotted to the class which has the most closest neighbors. As you increment the quantity of closest neighbors, the estimation of k, precision may increment Jesus Maillo and Sergio (2016).

SGD – SGD stands for stochastic gradient descent is capable for tuning the factors of massively over-parametrized models to fetch small training loss with good generalization despite the existence of numerous bad minima. This is especially surprising given DNNs are capable of overfitting random data with almost zero training loss Zhang et al. (2016). This behavior has been studied by Arpit et al. (2017); Advani and Saxe (2017) where they suggest that deep networks generalize well because they tend to fit simple functions over training data before overfitting noise. It has been further discussed that model parameters that are in a region of flatter minima generalize better Hochreiter and Schmidhuber (1997); Keskar et al. (2016); Wu et al. (2017),and that SGD finds such minima when used with small batch size and large learning rate Keskar et al.(2016); Jastrz¸ebski et al. (2017); Smith et al. (2017); Chaudhari and Soatto (2017).

2.1. Size of Dataset

The data set has been collected from different colleges with the help of cri.ai we generated the assessment test to know the capabilities of the students in which categories do they fix. The total data set we have collected is 5000 which has been used to train the model with the help of machine learning and neural networks. The below figure shows the data set obtained.

Table 1 Number of Unique ID with Students name.

Assessment	Unique ID	Student name
BRCET	123	Karthik talam
BRCET	15321A0205	Anjali reddy
BRCET	15321A0206	Archana
BRCET	15321A0207	Ashwini
BRCET	15321A0209	Bhargavi

3. RESULTS AND DISCUSSION

3.1. Architecture of the Proposed System (data pipeline) - The proposed system developed by using Kafka (Messaging bus), PySpark (distributed processing published data) technologies. Kafka is frequently utilized progressively conversational information structures to give constant assessment. Since Kafka is a quick, versatile, powerful, and fault tolerant distribute obtain in informing framework, Kafka is utilized being used situations where JMS, RabbitMQ, and AMQP may not be considered because of volume and responsiveness. Kafka has higher throughput, unwavering quality, and replication attributes, which makes it appropriate for things likeIoT sensor information. Kafka has worked with Spark Streaming and Spark for constant ingesting, investigation and handling of leaking information. Kafka is an information stream used to nourish Hadoop BigData. Kafka merchants support horrible significance streams for high transcribe assessment in Spark.In this project, Kafka is used to process the data through the PySpark. PySpark is a language used for performing historical data analysis, building machine learning data pipelines, and also by creating ETLs for a data platform.By using the pyspark, we can transfer the data stream from the kafka and the pyspark can process the large data. It will form the data stream as (key, value) pair. Through this pyspark, we can easily process the large data. ElasticSearch is fantastic for indexing plus filtering data.So, in this project we are using the elasticsearch, we can analyze the capabilities of the people from the historical data. To perform the elasticsearch, we need to transform the data before indexing. Here, we apply the machine learning classifiers to search with high accuracy. We used Linear Discriminant Analysis (LDA) and it is a dimensionality reduction technique used as a preprocessing step in Machine Learning and pattern classification applications. The majoraim of dimensionality reduction techniques is to reduce the dimensions by eliminating the redundantplus dependent features by transforming the features from higher dimensional space to a space with lower dimensions. Kibana is used to visualize, search, and work together with data stored in Elasticsearch indices. We can also perform complex data analysis plus visualize historical data . Were the kibana works on the port of local host 5601.Here, Kibana can build the index for the processing data to search and view the data from the PostgreSQL.

