VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JnanaSangama, Belgaum-590014

Final Report
On

"Prediction of Regional Language Implementation"

Submitted in Partial fulfillment of the Requirements for the VIII Semester of the Degree of

Bachelor of Engineering
In
Computer Science & Engineering
By

KARTHIK A N (1CE18CS030)
R LAKSHMI SAI CHETANA NATH(1CE18CS061)
SURABHI G R (1CE18CS084)
UDANKA AARUNJAIN (1CE18CS090)

Under the Guidance of
Mrs. Nandini S B
Asst Prof, Dept. of CSE

CITY ENGINEERING COLLEGE

Doddakallasandra, Kanakapura Road,

Bengaluru-560061

CITY ENGINEERING COLLEGE

Doddakallasandra, Kanakapura Road, Bengaluru-560061

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the Project work entitled **"PREDICTION OF REGIONAL LANGUAGE IMPLEMENTATION"** has been carried out by **KARTHIK A N (1CE18CS030), R LAKSHMI SAI CHETANA NATH (1CE18CS061), SURABHI G R(1CE18CS084) and UDANKA AARUNJAIN (1CE18CS090)** bonafide students of City Engineering College in partial fulfilment for the award of **Bachelor of Engineering in Computer Science and Engineering** of the Visveshvaraya Technological University, Belgaum during the year **2021-2022**. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library. The Project phase Report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said Degree.

Mrs. Nandini S B Internal Guide Asst. Prof, Dept. Of CSE	**Dr. Sowmya Naik** Head, Dept. of CSE	**Dr. H. N Thippeswamy** Principal, CEC

External viva

Name of Examiner — Signature with Date

1.

2.

ABSTRACT

This project is to apply Data Science, Artificial Intelligence and Machine Learning. In order to get to know the cons and pros of the new implementation of regional language for professional degree course, a nationwide survey conducted to get the data and mindsets of the people. By the huge collection of data, the analysis will be done by running particular ai/ml-based algorithms which includes linear regressions, KNN, k means and random forest for clustering the similar outcomes.

Linear Regressions helps in finding out the relationship between variables and forecasting algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. The K-means algorithm based on the approximate backbone and the shuffled frog leaping algorithm is used for the obtained clustering results.

By processing the acquired data, we can find the accuracy of the government's implication on the professional course, and also can predict the knowledge growth in the same.

This allows the people to think on before opting for the course and also get a clear picture on how the mindset of the people.

ACKNOWLEDGEMENT

While presenting this Project on “Prediction of Regional Language Implementation”, we feel that it is our duty to acknowledge the help rendered to us by various persons.

Firstly, I take this opportunity to thank my college “**CITY ENGINEERING COLLEGE”** for providing all the resource required to success the project report. I would like to express our heartfelt gratitude to **Dr. H N Thippeswamy**, Principal, CEC Bangalore and **Dr. Sowmya Naik**, Head of Dept, Computer Science and Engineering for extending their support.

I am very grateful to our guide, **Mrs. Nandini S B,** Asst. Prof., Department of Computer Science and Engineering, for her able guidance and valuable advice at every stage of our project which helped me in the successful completion of my project. Her guidance and support were truly invaluable

We would also have indebted to our Parent and Friends for their continued moral and material support throughout the course of project and helping me in finalize the presentation. Our hearty thanks to all those have contributed bits, bytes and words to accomplish this Project.

KARTHIK A N (1CE18CS030)

R LAKSHMI SAI CHETANA NATH (1CE18CS061)

SURABHI G R (1CE18CS084)

UDANKA AARUNJAIN (1CE18CS090)

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to natural intelligence displayed by animals including humans. Leading AI textbooks define the field as the study of "intelligent agents": any system that perceives its environment and takes actions that maximize its chance of achieving its goals.

1.1 Artificial Intelligence (AI):

Some popular accounts use the term "Artificial Intelligence" to describe machines that mimic "Cognitive" functions that humans associate with the human mind, such as "Learning" and "Problem solving", however, this definition is rejected by major AI researchers.

AI applications include advanced web search engines (e.g., Google), recommendation systems (used by YouTube, Amazon and Netflix), understanding human speech (such as Siri and Alexa), self-driving cars (e.g., Tesla), automated decision-making and competing at the highest level in strategic game systems (such as chess and Go). As machines become increasingly capable, tasks considered to require "intelligence" are often removed from the definition of AI, a phenomenon known as the AI effect. For instance, optical character recognition is frequently excluded from things considered to be AI, having become a routine technology. Figure 1.1 shows the block diagram of AI

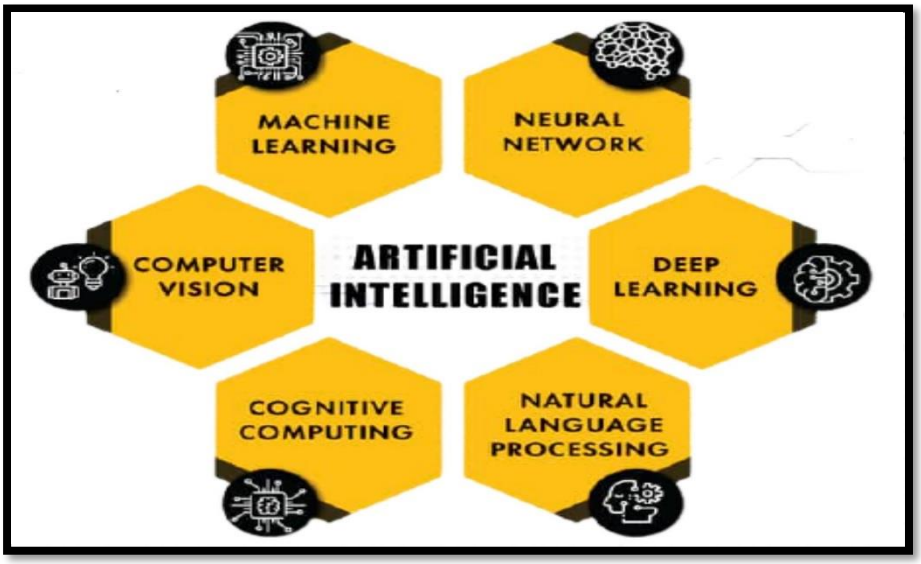

Figure 1.1: Block Diagram of Artificial Intelligence

1.2 Machine Learning (ML):

Machine learning (ML) is the study of computer algorithms as shown in Figure1.2 that can improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Some implementations of machine learning use data and neural networks in a way that mimics the working of a biological brain. In its application across business problems, machine learning is also referred to as predictive analytics.

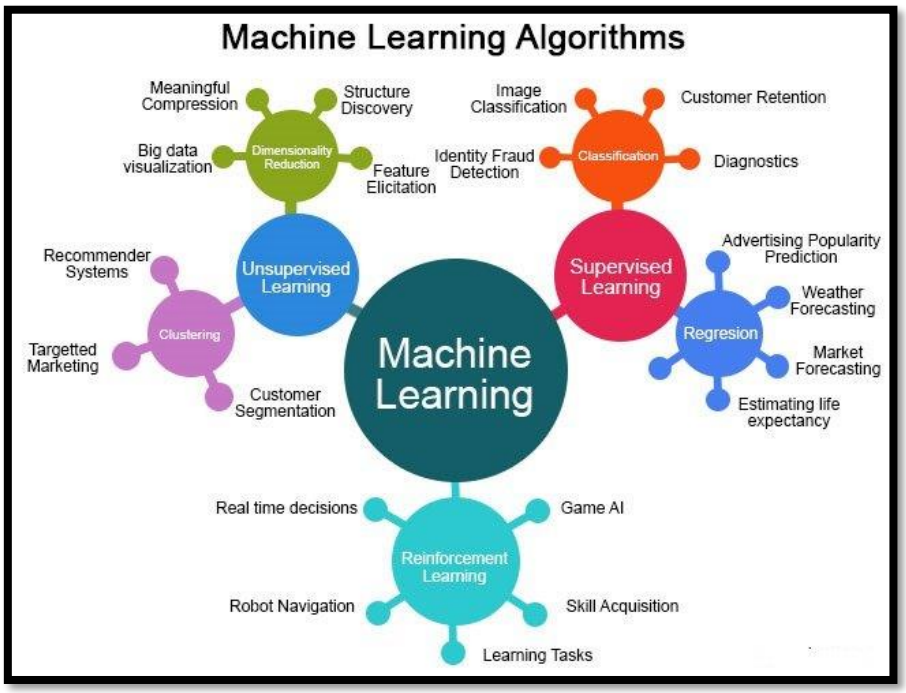

Figure 1.2: Machine Learning Algorithm

1.3 Data Science:

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains as shown in Figure 1.3. Data science is related to data mining, machine learning and big data.

Data science is a "concept to unify statistics, data analysis, informatics, and their related methods" in order to "understand and analyze actual phenomena" with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge. However, data science is different from computer science and information science.

A data scientist is someone who creates programming code, and combines it with statistical knowledge to create insights from data. Data science is an interdisciplinary field focused on extracting knowledge from data sets, which are typically large (see big data), and applying the knowledge and actionable insights from data to solve problems in a wide range of application domains. The field encompasses preparing data for analysis, formulating data science problems, analyzing data, developing data-driven solutions, and presenting findings to inform high-level decisions in a broad range of application domains.

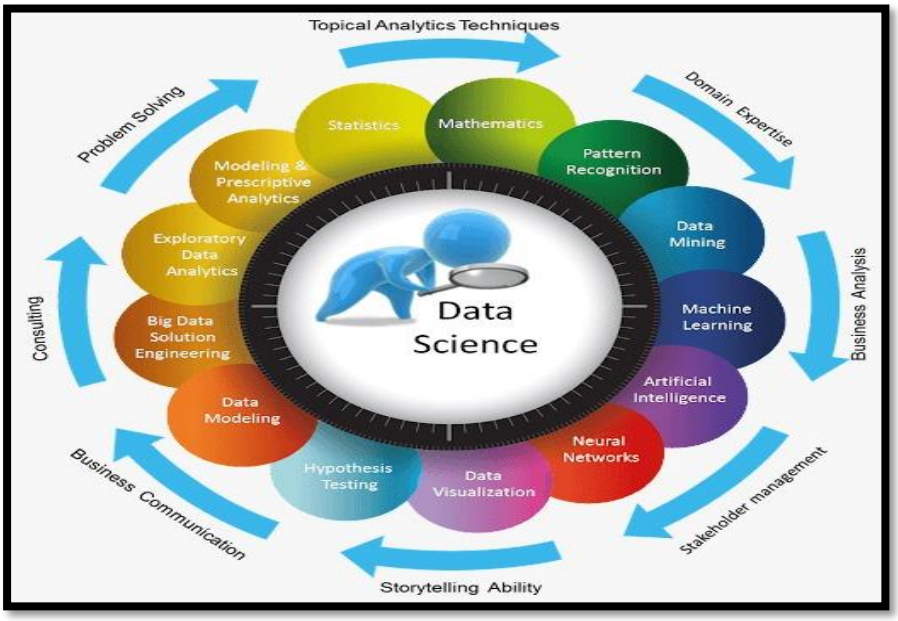

Figure 1.3: Data Mining, Machine Learning and Big Data

1.4 ALGORITHMS USED

These are the following algorithms used in this project:

- Linear Regression Algorithm
- K-Means Algorithm
- KNN Algorithm
- Random Forest Algorithm

1.4.1 LINEAR REGRESSION ALGORITHM

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated. Linear regression is one of the easiest and most popular Machine Learning algorithms as shown in Figure1.4. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

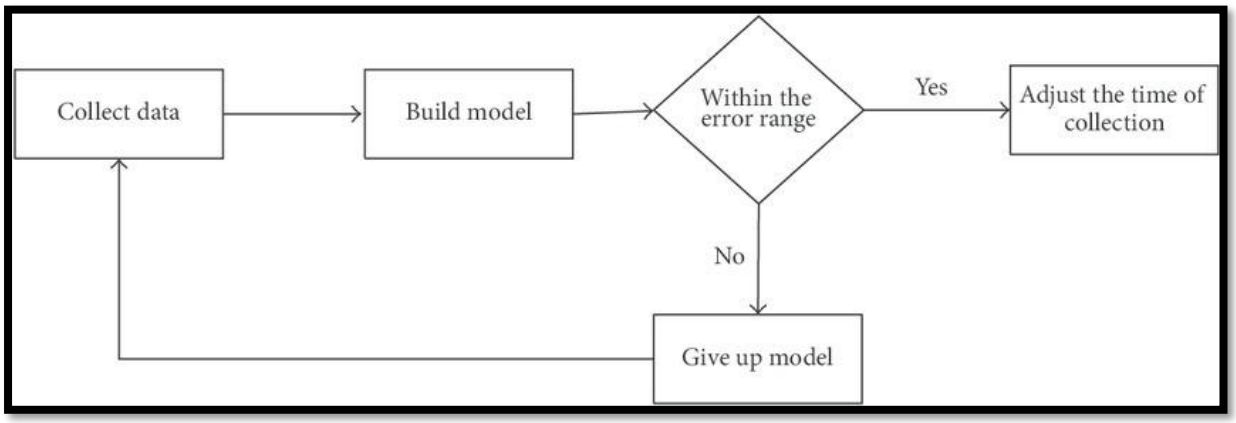

Figure 1.4: Linear Regression Algorithm

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. The linear regression model provides a sloped straight line representing the relationship between the variable

1.4.2 K-MEANS ALGORITHM

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training. It is a centroid-based algorithm as shown in Figure1.5, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

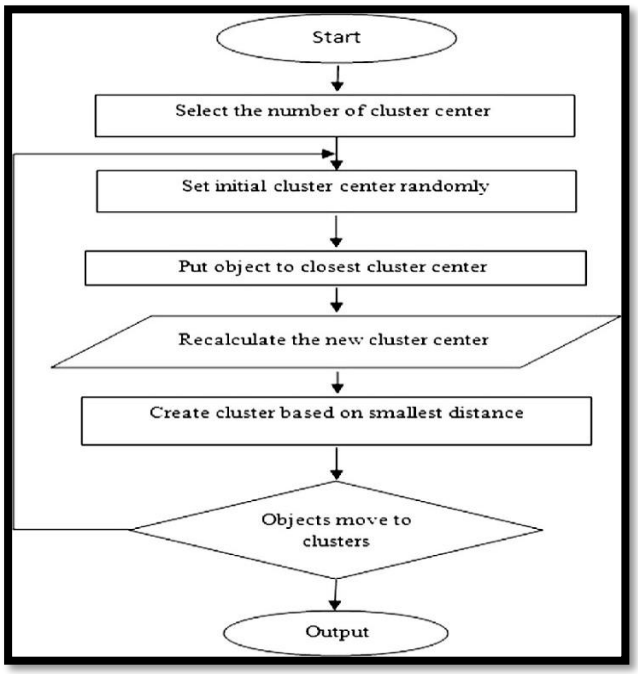

Figure 1.5: K-Means Algorithm

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

1.4.3 KNN ALGORITHM

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. This algorithm as shown in Figure1.6 assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

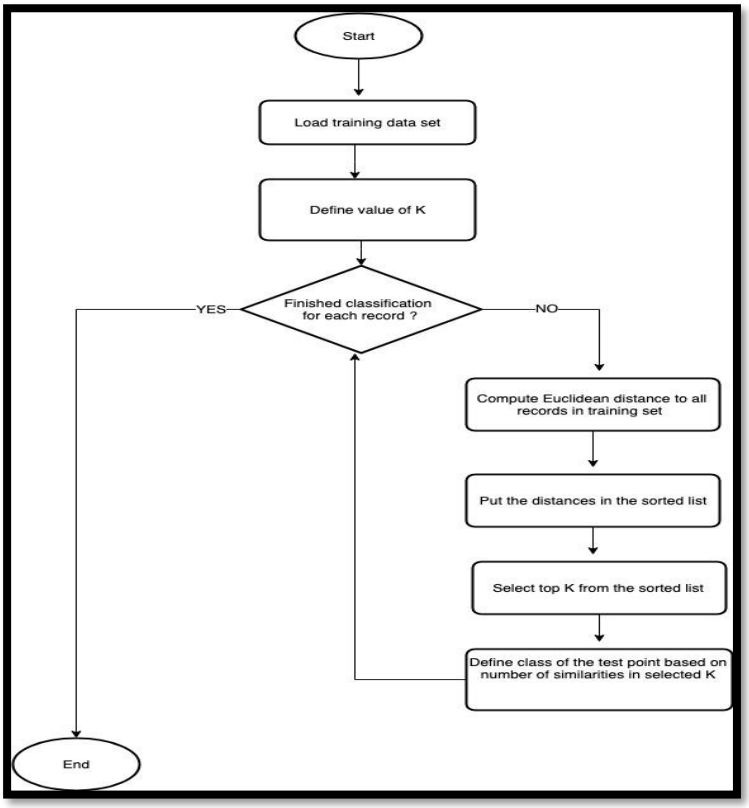

Figure 1.6: KNN Algorithm

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a “non-parametric algorithm”, which means it does not make any assumption on underlying data. It is also called a “lazy learner algorithm” because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at

the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

1.4.4 RANDOM FOREST ALGORITHM

Random Forest is a popular machine learning algorithm as shown in Figure1.7 that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

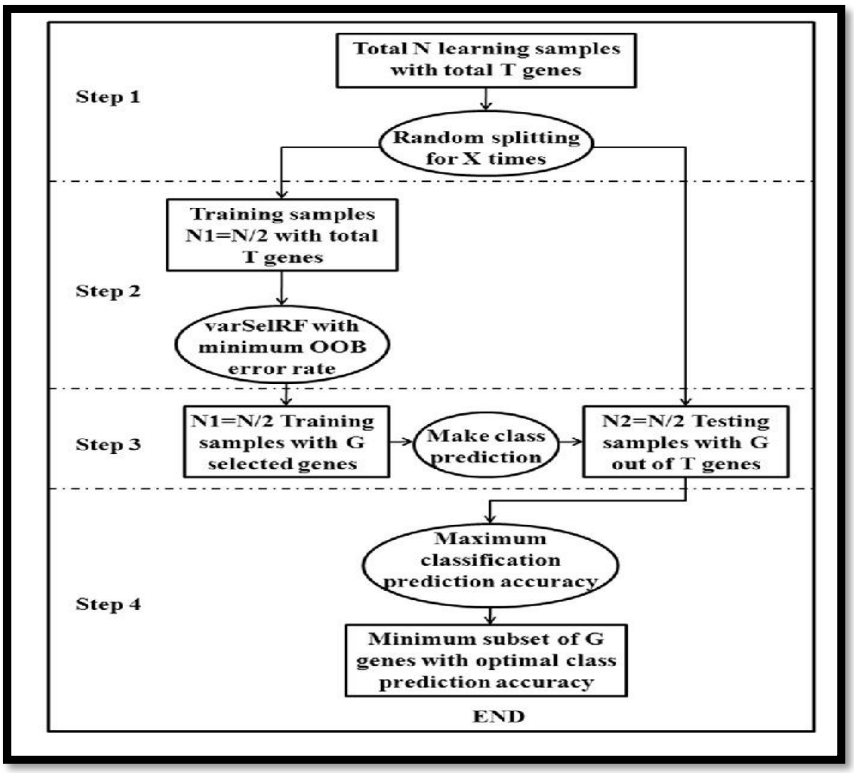

Figure 1.7: Random Forest Algorithm

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Chapter 2

LITERATURE SURVEY

[1] Analysis of computer science based on big data mining

Authors: Liu Xuan, Liu Chang

Publication Year: 2020

Publication: Huazhong University of science and Technology library, Wuhan

Description:

The scientific construction of a first-class discipline construction evaluation system is of great significance to the promotion of discipline construction. As an important evaluation reference system, the third-party evaluation system must pay attention to its underlying data sources and calculation methods. Using the massive underlying data of the Scopus database, through the analysis of the computer disciplines of four Chinese universities, the development trend is discussed from the aspects of overall academic output, scientific research quality, and hot topics.

Advantages:

- Marketing / Retail: Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have an appropriate approach to selling profitable products to targeted customers.
- Finance / Banking: Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer data, the bank, and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect the card's owner.
- Manufacturing: By applying data mining in operational engineering data, manufacturers can detect faulty equipment and determine optimal control parameters. For example, semiconductor manufacturers have a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are a lot the same and some for unknown reasons even has defects.
- Government: Data mining helps government agencies by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activities.

Disadvantages:

- Privacy Issues: The concerns about personal privacy have been increasing enormously recently especially when the internet is booming with social networks, e-commerce, forums, blogs. Because of privacy issues, people are afraid of their personal information is collected and used in an unethical way that potentially causing them a lot of trouble. Businesses collect information about their customers in many ways for understanding their purchasing behavior trends
- Security Issues: Security is a big issue. Businesses own information about their employees and customers including social security numbers, birthdays, payroll and etc. However, how properly this information is taken care of is still in question. There have been a lot of cases that hackers accessed and stole big data of customers from a big corporation.
- Misuse of Information/inaccurate information: Information is collected through data mining intended for ethical purposes can be misused. This information may be exploited by unethical people or businesses to take the benefits of vulnerable people or discriminate against a group of people.

[2] An overview on machine learning technologies and their use in E-learning

Authors: Ramzi Farhat, Yosra Mourli, Mohamed Jemni, Houchine Ezzedine

Publication Year: 2020

Publication: Latice research laboratory university Tunis, Tunisia, University Polytechnique Hauts-de-France, Valencinnes , France

Description:

Thanks to new technologies, internet, connected objects we produce a phenomenal amount of data. Putting these data in context, organizing them to be able to perceive, understand and reflect them is very important. Traditionally, human have analyzed data. However, as the volume of data surpasses, human increasingly turn to automated systems that can imitate him. Those systems able to learn from both data and changes in data in order to solve problems are called machine learning. Artificial intelligence has a major impact on e-learning research and

the machine learning based methods can be implemented to improve Technology Enhanced Learning Environments (TELE). This paper is an overview of the recent findings in this research field. At first, we introduce the key concepts related to machine learning. Then, we present some recent works using machine learning in e-learning context.

Advantages:

- Automation of Everything: Machine Learning is responsible for cutting the workload and time. By automating things, we let the algorithm do the hard work for us. Automation is now being done almost everywhere. The reason is that it is very reliable. Also, it helps us to think more creatively.

- Wide Range of Application: ML has a wide variety of applications. This means that we can apply ML on any of the major fields. ML has its role everywhere from medical, business, banking to science and tech. This helps to create more opportunities. It plays a major role in customer interactions.

- Scope of Improvement: Machine Learning is the type of technology that keeps on evolving. There is a lot of scope in ML to become the top technology in the future. The reason is, it has a lot of research areas in it. This helps us to improve both hardware and software.

- Efficient Handling of Data: Machine Learning has many factors that make it reliable. One of them is data handling. ML plays the biggest role when it comes to data at this time. It can handle any type of data

- Best for Education and Online Shopping: ML would be the best tool for education in the future. It provides very creative techniques to help students study. Recently a school has started to use ML to improve student focus. In online shopping, the ML model studies your searches. Based on your search history, it would provide advertisements. These will be about your search preferences in previous searches

Disadvantages:

- Possible of High Errors: In ML, we can choose the algorithms based on accurate results. For that, we have to run the results on every algorithm. The main problem occurs in the training and testing of data. The data is huge, so sometimes removing errors becomes early impossible. These errors can cause a headache to users. Since the data is huge,
- Algorithm Selection: The selection of an algorithm in Machine Learning is still a manual job. We have to run and test our data in all the algorithms. After that only we can decide what algorithm we want. We choose them on the basis of result accuracy. The process is very much time-consuming.
- Data Accusation: In ML, we constantly work on data. We take a huge amount of data for training and testing. This process can sometimes cause data inconsistency. The reason is some data constantly keep on updating. So, we have to wait for the new data to arrive. If not, the old and new data might give different results. That is not a good sign for an algorithm.
- Time and Space: Many ML algorithms might take more time than you think. Even if it's the best algorithm it might sometimes surprise you. If your data is large and advanced, the system will take time. This may sometimes cause the consumption of more CPU power. Even with GPUs alongside, it sometimes becomes hectic. Also, the data might use more than the allotted space.

[3] Anomaly Detection by Using Streaming K-Means and Batch K-Means

Authors: Zhuo Wang, Yanghui Zhou, Gangmin Li

Publication Year: 2020

Publication: 2020 5th IEEE International Conference on Big Data Analytics

Description:

This paper introduces K-Means algorithm as new technique for detecting anomaly. Data analysis has been applied to industry field widely and plays important role in it. However,

conventional data analysis method cannot process large-scale data in considerable time and waste lots of computing resources. Conversely, Batch processing and Stream processing are equipped with property of processing data in short time interval, especially stream processing, can process data in real-time. This paper also compares Batch K-Means processing with Streaming K-Means processing according to distance, cost value and cluster distribution factors. Moreover, this paper also discusses how to reach optimized K value of Batch K-means model and Streaming K-means model, analyzes attributes of Batch K-Means processing and Streaming K-Means processing and finds limitations of these two processing models. Finally, the paper proposes limitations of research experiment and future improvement of clustering technique.

Advantages:

- Relatively simple to implement.
- Scales to large data sets.
- Guarantees convergence.
- Can warm-start the positions of centroids.
- Easily adapts to new examples.
- Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

Disadvantages:

- Choosing k manually.
- Being dependent on initial values.
- Clustering data of varying sizes and density.
- Clustering outliers.
- Scaling with number of dimensions.

[4] Improved random forest classification approach based on hybrid clustering selection

Authors: Dong Yuan, Jian Huang, Xu Yang, Jiarui Cui

Publication Year: 2020

Publication: 2020 Chinese Automation Congress

Description:

The random forest algorithm is an ensemble learning method, with the decision tree as its base classifier. In the ensemble model, it is not always true that the more the base classifiers, the better the classification effect, since if there are more base classifiers with poor performance in the model, they may have negative impacts on the final classification result. In order to modify the random forest classification method under the premise of ensuring the diversity of the random forest model, based on the random forest algorithm of cluster integration selection and personal indoor thermal preference model, this paper proposes a random forest method of clustering ensemble selection with Dunn index. Considering the shortcomings of the irrevocable merging strategy of hierarchical clustering algorithm, a random forest method of hybrid clustering ensemble selection based on hierarchical clustering and kmedoids partition clustering is developed. The effectiveness of the proposed methods is verified by classifying personal indoor thermal preferences.

Advantages:

- No feature scaling required: No feature scaling (standardization and normalization) required in case of Random Forest as it uses rule-based approach instead of distance calculation.
- Handles non-linear parameters efficiently: Nonlinear parameters don't affect the performance of a Random Forest unlike curve-based algorithms. So, if there is high non-linearity between the independent variables, Random Forest may outperform as compared to other curve-based algorithms.
- Random Forest can automatically handle missing values.
- Random Forest is usually robust to outliers and can handle them automatically
- Random Forest algorithm is very stable. Even if a new data point is introduced in the dataset, the overall algorithm is not affected much since the new data may impact one tree, but it is very hard for it to impact all the trees.
- Random Forest is comparatively less impacted by noise

Disadvantages:

- It requires much computational power as well as resources as it builds numerous trees to combine their outputs.
- It also requires much time for training as it combines a lot of decision trees to determine the class
- Due to the ensemble of decision trees, it also suffers interpretability and fails to determine the significance of each variable.

[5] Approach to Determining the Boundaries of the Search Range for the Number of Trees in the Random Forest Algorithm

Authors: Liliya Demidova, **Maria Ivkina**

Publication Year: 2020

Publication: 2020 9th MEDITERRANEAN CONFERENCE ON EMBEDDED COMPUTING, BUDVA, MONTENEGRO

Description:

The problem of determining the search ranges for the optimal values of the trees' number for the random forest (RF) in order to reduce the time spent on its development has been considered. The aim of the work is to obtain the formulas for determining the search range for the number of trees. The formulas have been obtained based on the results of the experimental studies on the development of the RF classifiers based on various datasets from the machine learning data repositories. The results of experimental studies on the development of the RF classifiers using the training and test sets based on the analyzed datasets have been presented. The formulas for the graphical dependencies have been obtained in the general form for assessing the classification quality at the test set and the development time.

Advantages:

- Random Forest algorithm is less prone to overfitting than Decision Tree and other algorithms.
- Random Forest algorithm outputs the importance of features which is a very useful.
- Random Forest can be used to solve both classification as well as regression problems.
- Random Forest works well with both categorical and continuous variables.
- Random Forest can automatically handle missing values.

Disadvantages:

- Random Forest algorithm may change considerably by a small change in the data.
- Random Forest algorithm computations may go far more complex compared to other algorithms.
- Complexity: Random Forest creates a lot of trees (unlike only one tree in case of decision tree) and combines their outputs. By default, it creates 100 trees in Python sklearn library. To do so, this algorithm requires much more computational power and resources. On the other hand, decision tree is simple and does not require so much computational resources.
- Longer Training Period: Random Forest require much more time to train as compared to decision trees as it generates a lot of trees (instead of one tree in case of decision tree) and makes decision on the majority of votes.

[6] Optimization of regression algorithms using learning curve in wsn

Authors: Vivek Kumar Verma, Vinod Kumar

Publication Year: 2021

Publication: ABES engineering college Ghaziabad, UP, India and SRM institute of science and technology

Description:

Regression algorithm used for the prediction of output with given features and it is a supervised learning algorithm. In applying regression algorithms such as linear regression, Regression using ANN, Regression using deep learning there comes a problem of High bias or under fitting due to simple model or High Variance due to complex model. High variance problems occur when we take high order polynomial in linear regression, more hidden layers in ANN, or Deep Learning. So, to find which types of problems occur i.e. High bias or variance, one has to use the learning curve to find an acceptable error. This paper will be helpful for those researchers who are working in the regression algorithm and have lesser knowledge in the mathematical treatment of the regression algorithm. In this, simulation has done with the GNU octave simulator and verified the concept of the learning curve.

Advantages:

- Linear Regression performs well when the dataset is linearly separable. We can use it to find the nature of the relationship among the variables.
- Linear Regression is easier to implement, interpret and very efficient to train.
- Linear Regression is prone to over-fitting but it can be easily avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques and cross-validation.

Disadvantages:

- Main limitation of Linear Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. It assumes that there is a straight-line relationship between the dependent and independent variables which is incorrect many times.
- Prone to noise and overfitting: If the number of observations are lesser than the number of features, Linear Regression should not be used, otherwise it may lead to overfit because is starts considering noise in this scenario while building the model.
- Prone to outliers: Linear regression is very sensitive to outliers (anomalies). So, outliers should be analyzed and removed before applying Linear Regression to the dataset.
- Prone to multicollinearity: Before applying Linear regression, multicollinearity should be removed (using dimensionality reduction techniques) because it assumes that there is no relationship among independent variables. of one tree in case of decision tree) and makes decision on the majority of votes.

Chapter 3

PROBLEM STATEMENT

According to the recent implementation by AICTE and National Education Policy (NEP) which was launched by the government of India which said that Professional Degree Course can also be taken in the regional languages, which will be reflected for the academic year 2022-2023. Whether the AICTE's Regional Language Implementation will be knowledge one or not?

3.1 EXISTING METHOD

Till 2021 the professional Degree course was allowed to take up only in the medium of English. From the year 1875 the following professional degree course saw its foundation in India that time the course was allowed only in English. Till the following time following rule or implementation was in charge.

DISADVANTAGES:

- People who studied their intermediates and their schooling in the following regional languages makes it difficult for them to understand things and process the complicated English terms.
- Existing system it's tough for rural students get adapted to English language hence their processing speed would be less.

3.2 PROPOSED SYSYTEM

Offering higher education in regional languages has been the news of the hour. The newest implementation has been in the news for the academic year 2022-2023. So, the following survey is done in order get the mindset of the people and the data is then processed by particular algorithms which include LINEAR REGRESSION, KNN, K-means, Random Forest by this we can predict, analyze and get the accuracy of the following implications and the policy.

As its the new implementation now using the above algorithms we will assume if it is going be knowledgeable or not. The above algorithms are used to collect the data, store them and analysis the data collected to check if it is going be useful in future. Using this prediction we can come to a conclusion whether the imposed rule is useful for the Professional degree students or not.

ADVANTAGES:

- Can get to know the concepts better when done in the regional language, because this connects better when thought in the regional language, the algorithms give out the spot-on analysis, it tells the accuracy
- Student friendly.
- As mother tongue connects them very well, it would be easier for them to learn things

Chapter 4

SYSTEM REQUIRMENT SPECIFICATION

System requirement specification is a description of a software system to be developed. It lays out functional and non-functional requirements, and may include a set of use case that describes user implementations that the software must provide.

4.1 HARDWARE REQUIREMENTS

These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements minimum and recommended.

Processor

A Processor is the logic circuitry that responds to and processes the basic instructions that drive a computer. The four primary functions of a processor are fetch, decode, execute and write back. Intel i3 1.9GHz processor

RAM

RAM (random access memory) is the place in a computing device where the operating system (OS), application programs and data in current use are kept so they can be quickly reached by the device's processor. RAM is much faster to read from and write to than other kinds of storage in a computer, such as a hard disk drive (HDD), solid-state drive (SSD) or optical drive. Data remains in RAM as long as the computer is running. When the computer is turned off, RAM loses its data. When the computer is turned on again, the OS and other files are once again loaded into RAM, usually from an HDD or SSD.RAM 4GB.