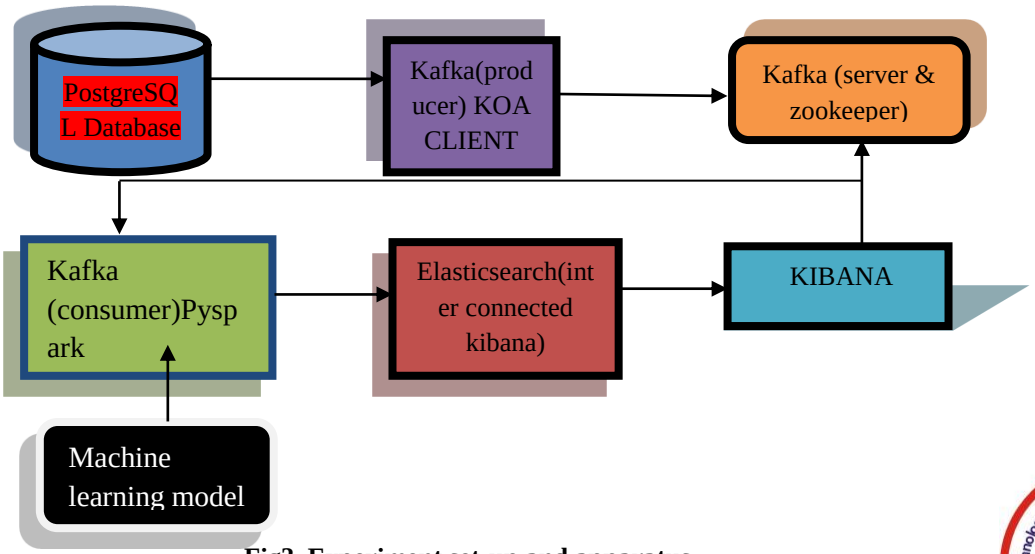

Fig3. Experiment set-up and apparatus

3.2 Experimental result

The above architecture has been used as data pipeline to flow data from the generating point to the visualization part. The data has been generated fron the online test assessment which has been conducted on the cri.ai these results as been stored in the postgre sql from where the data will be fetched by the koa client and passed through the data pipeline and the visualization part will be at kibana . the ML has been used which is an inbuilt part of an kibana with the help of premieum settings. And the algorithm used in ML is the SGD i.e stochastic gradient descent which supports the learning rate , momentum, decay and the nesterov. The type of model used in ML is linear regression with the bit concept of neural networks. The assessment view in graphical form is given as bellow.