Hard Disk

hard disk is part of a unit, often called a disk drive, hard drive, or hard disk drive, that stores and provides relatively quick access to large amounts of data on an electromagnetically charged surface or set of surfaces. Today's computers typically come with a hard disk that contains several billion bytes (gigabytes) of storage.500GB hard disk.

- System : Intel i3 1.9Ghz

- Hard Disk : 500 GB
- Ram : 4 GB
- Any desktop/ Laptop system with above configuration or higher level

4.2 SOFTWARE REQUIREMENTS

Software Requirements is a field within software engineering that deals with establishing the needs of stakeholders that are to be solved by software. The IEEE Standard Glossary of Software Engineering Terminology defines a requirement. A condition or capability needed by a user to solve a problem or achieve an objective. A condition or capability that must be met or possessed by a system or system component to satisfy a contract, standard, specification, or other formally imposed document. A document representation of a condition or capability**.**

Windows

A window is a separate viewing area on a computer display screen in a system that allows multiple viewing areas as part of a graphical user interface (GUI. Windows are managed by a windows manager as part of a windowing system. A window can usually be resized by the user. On today's multitasking operating systems, you can have a number of windows on your screen at the same time, interacting with each whenever you choose. The window first came into general use as part of the Apple Macintosh. Later, Microsoft made the idea the foundation of its Windows operating system. The X Window System was developed as an open cross platform windowing system for use in networks. It allows a client application in one computer to request windowing services at a user's workstation computer. Front end technologies are Windows builder (swings and applets).

Python

Python is a high-level, interpreted, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library

Database

A database is a collection of information that is organized so that it can be easily accessed, managed and updated. Data is organized into rows, columns and tables, and it is indexed to make it easier to find relevant information. Data gets updated, expanded and deleted as new information is added. Databases process workloads to create and update themselves, querying the data they contain and running applications against it.

Tool

An item or implement used for a specific purpose. A tool can be a physical object such as mechanical tools including saws and hammers or a technical object such as a web authoring tool or software program. Furthermore, a concept can also be considered a tool.

- Operating System : Windows 7
- Coding language : Python
- IDE : Anaconda Jupiter
- Database : Microsoft excel

Chapter 5

HIGH LEVEL DESIGN

System analysis and design is a methodology applies in computer system to develop a new system or to enhance given system which can solve a given problem. System analysis refers to understanding the present system, future requirements and defining solutions to meet the user's requirements within constraints and time frame

5.1 SYSTEM ARCHITECTURE

Machine learning architecture as shown in Figure 5.1 majorly has datastore and machine learning engine these engines are responsible for the following calculations, transformation, and giving a particular output. This architecture shows in manner which the dataflows in order to provide the output for the following huge amount of data in order to provide the exact outputs that satisfies the system.

The next major part is the performance tuning here it checks for the errors and make the process a bit eases out to give the nearest proximities. The data from the datastore is based on the survey that's been conducted by the developer(us). Then the datastore is being sent to the data transformation here the data will be cleaned and clustering of the data happens, this provides us the clean format of data by removing the missing data in the list. After the data Transformation the data will be sent to the feature generation and hence the output will be given in the as the model output.

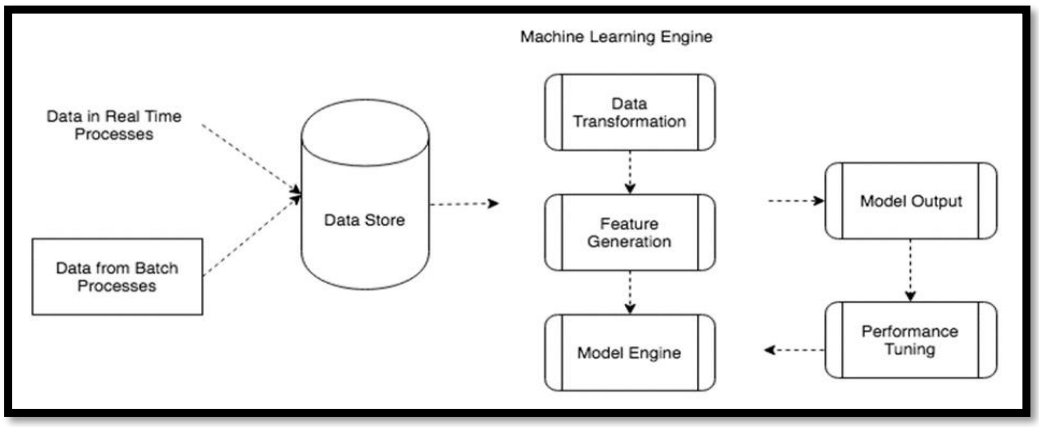

Figure 5.1: System Architecture

Chapter 6

DETAIL DESIGN

Detailed design is the phase where the design is refined and plans, specifications and estimates are created. Detailed design will include outputs such as 2D and 3D models, P & ID's, cost build up estimates, procurement plans etc. This phase is where the full cost of the project is identified.

6.1 Flow Charts of Algorithms

A flowchart is a type of diagram that represents a workflow or process. A flowchart can also be defined as a diagrammatic representation of an algorithm, a step-by-step approach to solving a task

LINEAR REGRESSION ALGORITHM

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated. Linear regression is one of the easiest and most popular Machine Learning algorithms as shown in Figure 6.1. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

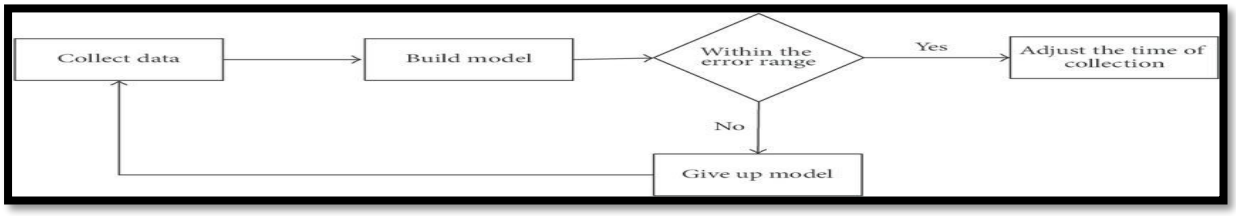

Figure 6.1: Linear Regression Algorithm Flow Chart

K-MEANS ALGORITHM

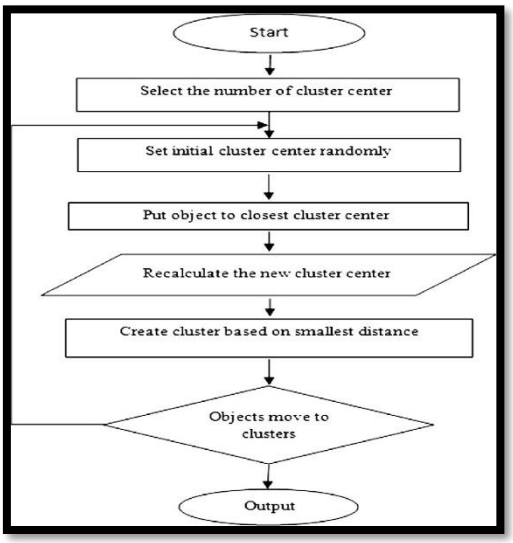

Figure 6.2: K-Means Algorithm Flow Chart

K-Means Clustering is an Unsupervised Learning algorithm as shown in Figure 6.2, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

KNN ALGORITHM

K-Nearest Neighbour is one of the simplest Machine Learning algorithm as shown in Figure 6.3 based on Supervised Learning technique. This algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

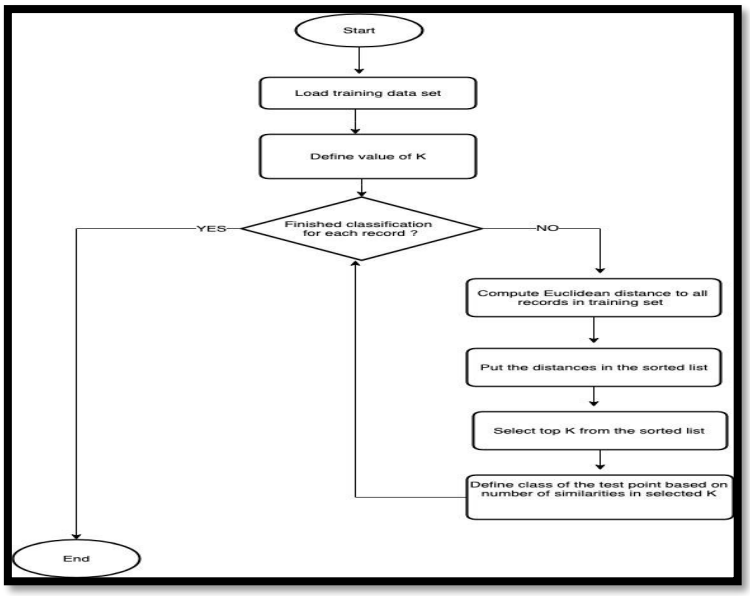

Figure 6.3: KNN Algorithm Flow Chart

RANDOM FOREST ALGORITHM

Random Forest is a popular machine learning algorithm as shown in Figure 6.4 that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

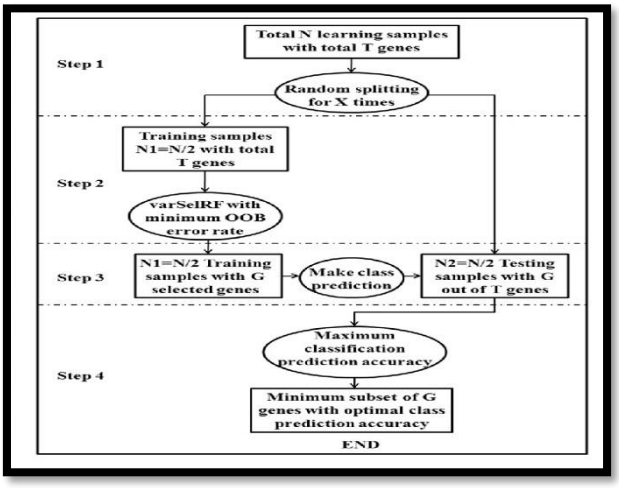

Figure 6.4: Random Forest Algorithm Flow Chart

6.2 Survey Module

The survey module plays the major part in the following project and is responsible for getting information that's required for the predictions to happen. Survey module is responsible for the following activities:

- **Collection of Information:** As the survey name tells, here the information will be collection by the set of questionnaires that's been made by the developers for plotting up the required predictions.
- **Creation of the dataset:** After getting the information, in this part of module the data created by the survey will be stored into the dataset as .CSV file extension.

6.2.1 Sequence Diagram

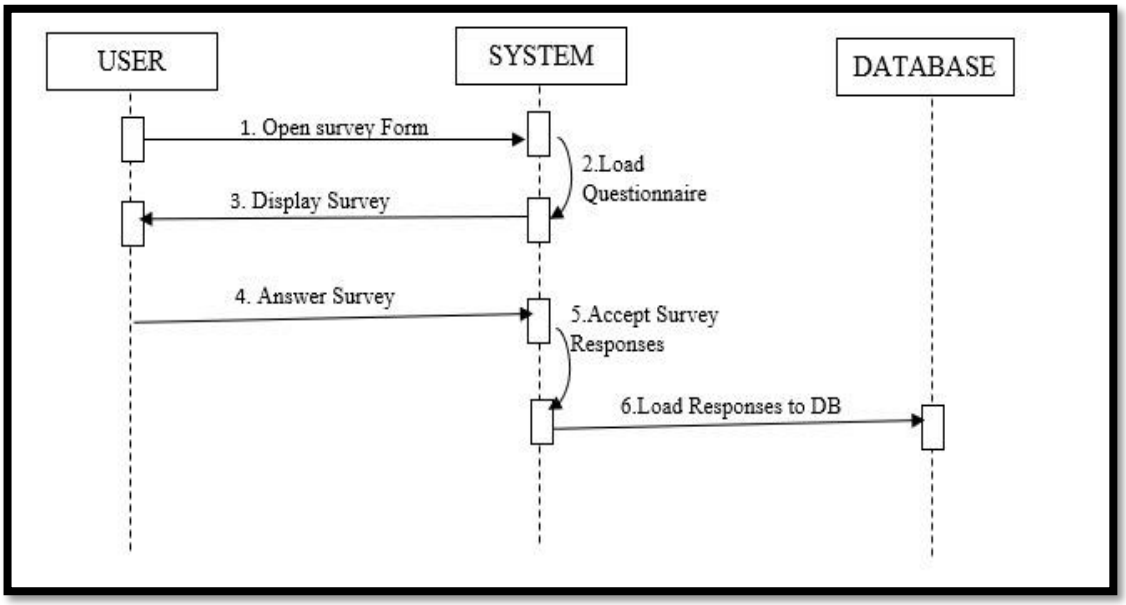

Figure 6.5: Sequence Diagram for Survey

The working of the entire process of the survey is as follows. As shown in Figure 6.5 the user will have to open the survey form. The questionnaires will get loaded. After the questionnaires are loaded the survey is displayed to the user. The user can start filling out the survey form. Once the user fills the survey form of all his answers for the entire questions the survey responses will be accepted. All the responses will get saved to the database. The entire information i.e. all the responses collected from the people will be saved in the database.

6.2.2 Data Flow Diagram

The dataflow diagram shown in figure 6.6 Which is for the survey module, here the developer firstly forms the question which is the first step that is formation of questions as shown in the diagram. Then step two is the conduction of the survey which is shown in the diagram as step 2. once the survey is formed the public firstly opens the survey and attend questionnaires uploaded by the developer, these are the step 3 and 4 as shown in the diagram. Then the step 5 is that the information from the public are been collected by the developer as shown in the above figure. Finally, once the public responds to the survey, the datasets are formed which is the final and last step in this process as shown as step 6 in the diagram

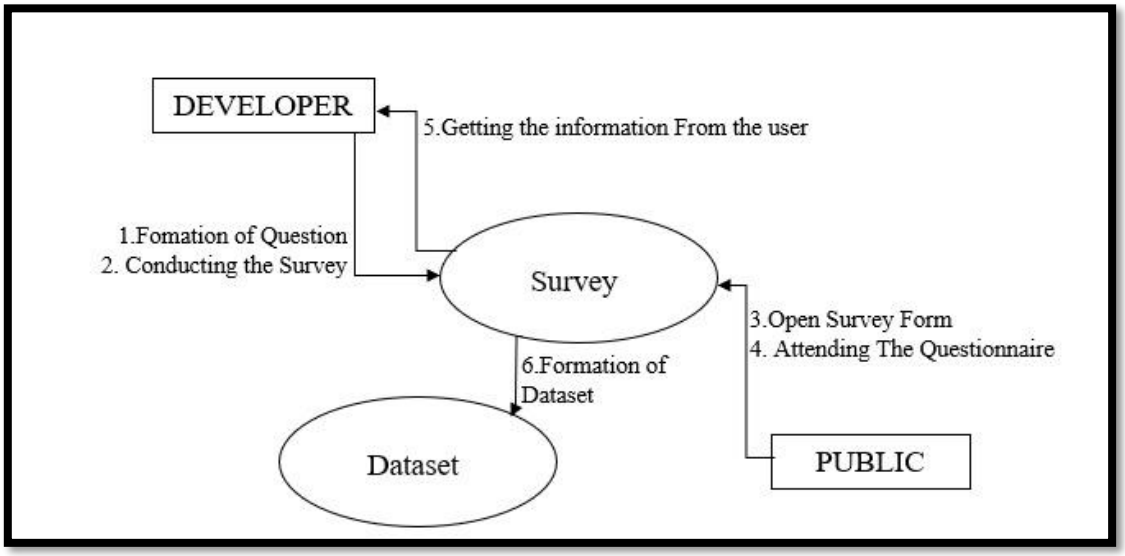

Figure 6.6: Data flow Diagram for Survey

6.2.3 Activity Diagram

The developer will have to prepare the survey form and this survey form will be circulated among the people. Once the survey form is circulated the user will have to open the survey form and start filling out the form. Once the survey form is completely filled the responses are accepted. By this the dataset is prepared.as shown in Figure 6.7

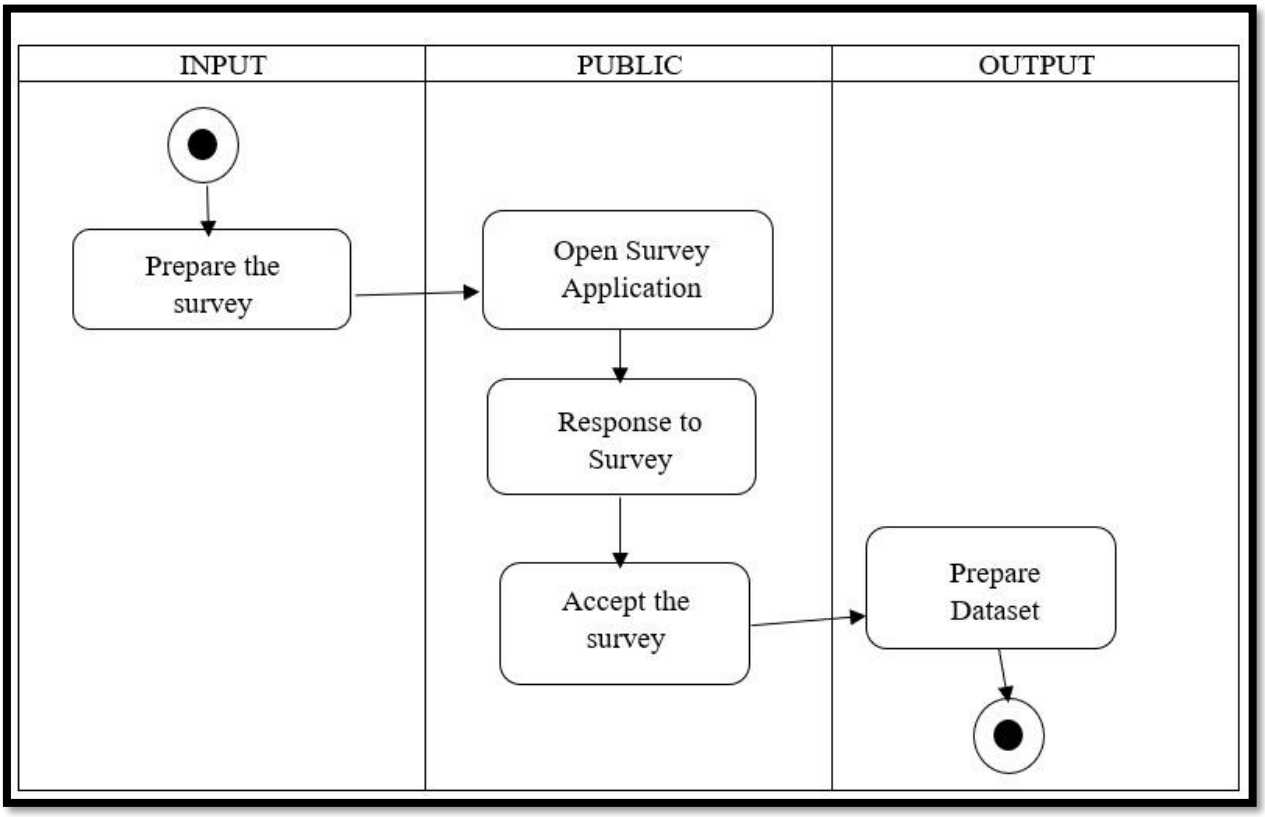

Figure 6.7: Activity Diagram for Survey

6.2.4 Use case Diagram

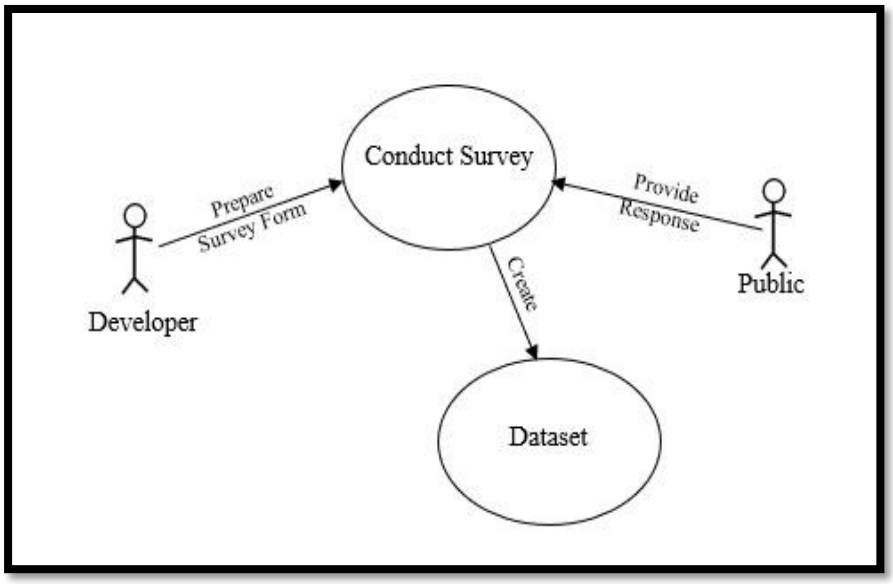

Figure 6.8: Use Case Diagram for Survey

So, the above figure 6.8 Shows the use case diagram of the survey module. Here firstly the developer prepares a survey form with the number of questionnaires that are uploaded by the developer. Once the developer prepares the survey form he/she conducts the survey and circulates the forms among the public. Now the public responds to the survey created by the developer and sends their responses. Finally, the datasets are collected.

6.3 Processing Module:

The Processing module does all the cleaning and training of the dataset and filtering by using the certain algorithm by importing them into the system. The following activities takes place in a processing module:

- **Importing Libraries:** In order to perform the necessary actions importing the libraries that support the following is must.
- **Importing Dataset:** Import the datasets which we have collected for our machine learning project.
- **Data cleaning**: is one of the important parts of machine learning. It plays a significant part in building a model.
- **Splitting the Dataset into the Training set and Test set:** divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model.

6.3.1 Sequence Diagram

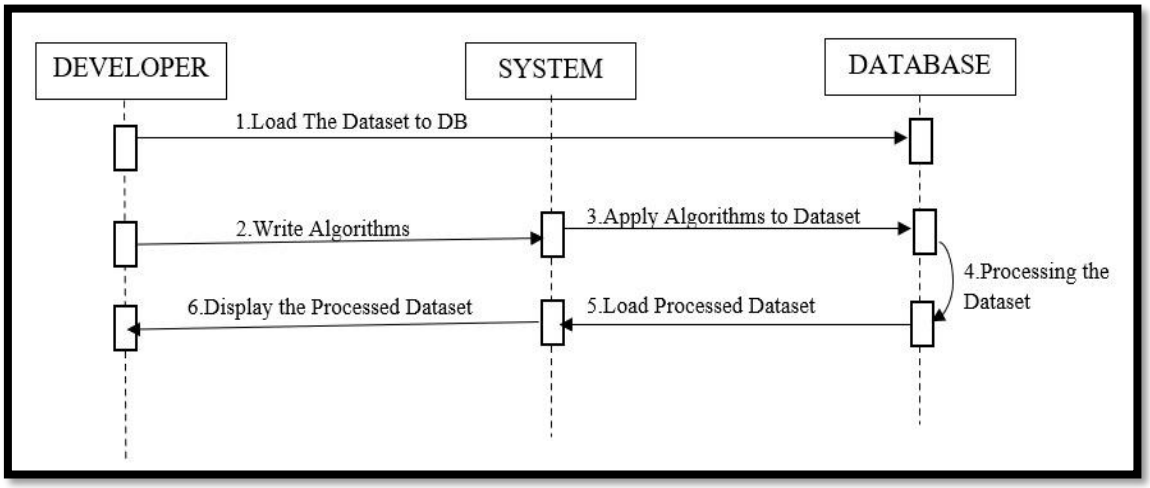

Figure 6.9: Sequence Diagram for Processing

Once the survey is held and the entire responses are collected and saved in the database. All the responses will be loaded in the database i.e. the dataset. Then the machine learning algorithms i.e. random forest algorithm, KNN algorithm, K-means algorithm and liner regression algorithm will be applied to the dataset. Once all the machine learning algorithms are applied the dataset is processed with the algorithms. Then the dataset is loaded i.e. the processed dataset lastly the processed dataset is displayed as shown in Figure 6.9

6.3.2 Data Flow Diagram

This is one of the modules in our project. Here firstly the developer writes the ML algorithm to the collected dataset which shows the step 1 in the Figure 6.10 Then the used ML algorithms are linear regression, K means, KNN, Random-forest. Then these algorithms are applied on to the collected datasets which is shown as step 2 in the diagram. Then these collected datasets are processed which is the step 3 which shows the processing of collected dataset. Finally, the processed datasets are displayed.

.

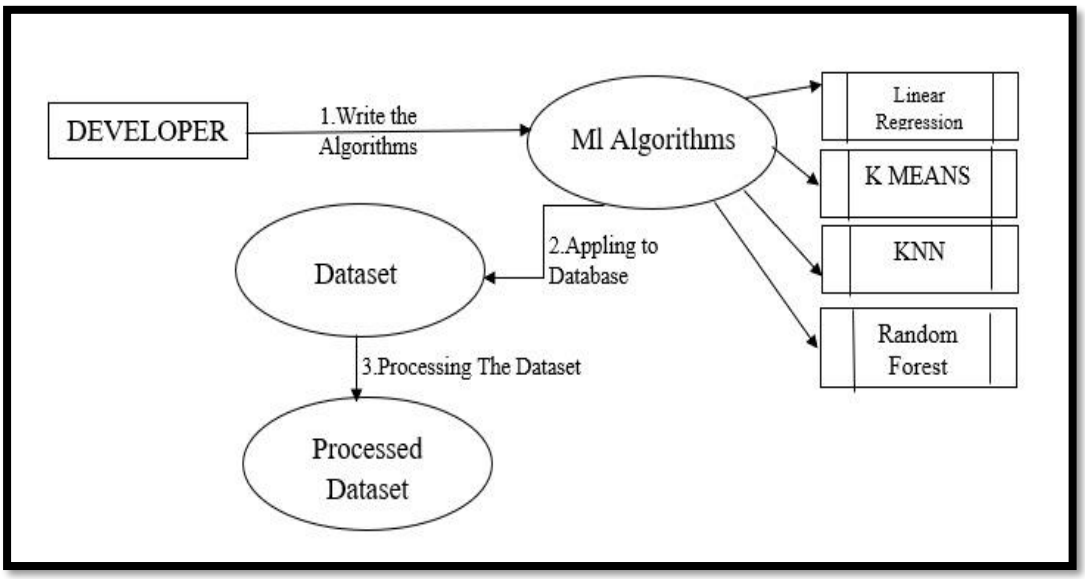

Figure 6.10: Data Flow Diagram for Processing

6.3.3 Activity Diagram

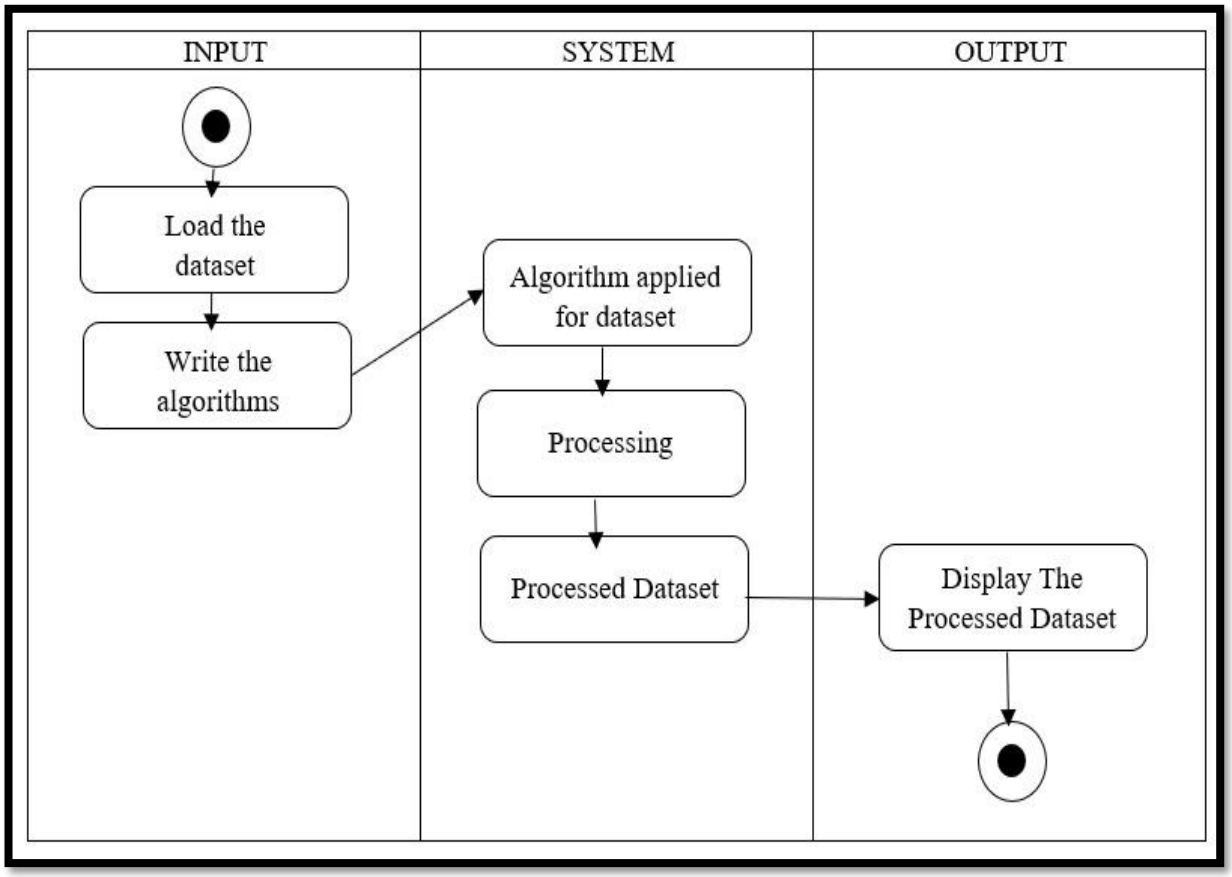

Figure 6.11: Activity Diagram for Processing

The figure 6.11 Shows the activity diagram for the processing module. Here firstly the process is started in the column of input the datasets are loaded and algorithms like KNN, K-Means, linear regression and random forest are applied on to the datasets that are stored in the system. Then the datasets are processed in the system end then these processed Datasets are collected and sent on to output end then finally the processed datasets are displayed in output end and the process is lastly stopped.

6.3.4 Use Case Diagram

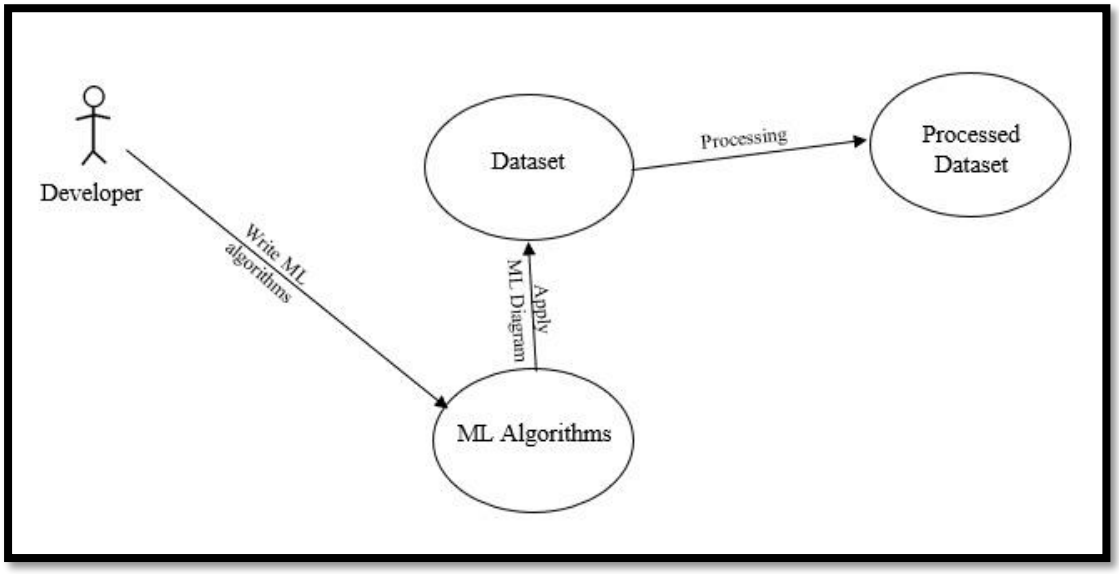

Figure 6.12: Use Case Diagram for Processing

The developer will write the Machine learning algorithms. All these machines learning algorithms i.e. random forest algorithm, k means algorithm, KNN algorithm and liner regression will be applied to the dataset that have been collected from the survey about the regional language implementation on the professional degree courses. Once the machine learning algorithms are applied to the dataset, the dataset will be processed which means algorithm once applied will do the working nd then we finally get the processed dataset as shown in Figure 6.12

6.4 Predicting Module

The predicting module is another major module in the project. Here, the trained example will be tested and the clustering will be done. The following activities takes place in predictive module:

- **Decision trees making:** Decision trees are a simple, but powerful form of multiple variable analysis. They are produced by algorithms that identify various ways of splitting data into branch-like segments
- **New data Clustering:** Clustering the new data is the most important phase of this module here the data will be classified according to the trained example by the dataset.