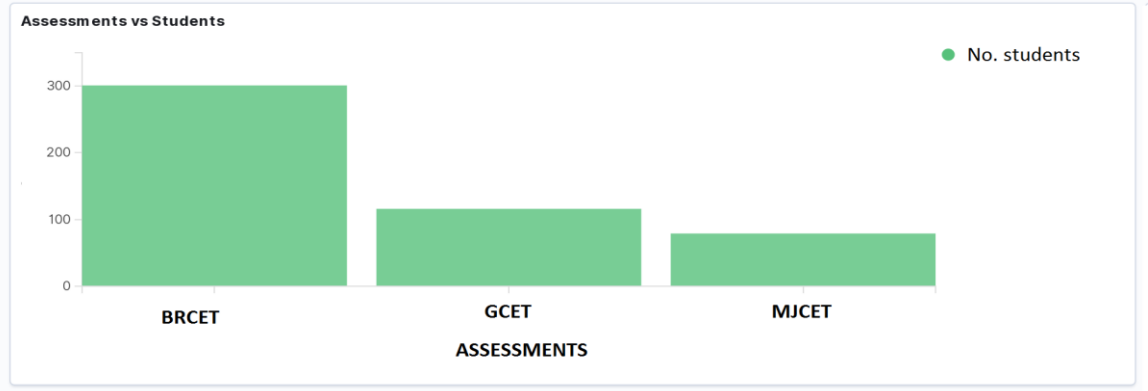

Fig4.(a) ass. Vs students

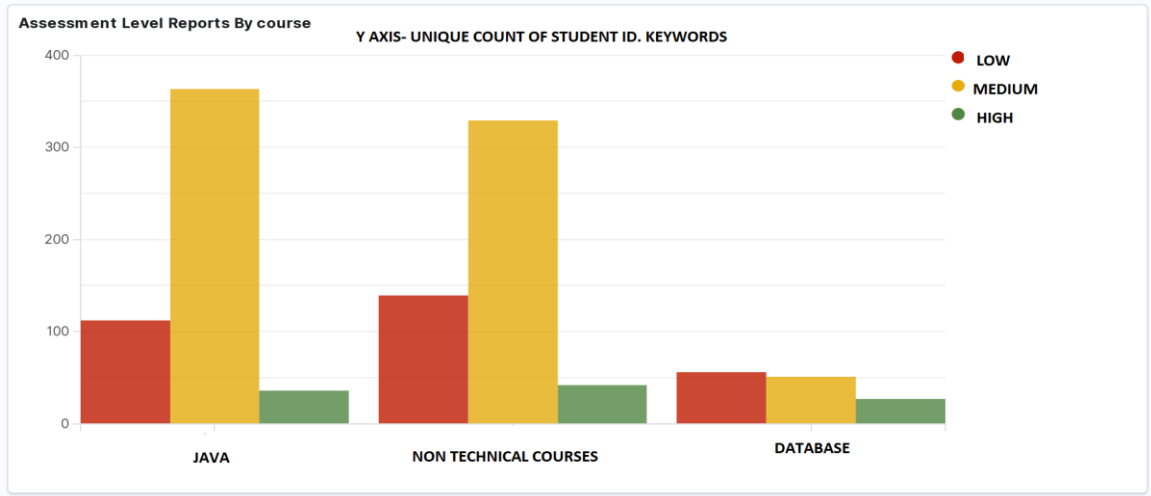

Fig4.(b) ass. Vs course reports

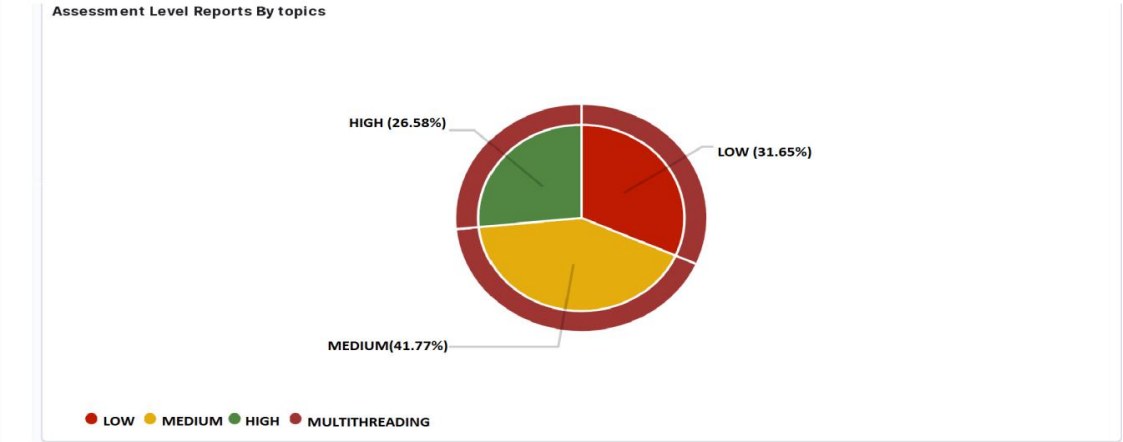

Fig4.(c) ass. Vs topics reports

The above result has been found out with the help of 5000 dataset which was generated from the various colleges by conducting assessment test for the students. Here the model has been trained with the help of machine learning and neural networks where the algorithm used here is the SGD as it gives more perfect result as compare to other algorithms. The training odf dataset has been done on the ratio of 8:2 , where the training dataset is of 80% and test dataset is of 20% the table is given as below.

Training data	Test data	Total data
80	20	100
100	50	150
150	50	200
200	100	300
-----	-------	---------------

Table 2. division of data in test and train and model train cycle.

In our experiment, we used Postgre SQL database to get data for the purpose of skill assessment. From the Postgre SQL the data is fetched by the koa client to the kafka were the zookeeper plays an vital role in implementing the data pipeline as it works on the port 2181, pyspark is used for creating key_value pair , elastic search is used which will be running on port 9200 and finally the visualization part will be on kibana which will show the result on the localhost port 5601.This entire data pipeline is stored in docker hub which can be pulled using docker on any system to make it as an set up. Docker is used to avoid system harmness which makes this connection to be in image form so that the memory required will be too less.

4. CONCLUSION

We conclude that in this paper, we implemented skill assessment program for the human capabilities. By using the proposed system, we can detect the people capabilities. This project is implemented by using the data pipeline which consist of kafka, pyspark , elastic search and kibana. Were the ML uses SGD to make system train which is the part of linear regression in supervised learning , we proved that the proposed system is very efficient to human capability screening as it has been trained from very basic till the end

5. ACKNOWLEDGEMENT

This project was done under my guide and also the intern company special thanks to both of them. the company has helped me a lot to understand the entire features and also the process, with the help of them I am able to complete my paper as well as project.

REFERENCES

1. Irina Pak and Phoey Lee Teh, December 2016 "Machine Learning Classifiers: Evaluation of the Performance in Online Reviews", Indian Journal of Science and Technology, Vol 9(45), DOI: 10.17485/ijst/2016/v9i45/100703.
2. R Xiao , J. C Wang, Z X Sun, 2002 ,. An approach to incremental SVM learning algorithm. Journal of Nanjing University (Natural Sciences) , 38 (2) :152-157.
3. C. D Guo, S. Z Li. Control 2003, 14 (1) , - based Audio Classification and Retrieval by Support Vector Machines. IEEE Trans. on Neural Network , pp.209-115
4. S. Choudhury and A. Bhowal , 2015 [3]Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection´, Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM).
5. N. Chand et al., 2016,[3]A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection´, Advances in Computing, Communication, & Automation (ICACCA).
6. Itay Safran and Ohad Shamir, 2016 On the quality of the initial basin in overspecified neural networks. InInternational Conference on Machine Learning, pages 774–782.
7. Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou, , 2017 Empirical analysis of thehessian of over-parametrized neural networks. arXiv preprint arXiv:1706.04454.
8. Karen Simonyan and Andrew Zisserman , 2014 Very deep convolutional networks for large-scale imagerecognition. arXiv preprint arXiv:1409.1556.
9. Leslie N Smith , 2017 Cyclical learning rates for training neural networks. In Applications of ComputerVision (WACV), 2017 IEEE Winter Conference on, pages 464–472. IEEE.
10. Leslie N Smith and Nicholay Topin, , 2017. Super-convergence: Very fast training of residual networksusing large learning rates. arXiv preprint arXiv:1708.07120
11. Samuel L Smith and Quoc V Le , 2017 Understanding generalization and stochastic gradient descent. arXivpreprint arXiv:1710.06451.
12. Jesus Maillo a , ∗, Sergio Ramírez a , Isaac Triguero (2016) Francisco Herrera a , b kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data
13. Hassan Nazeer, Waheed Iqbal, Fawaz Bokhari, 12 dec 2017 Real-time Text Analytics Pipeline Using Open-source Big Data Tools arXiv:1712.04344v1.