6.4.1 Sequence Diagram

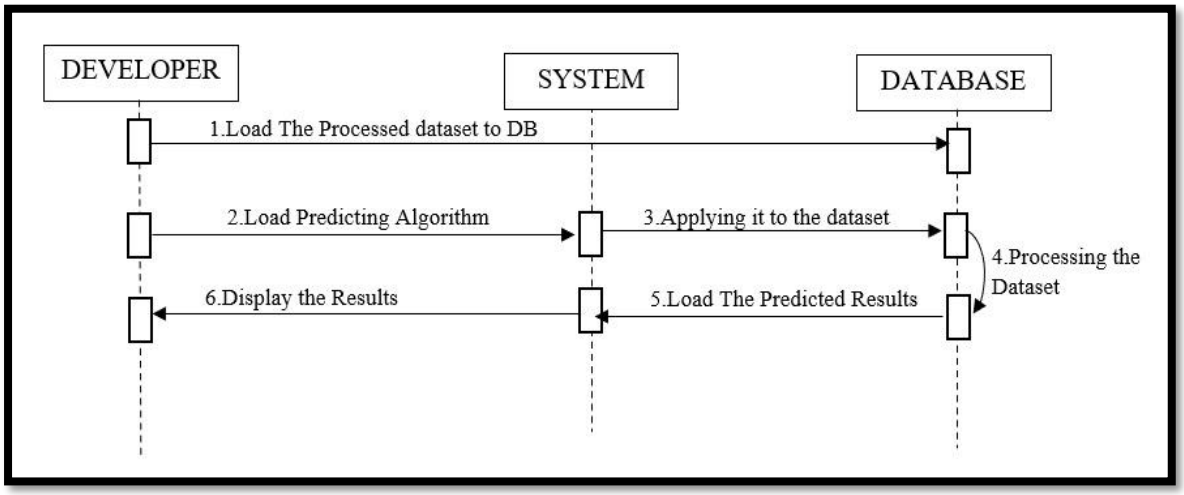

Figure 6.13: Sequence Diagram for Predicting

The above figure 6.13 Shows the sequence diagram of the predicted module. Once the survey is held and the entire responses are collected, saved in the database. All the responses will be loaded in the database i.e. the dataset. Then the predicted algorithms i.e. random forest algorithm, KNN algorithm, K- Means algorithm and liner regression algorithm will be applied to the dataset. Once all the algorithms are applied the dataset, the datasets are processed in the database and the loaded predicted results are sent to the system, Finally the system displays the results to the developer.

6.4.2 Data Flow Diagram for Predicting

As shown in Figure 6.14 the developer writes the prediction algorithm which are k-means, KNN linear regression and random forest to the overall collected datasets and then applies that in order to process the algorithms applies dataset then the datasets are processed, the collected datasets are then used in the phase processing. Finally, after processing the complete dataset the predicted results are displayed to both the user and the developer.

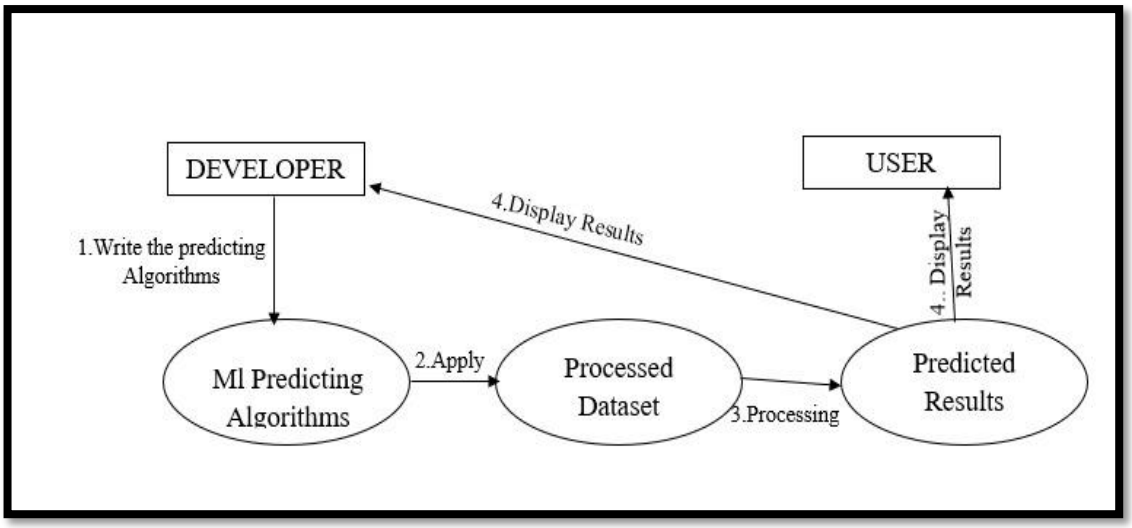

Figure 6.14: Data Flow Diagram for Predicting

6.4.3 Activity Diagram for Predicting

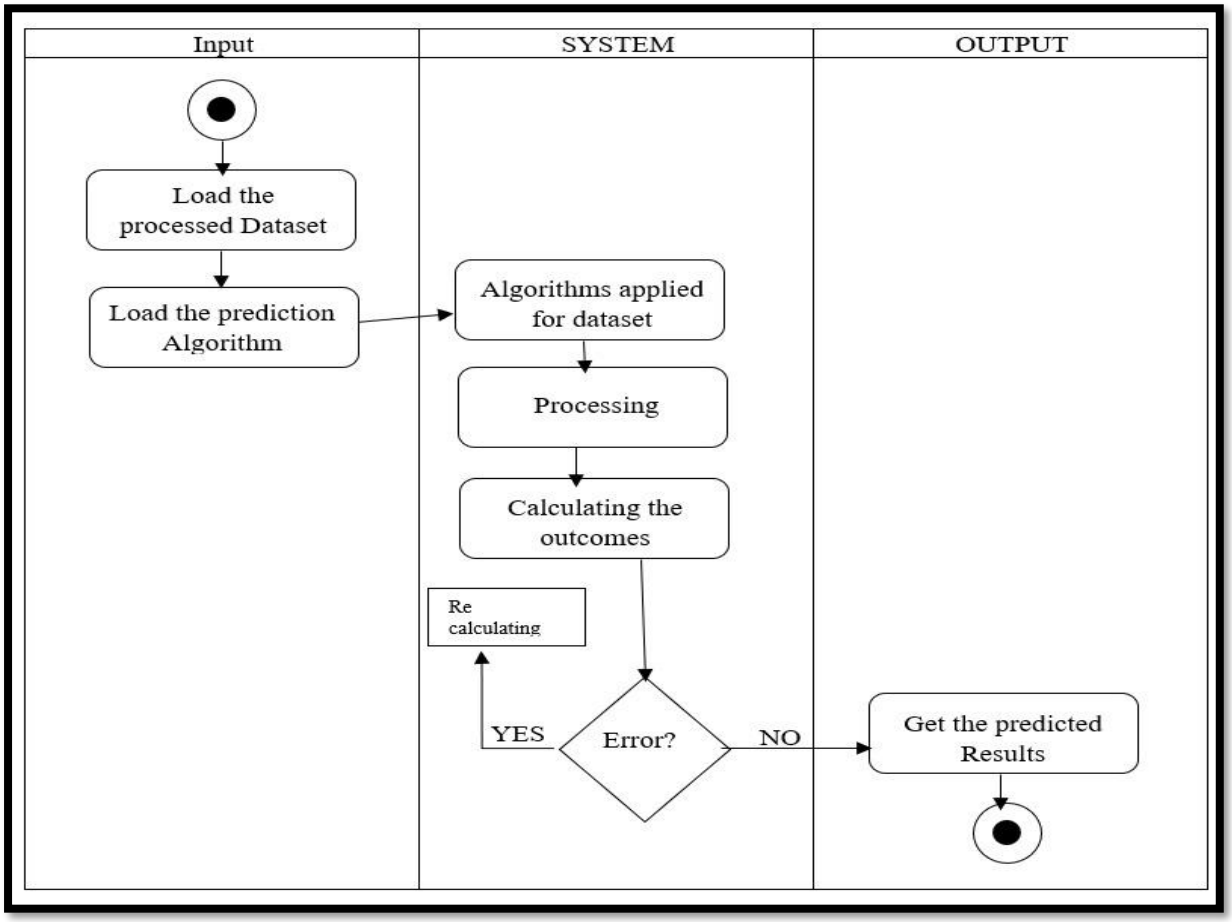

Figure 6.15: Activity Diagram for Predicting

The present dataset has to be loaded. Once the processed dataset is loaded then the prediction algorithms has to me loaded i.e. K means algorithm, KNN algorithms, linear regression and random forest algorithm. Once these algorithms are applied to the dataset then it will be processed. After the algorithms are applied the outcomes are calculated and if there is an error it will recalculate the outcomes again and then again it checks if there is an error. Only when there is no error we get the prediction results. Finally, we get the predicted results as shown in Figure 6.15

6.4.4 Use Case Diagram for Predicting

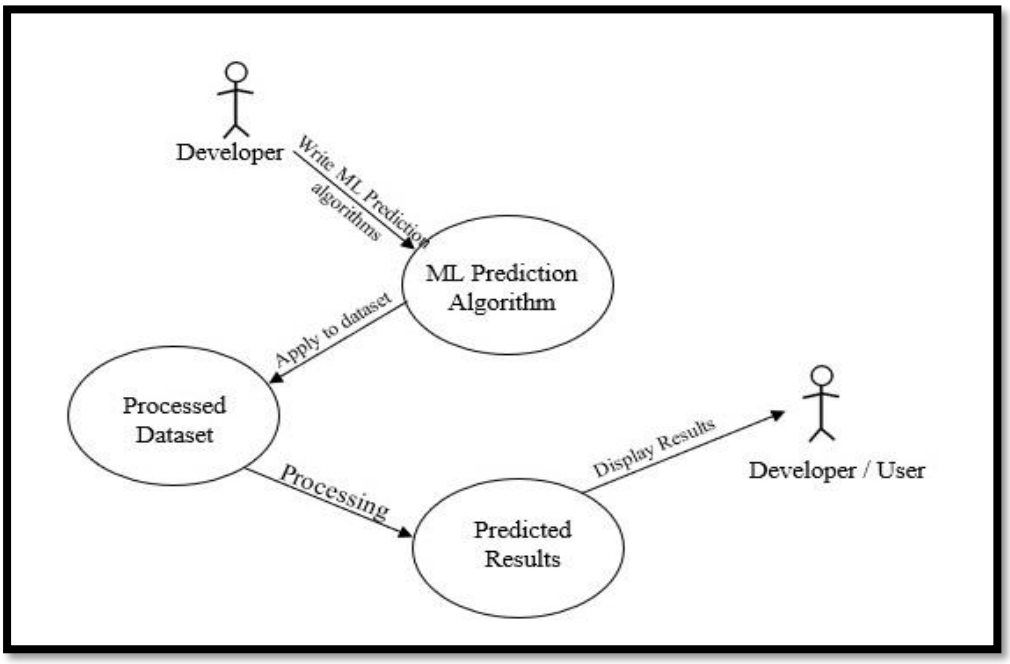

Figure 6.16: Use Case Diagram for Predicting

As shown in Figure 6.16 the developer will write the Machine learning algorithms for the prediction. The algorithms are linear regression algorithm, KNN algorithm, and random forest algorithm will be written by the developer. All the machine learning algorithms for prediction will be applied to the processed dataset. This processed dataset will undergo some working and from that we will get the predicted results. After the required results are out this can be displayed to the developer or the user. The end result will be the prediction of students out there are comfortable in pursuing the professional degree course in their regional language or not. This result is displayed in the end.

Chapter 7

SYSTEM IMPLEMENTATION

System implementation is the process of construction of the new system and delivery of that system into production construction and delivery phases of life cycle. The construction phases do two things build and tests a functional system that fulfills business or organizational design requirements, and implements the interface between the new system and existing production system. Implementation is the realization of an application, or execution of plan, idea, model, design, specification, standard, algorithm, or policy. In other words, an implementation is a realization of technical specification or algorithm as a program, software component, or other computer system through programming and deployment. Many implementations may exist for a given specification or standard.

Implementation is one of the most important phases of the Software Development Life Cycle (SDLC). It encompasses all the processes involved in getting new software or hardware operating properly in its environment, including installation, configuration, running, testing, and making necessary changes. Specifically, it involves coding the system. This phase of the system is conducted with the idea that whatever is designed should be implemented, keeping in mind that it fulfills user requirements, objective and scope of the system. The implementation phase produces the solution to the user problem.

7.1 Overview of System Implementation

The project is implemented considering the two following aspects:

- Usability Aspect
- Technical Aspect

7.1.1 Usability Aspect

The usability aspect of implementation of the project is realized using two principles:

The Project is being implemented as a Jupyter Notebook (.ipynb)

There are may be many ways to implement ML algorithms and data analytics project with Python as the primary Programming Language used. We have chosen to implement in Jupyter Notebook as Jupyter provides a wonderful platform to run the algorithms and get the exact output and no need of downloading all the libraries at once it can be imported directly from the platform. Most importantly Jupyter is an independent application that means once you get the code downloaded along with your dataset it can run in any of the platform where Jupyter and the supporting anaconda is installed. Lastly the. ipynb is such a flexible file that it can be opened in any of the browser and doesn't need an specific installation to view them. .ipynb also allows you see the coding part of the project for better clarity on how the project is being implemented

7.1.2 Technical Aspects

The technical aspects of the projects is realized as explained below:

Anaconda For setting up Notebook

Anaconda is a distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS. The default installation of Anaconda2 includes Python 2.7 and Anaconda3 includes Python 3.7. However, it is possible to create new environments that include any version of Python packaged with anaconda

Jupyter Notebook

Project Jupyter is a project and community whose goal is to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages". A Jupyter Notebook document is a browser-based REPL containing an ordered list of input/output cells which can contain code, text (using Markdown), mathematics, plots and rich media. Underneath the interface, a notebook is a JSON document, following a versioned schema, usually ending with the ".ipynb" extension. A Jupyter Notebook can be converted to a number of open standard output formats (HTML, presentation slides, LaTeX, PDF, ReStructuredText, Markdown, Python) through "Download As" in the web interface, via the nbconvert library.

7.2 Database

The project is based on nationwide survey which was conducted for weeks and it was being circulated via Google Forms. Google forms allows you to store them in a structed format for formatting in a excel sheets and the same can be converted into .csv file and that makes us the dataset for running the algorithms. A Dataset is the basic data container in PyMVPA. It serves as the primary form of data storage, but also as a common container for results returned by most algorithms. The dataset will then be stored in Jupyter notebook by the upload option Object storage and Jupyter Notebooks for general-purpose usage are on the rise. The demand for big data and data elements is continuously increasing. While object storage is required to store the enormous data elements on a continuous rise, Jupyter Notebooks can compute and analyze these datasets and data.

7.3 Implementation Support

Implementation Support is a planned approach to integrate new or upgraded software or systems into the existing workflow of an organizational structure to help ensure the success of a Project overall system.

7.3.1 Installation of Anaconda

- Download the Anaconda installer**.**
- Double-click the installer to launch and click next
- Read the licensing terms and click I Agree.
- Select an install for Just Me unless you're installing for all users (which requires Windows Administrator privileges) and click Next.
- Select a destination folder to install Anaconda and click the Next button
- Choose whether to add Anaconda to your PATH environment variable. We recommend not adding Anaconda to the PATH environment variable, since this can interfere with other software. Instead, use Anaconda software by opening Anaconda Navigator or the Anaconda Prompt from the Start Menu.
- Choose whether to register Anaconda as your default Python. Unless you plan on installing and running multiple versions of Anaconda or multiple versions of Python, accept the default and leave this box checked.

- Click Install. If you want to watch the packages Anaconda is installing, click Show Details. And click next
- After a successful installation you will see the “Thanks for installing Anaconda” dialog box:
- If you wish to read more about Anaconda.org and how to get started with Anaconda, check the boxes “Anaconda Distribution Tutorial” and “Learn more about Anaconda”. Click the Finish button.

7.3.2 Installation of Python

- **Select Version of Python to Install:**

1. The installation procedure involves downloading the official Python .exe installer and running it on your system.
2. The version you need depends on what you want to do in Python. For example, if you are working on a project coded in Python version 2.6, you probably need that version. If you are starting a project from scratch, you have the freedom to choose.
3. If you are learning to code in Python, we recommend you download both the latest version of Python 2 and 3. Working with Python 2 enables you to work on older projects or test new projects for backward compatibility.

- **Download Python Executable Installer:**

1. Open your web browser and navigate to the Downloads for Windows section of the official Python website.
2. Search for your desired version of Python. At the time of publishing this article, the latest Python 3 release is version 3.7.3, while the latest Python 2 release is version 2.7.16.
3. Select a link to download either the Windows x86-64 executable installer **or** Windows x86 executable installer**.** The download is approximately 25MB.

- **Run Executable Installer:**

1. Run the **Python Installer** once downloaded. (In this example, we have downloaded Python 3.7.3.)
2. Make sure you select the **Install launcher for all users** and **Add Python 3.7 to PATH** checkboxes. The latter places the interpreter in the execution path. For older versions of Python that do not support the **Add Python to Path** checkbox.
3. Select **Install Now** – the recommended installation options.

7.4 Algorithm

An algorithm is a self-contained step by step set of operations to be performed. Algorithm is an effective method that can be expressed within a finite amount of space and time and in a well-defined formal language for calculating the function starting from an initial starte and initial input the instructions described a computation that when executed, proceeds through a finite number of well-defined successive states, eventually producing output and terminating at final ending state.

Step 01 : Start the Jupyter Python Application
Step 02 : Upload the dataset.
Step 03 : Import the necessary libraries.
Step 04 : Fetch the values.
Step 05 : Apply the data analysis codes.
Step 06 : Run it on kernel.
Step 07 : Get the Output for getting the insights of the datasets and for further analysis.
Step 08 : After the analysis the dataset must be split into training and testing datasets.
Step 09 : For feature selection using Random forest classifier.
Step 10 : After extracting the features build a model for evaluation.
Step 11 : Build the model using linear regression.
Step 12 : Scoring new data on training dataset and testing dataset.
Step 13 : Predicting the output for the new data.

7.5 Pseudo Code

Pseudocode is an artificial and informal language that helps programmers develop algorithms. Pseudocode is a "text-based" detail (algorithmic) design tool. The rules of Pseudocode are reasonably straightforward. All statements showing "dependency" are to be indented. These include while, do, for, if, switch

7.5.1: Data Analysis and Feature Engineering

Importing all the necessary packages

All the important modules such as pandas,numpy,matplotlib.pyplot,seaborn is imported.

Assign dataset to survey_data by using pd.read_excel()

This function is used to import dataset into the engine.The dataset was generated by Google form hence the format of the dataset is .xlsx

survey_data.head()

To display the imported dataset with coloumn names. Print the count of the target variable by using.

.values_count()

Here the target variable is knowledge_improvement. By using the function, it enables to get the number of count the responses that we go to the particular attribute.

For visualizing the target variable as a bar graph

Using the function

.plot_bar()

A matplotlib.pyplot function used to draw graphs for the given values

survey_data.info()

To get to know about the variables in the dataset that includes the count of values, datatype of each variable. Assign the cleansed dataset to survey_data by using

.drop

This function used to clean the dataset, in this case as timestamp and the respondet name is not required those attribute columns are dropped

Get the catergorical features

This helps in getting the number of variable/attribute that the dataset has after cleaning.

Ploting the graphs for the features extracted

plt.figure(figsize=())

To set the size of the graph plot area

Intialize variable ax <- sns.countplot()

The Seaborn Countplot function creates bar charts of the number of observations per category. The catergories are passed through variable x inside the function by passing the survey_data as the dataset.

Feature visualization:

x = mother_tounge_rating

This feature gives the count of how people are good with their mother tounge

mother_tounge_rating and gender

no_of_fluet_languages

mother tounge and no of language fluent it

categorize features

minimizing the unique values

For the extra features in the dataset the graphs and the analytics has to be made hence ploting of remaining features

remaining_features include

'teaching_language_preference',

'notes_language_preference',

'questionpaper_language_preference',

'exam_in_regional_language',

'teaching_in_regional_language'

7.5.2 Model Building:

sImporting libraries

import pandas as pd

import confusion matrix , accuracy_score from sklearn_metrics

Use **pd.read_csv()** to import training and testing data

this function is used to import training and testing dataset for performing operation.

Logistics Regression

Import Libraries

From sklearn.linear_model import LogisticRegression module

Which enables us to perform

Classifier.Fit()

Fit the dataset into classifier by passing x_train and y_train datasets

Classifier.predict()

To predict the outcomes of the classifier by passing training and testing dataset

roc_auc_score()

Finds the roc_auc accuracy score

From yellowbricks.classifier **import ROCAUC**

To visualise the classifier and this function helps to differentiate the new values in which region it falls in.

Visualizer.fit()

Fit the training and testing data into the graph.

Scoring new data

Classifier.predict()

By passing new values to the function it predicts the outcomes based on the ploted graph by looking on the position where the value plot on.

Chapter 8

SOFTWARE TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, subassemblies, assemblies and/or finished product. It is the process of exercising software with the intent of ensuring that the software system meets its requirements and user expectations and does not fail in an unacceptable manner. The system has been verified and validates by running the test data and live data.

8.1 Levels of Testing

Testing levels are the procedure for finding the missing areas and avoiding overlapping and repetition between the development life cycle stages. We have already seen the various phases such as Requirement collection, designing, coding testing, deployment, and maintenance of SDLC (Software Development Life Cycle).

8.1.1 Unit Testing

Unit testing is a development procedure where programmers create tests as they develop software. The tests are simple short tests that test functionally of a particular unit or module of their code, such as a class or function. Using open source libraries like unittest, cunit, oppunit and nun it these tests can be automatically run and any problems found quickly. As the tests are developed in parallel with the source unit test demonstrates its correctness.

Test case: Processing model

Steps	Test Action	Results
Step 1	Load the Dataset	The Dataset has been loaded
Step 2	Write the required Algorithms	The required Algorithms have been written
Step 3	Apply the required algorithms for the dataset	The algorithms have been applied

Step 4	Process the dataset	The dataset has been processed
Step 5	Load the processed dataset	The Processed Dataset has been loaded
Step 6	Display the dataset	The dataset has been displayed on the screen.

8.1.2 User Input

In User Interface the source program from any of the languages are provided as the user input**.**

8.1.3 Error Handling

In the system, we have tried to handle all the errors that occurred while running the application. The common errors we saw were reading a tuple with an attribute set to null and database connection getting lost. For testing we used Top-Down design a decomposition process which focuses as the flow of control, at latter strategies concern itself with code production. The first step is to study the overall aspects of the tasks at hand and break it into a number of independent modules. The second step is to break one of these modules further into independent sub modules. One of the important features is that each detail at lower levels are hidden. So, unit testing was performed first and then system testing.

8.1.4 Integration Testing

Data can be lost across an interface, one module can have an adverse effect on the other sub function, when combined may not produce the desired functions. Integrated testing is the systematic testing to uncover the errors with an interface. This testing is done with simple with data and developed system has run successfully with this simple data. The need for integrated system is to find the overall system performance.

Top down testing can proceed in a depth-first or a breadth-first manner. For depth-first integration each module is tested in increasing detail, replacing more and more levels of detail with actual code rather than stubs. Alternatively, breadth-first would processed by refining all the modules at the same level of control throughout the application .in practice a combination

of the two techniques would be used. At the initial stages all the modules might be only partly functional, possibly being implemented only to deal with non-erroneous data. These would be tested in breadth-first manner, but over a period of time each would be replaced with successive refinements which were closer to the full functionality. This allows depth-first testing of a module to be performed simultaneously with breadth-first testing of all the modules.

The other major category of integration testing is Bottom Up Integration Testing where an individual module is tested form a test harness. Once a set of individual modules have been tested they are then combined into a collection of modules, known as builds, which are then tested by a second test harness. This process can continue until the build consists of the entire application. In practice a combination of top down and bottom-up testing would be used. In a large software project being developed by a number of sub-teams, or a smaller project where different modules were built by individuals. The sub teams or individuals would conduct bottom-up testing of the modules which they were constructing before releasing them to an integration team which would assemble them together for top-down testing.

Steps to perform integration testing

Step 1: Create a test plan

Step 2: Create Test Cases and Test Data

Step 3: Once the components have been integrated execute the test cases

Step 4: Fix the bugs if any and re test the code

Step 5: Repeat the test cycle until the components have been successfully integrated

Name of the test	**Integration Testing**
Test Plan	To check whether the system works properly when all the modules are integrated
Test Data	Source program from any of the supported languages

8.1.5 System testing

Ultimately, system is included with other system components and the set of the system validation and integration tests are performed. System testing is a series of different tests whose main aim is to fully exercise the computer-based system. Although each test has different role all work should verify that all system elements are properly integrated and formed allocated function.

Name of the test	**System Testing**
Sample input	Source program from any of the supported languages
Expected Output	All the modules like login, execution, etc.
Actual Output	Application reacts to user inputs in expected manner
Remarks	Successful

8.1.6 Validating Testing

Validation testing is a concern which overlaps with integration testing. Ensuring that the application fulfils its specification is a major criterion for the construction of an integration rest. Validation testing also overlaps to a large extent with System Testing, where the application is tested with respect to its typical working environment. Consequently, for many processes no clear division between validation and system testing can be made.

8.1.7 Output Testing

After performing validation testing, the next step is output testing of the proposed system. Since the system cannot be useful if it does not produce the required output. Asking the user about the format in which the system is required tests the output displayed or generated by the system under consideration. The output format is considered in two ways, one is on screen format and the other is printed format. The output format on the screen is found to be corrected as the format was designated in the system according to the user needs.

8.1.8 Test Data and Output

Taking various kind soft data plays a vital role in system testing. After preparing the test data system under study is tested using the test data. While testing, errors are again uncovered and corrected by using the above steps and corrections are also noted for future use.

8.1.9 User Acceptance Testing

User acceptance testing of the system is the key factor for the success of system. A system under consideration is tested for user acceptance by constantly keeping in touch with perspective system at the time of development and making change whenever required. This is done with regard to the input screen design and output screen design**.**

Steps	Test Action	Expected Results	Actual Result	Remark
Step 1	Load the Dataset	The Dataset has to be loaded	The Dataset has been loaded	Pass
Step 2	Write the required Algorithms	The required Algorithms has to be written	The required Algorithms have been written	Pass
Step 3	Apply the required algorithms for the dataset	The algorithms have to be applied	The algorithms have been applied	Pass
Step 4	Process the dataset	The dataset has to be processed	The dataset has been processed	Pass
Step 5	Load the processed dataset	The Processed Dataset has to be loaded	The Processed Dataset has been loaded	Pass
Step 6	Display the dataset	The dataset has to be displayed on the screen.	The dataset has been displayed on the screen.	Pass

Chapter 9

RESULT & DISCUSSION

Finally, this project was conducted in order to receive the opinions of the people about the implementation of regional languages from all around the nation. We conducted a survey to get all the responses, the outcomes of the responses are shown above. The people from all over gave their opinions on the implementation of regional languages in professional degree course.

Few choose regional language over English language, few choose hybrid language which means the combination of both regional and English language, wherein few yet choose English as their convenience. This implementation was done with the minimal amount of data or responses generally but this can also be used for a proper feedback from people by collection a huge amount data or responses. The result we obtained was by jut the minimal amount of responses. So, the result was general and result outcomes are shown above with the help of snapshots that we received by conducting the survey.

Chapter 10

SNAPSHOTS

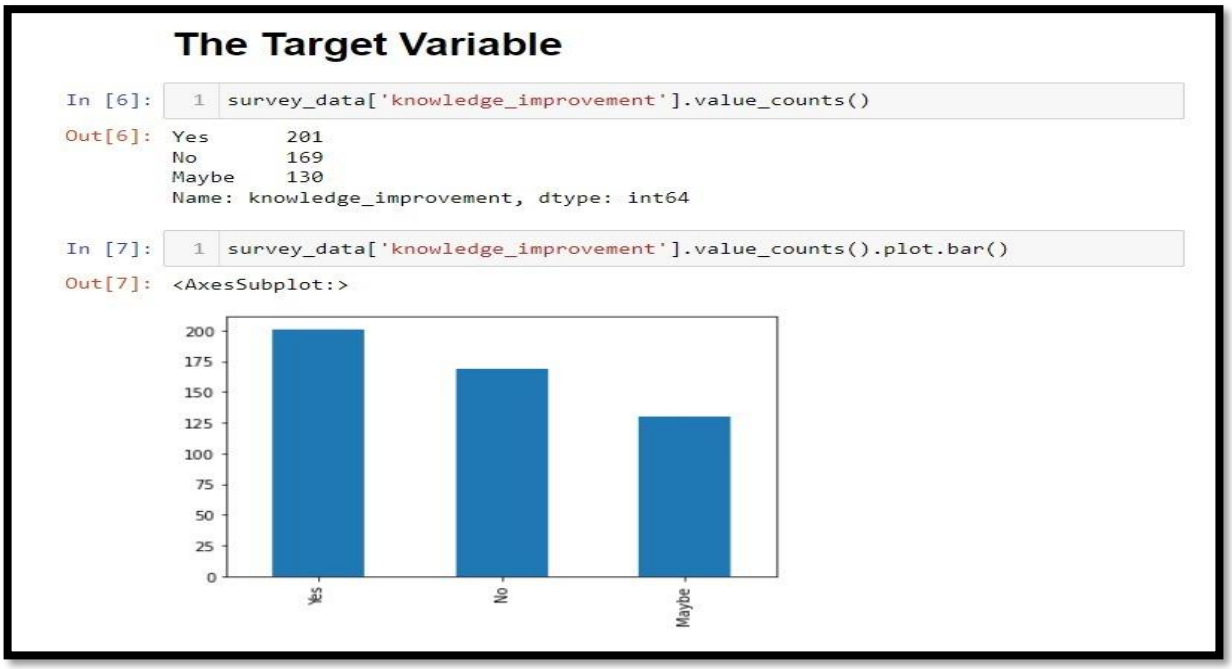

Figure 9.1: Target Variable Count

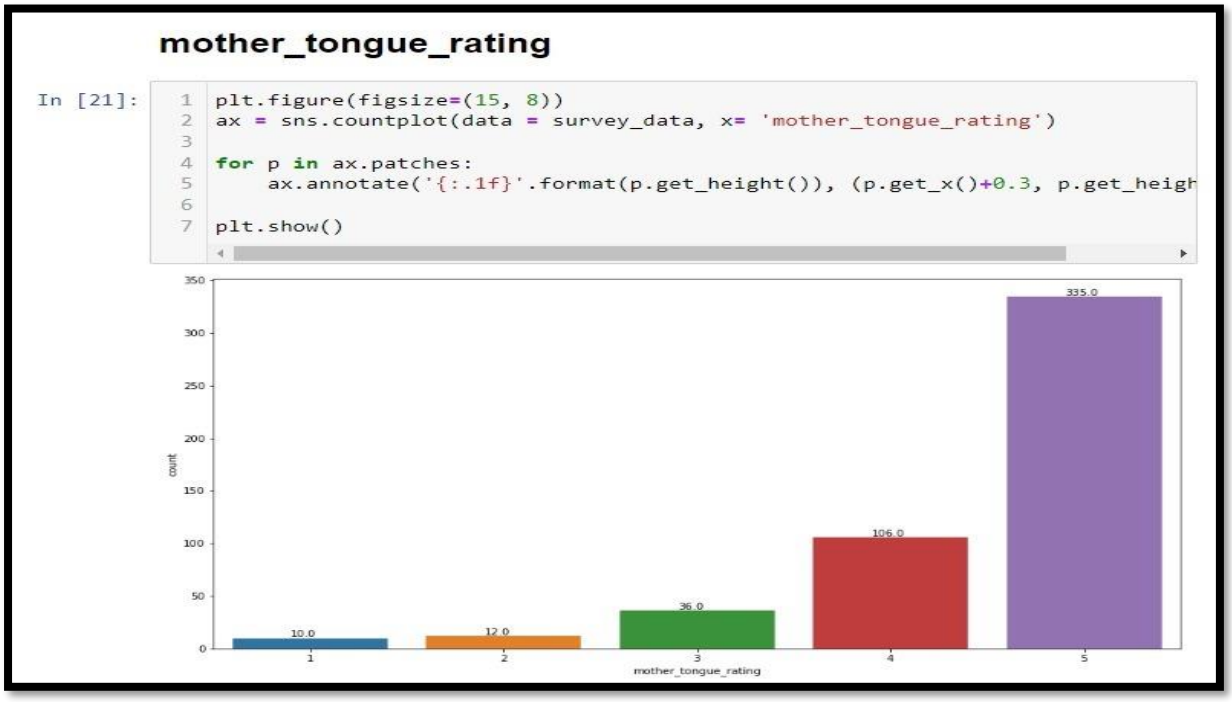

Figure 9.2: Mother Tongue Rating Count

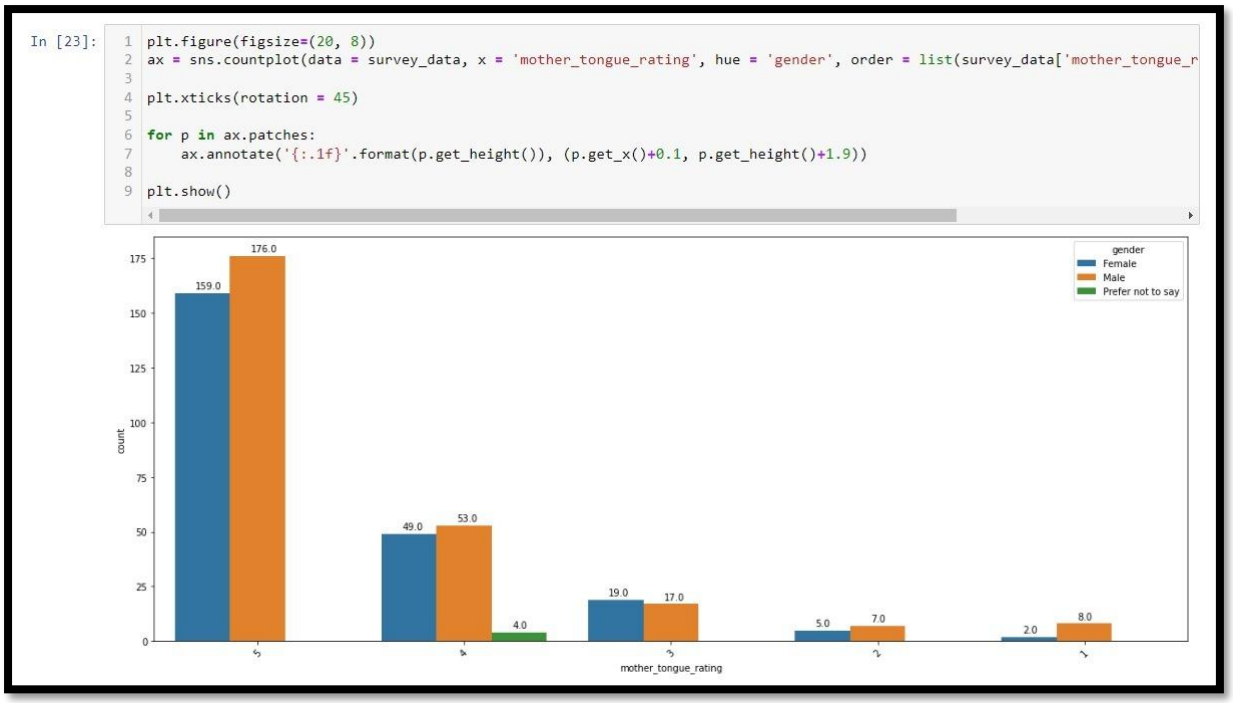

Figure 9.3: Gender Based Mother Tongue Rating

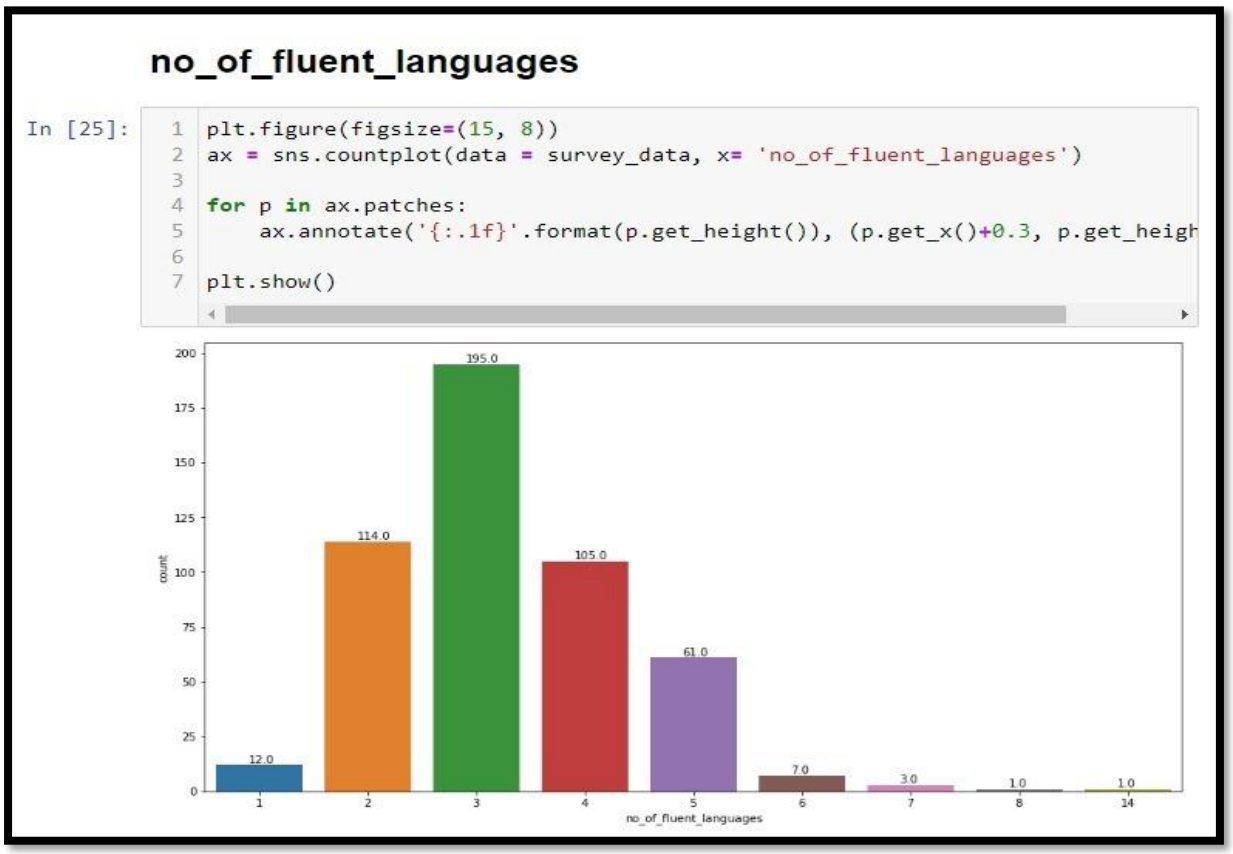

Figure 9.4: Number of Fluent Languages

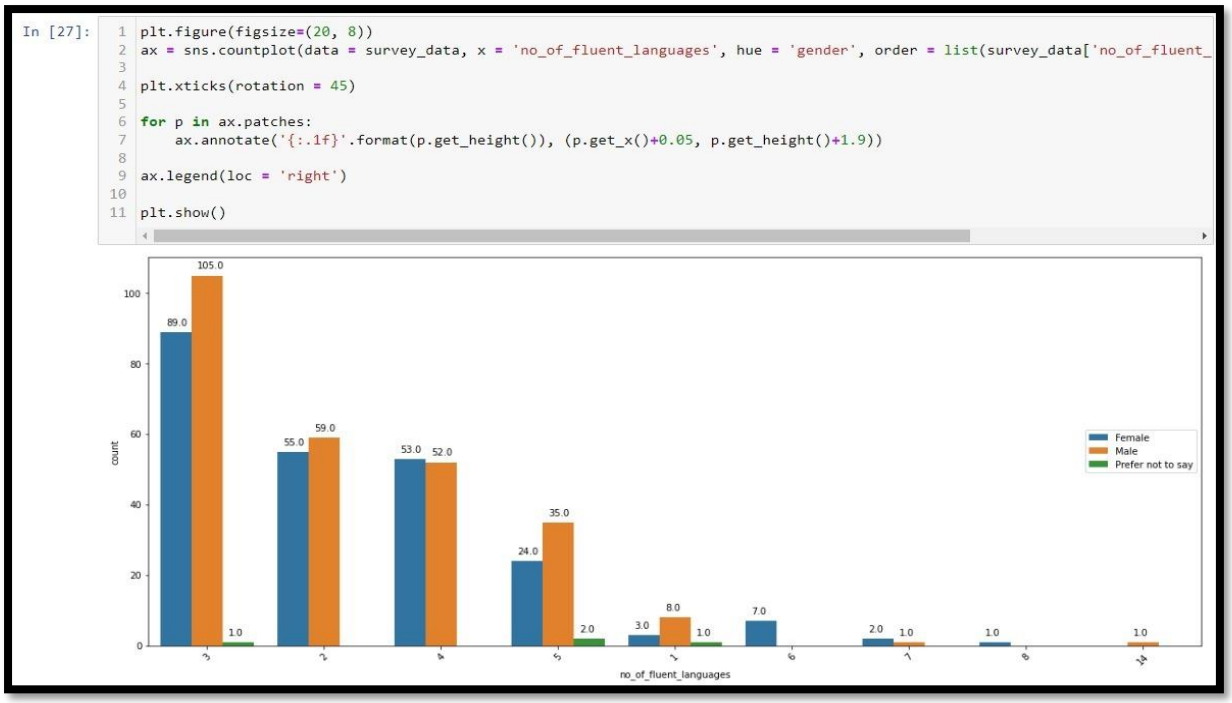

Figure 9.5: Gender Based Number of Fluent Languages

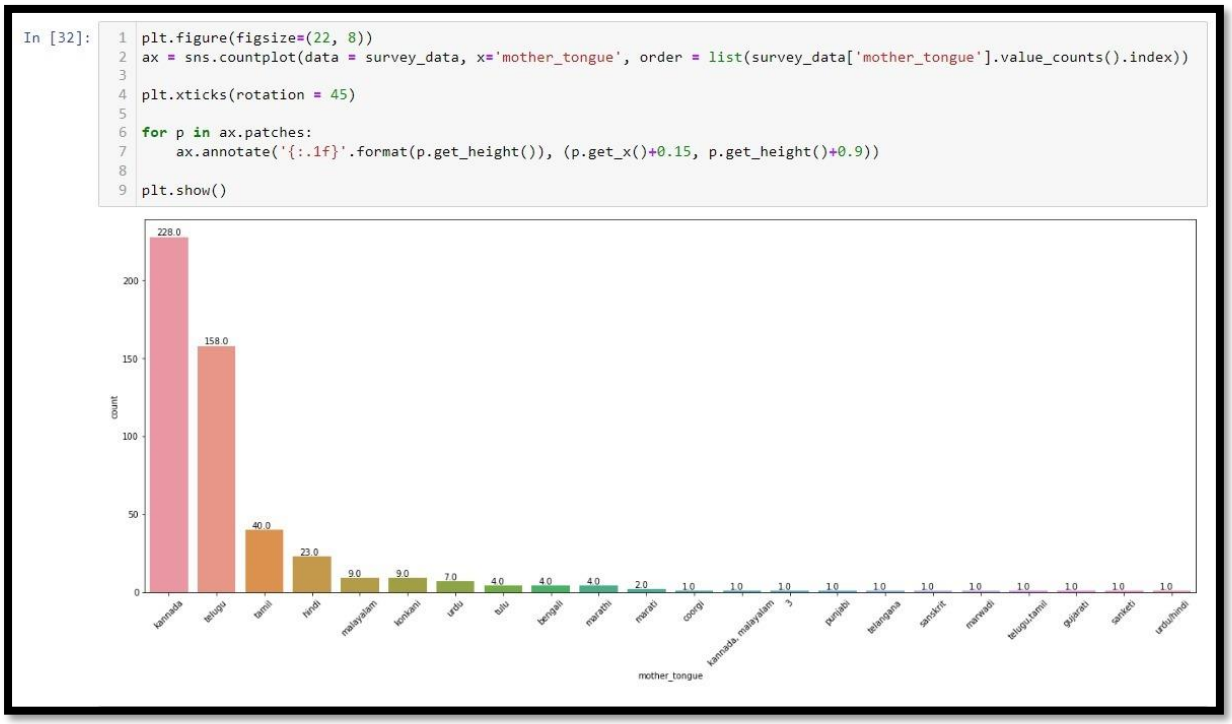

Figure 9.6: Region Based Mother Tongue

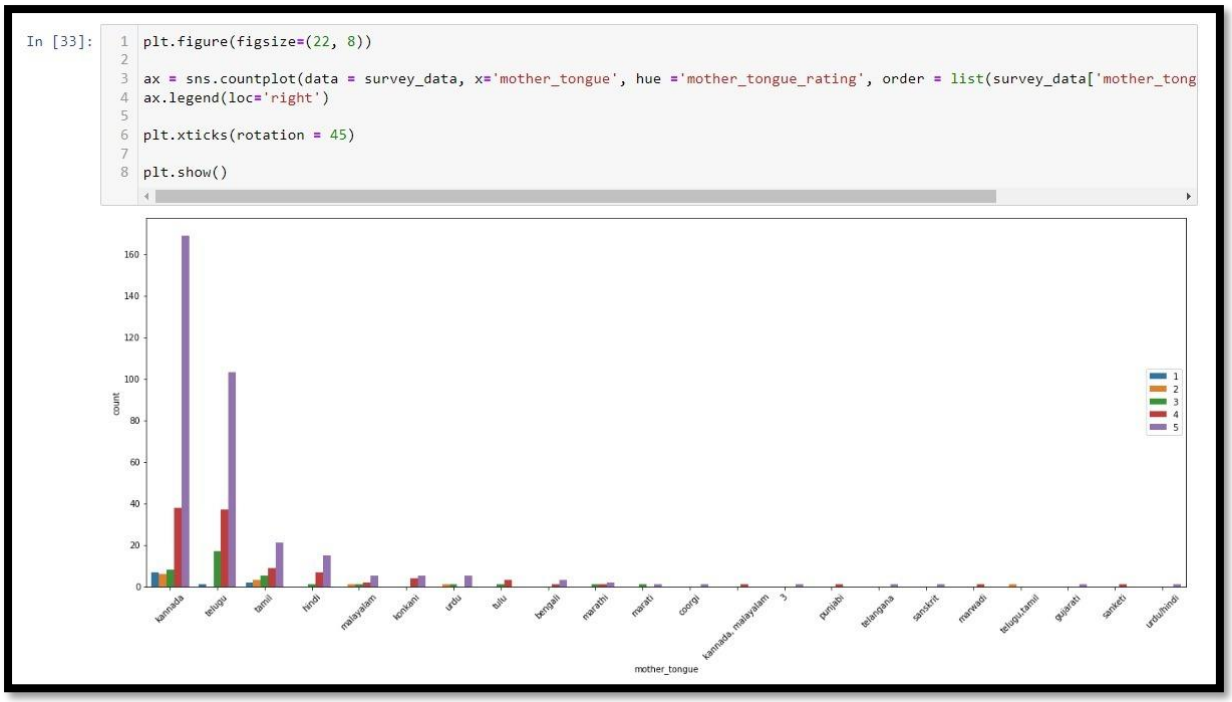

Figure 9.7: Mother Tongue Rating

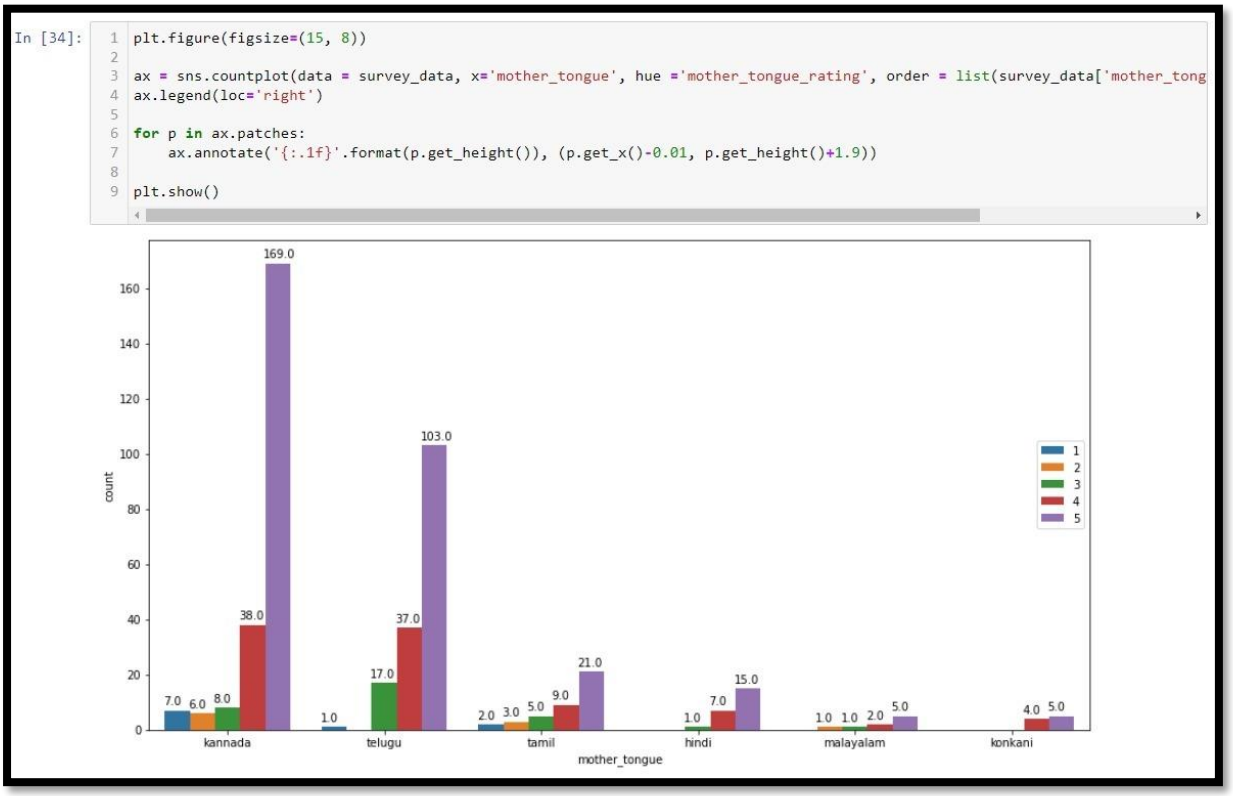

Figure 9.8: Mother Tongue Rating (Top Models)

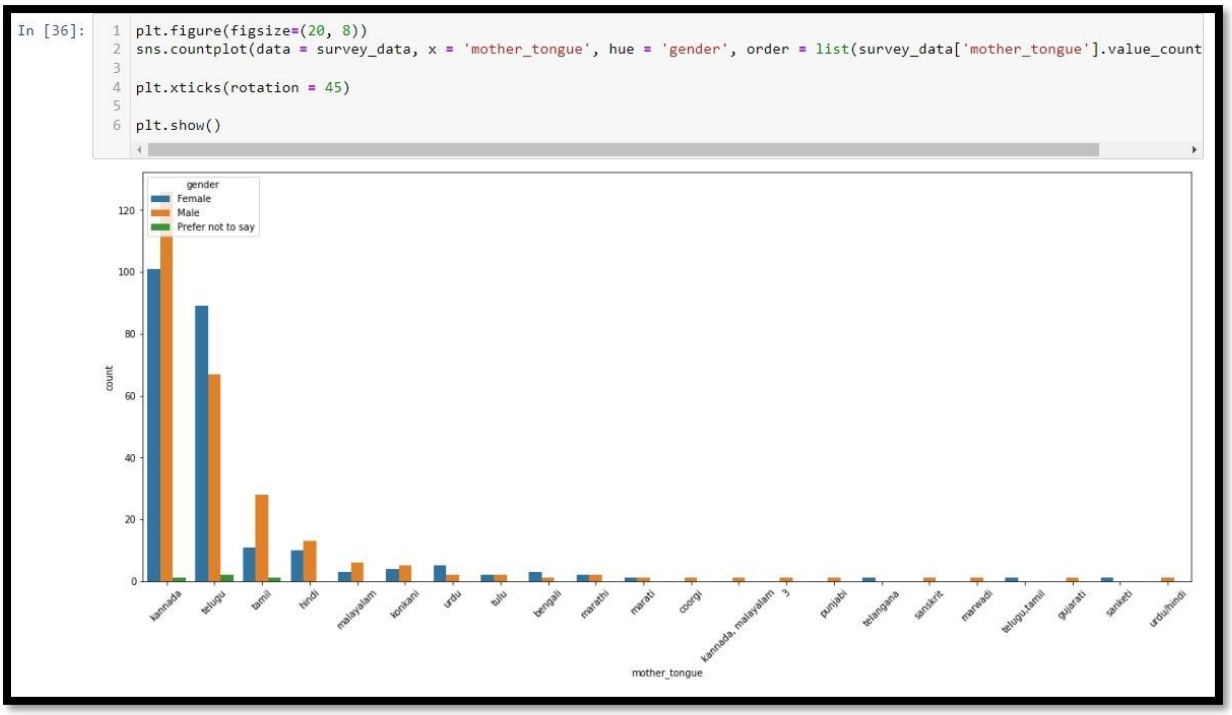

Figure 9.9: Gender Based Mother Tongue

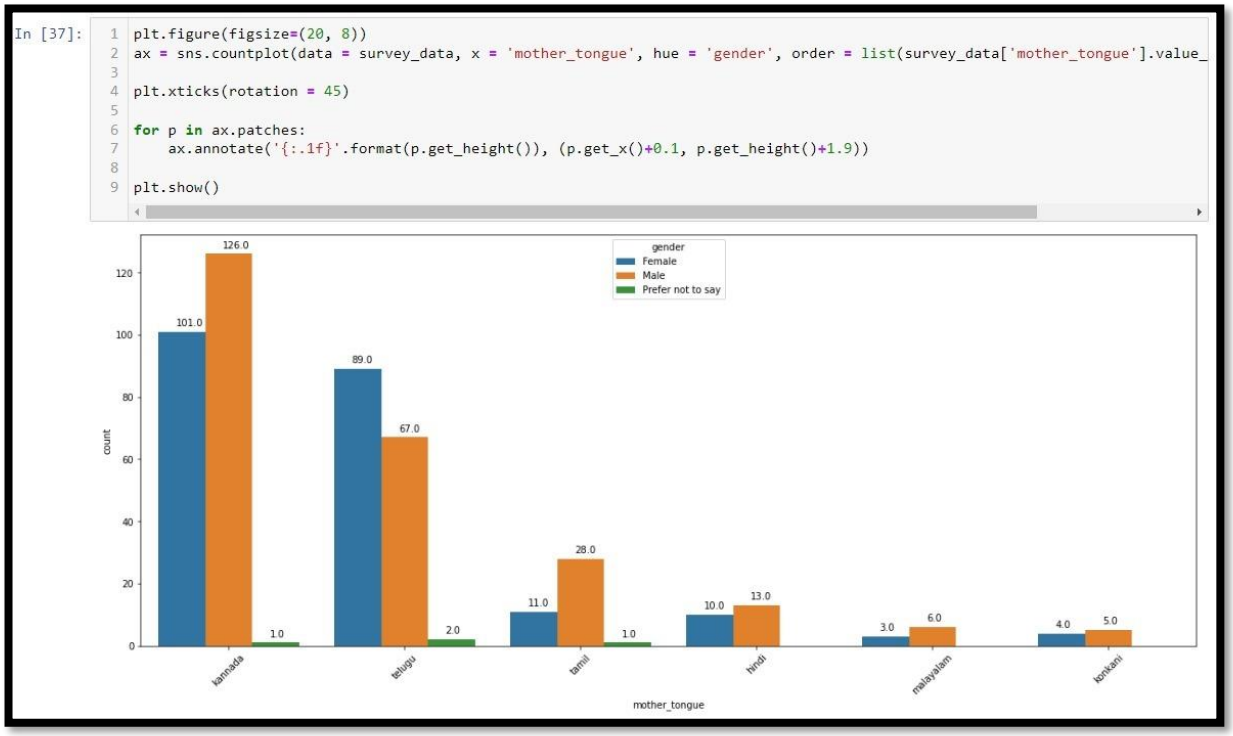

Figure 9.10: Gender Based Mother Tongue (Top Models)

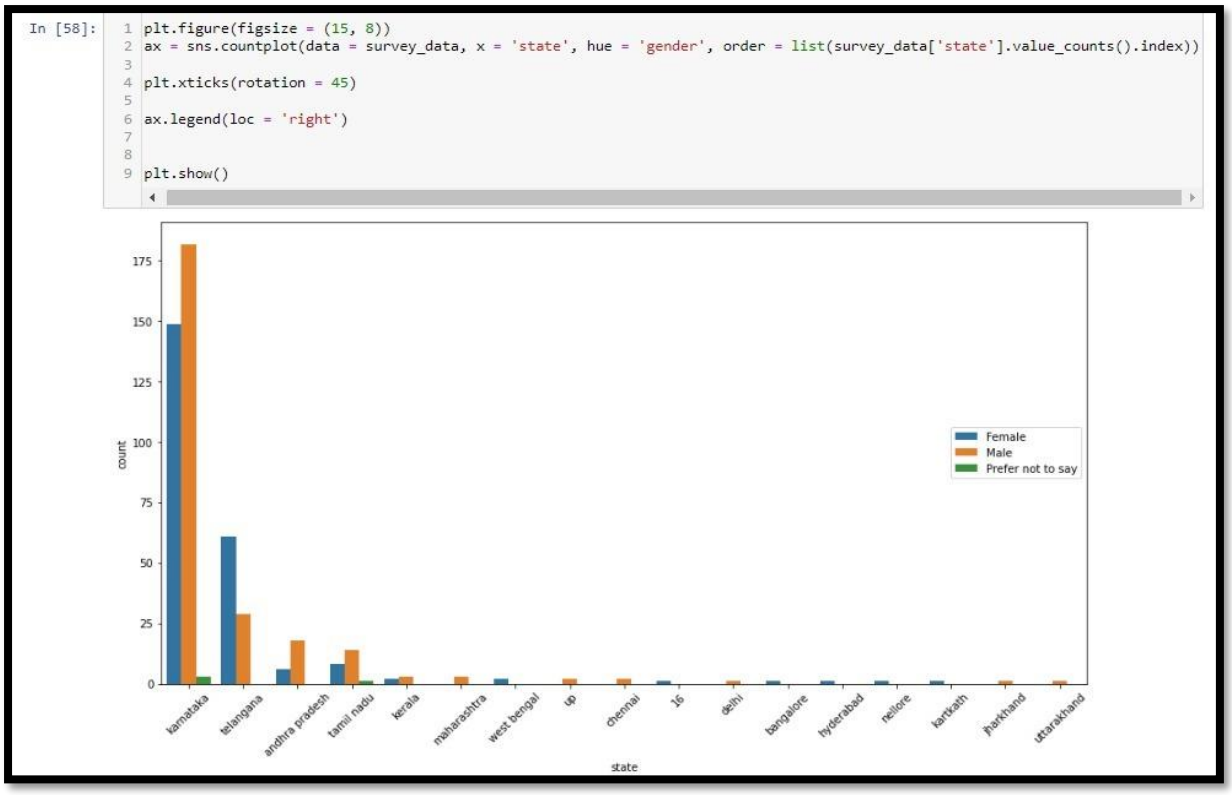

Figure 9.11: States Based on Gender

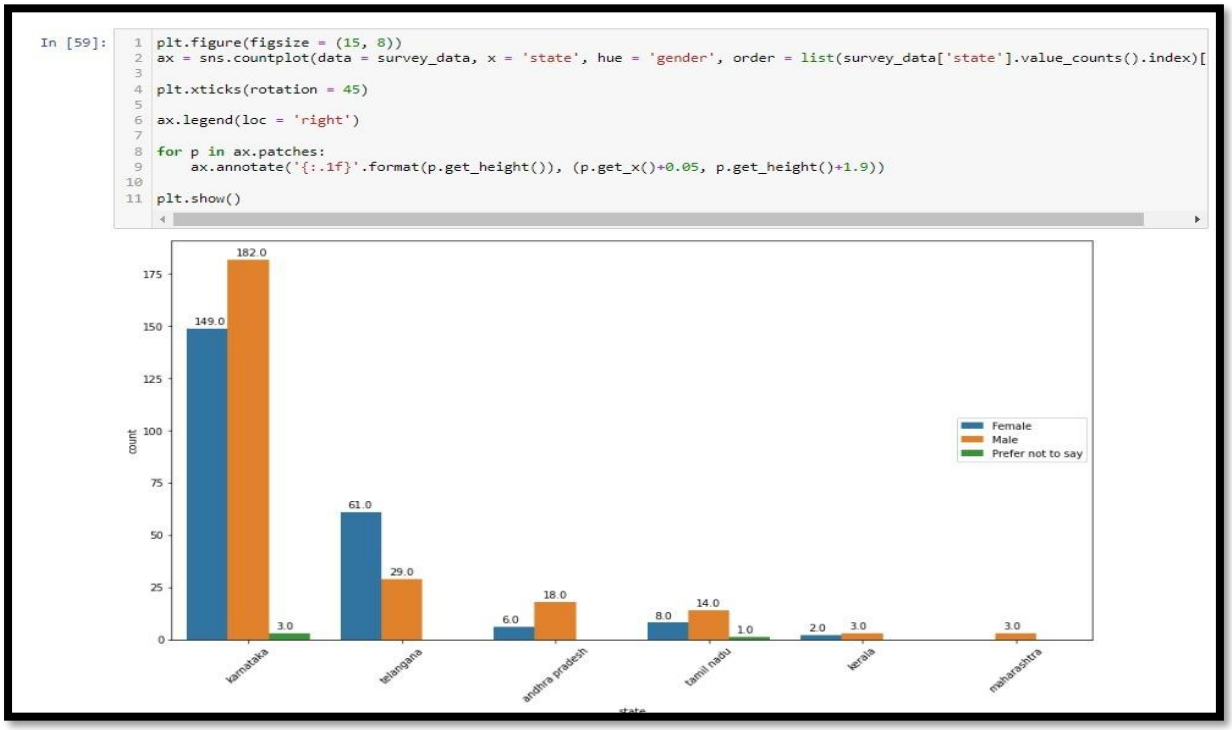

Figure 9.12: States Based on Gender (Top Models)

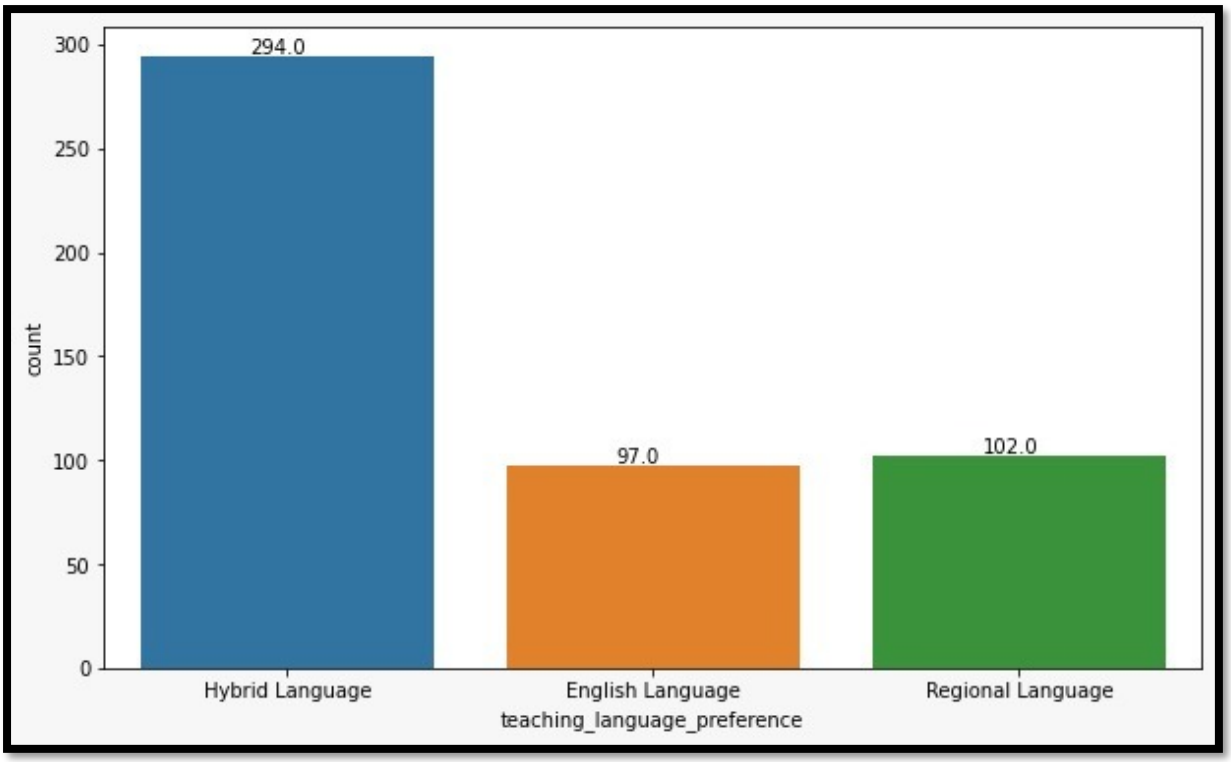

Figure 9.13: Teaching Language Preference

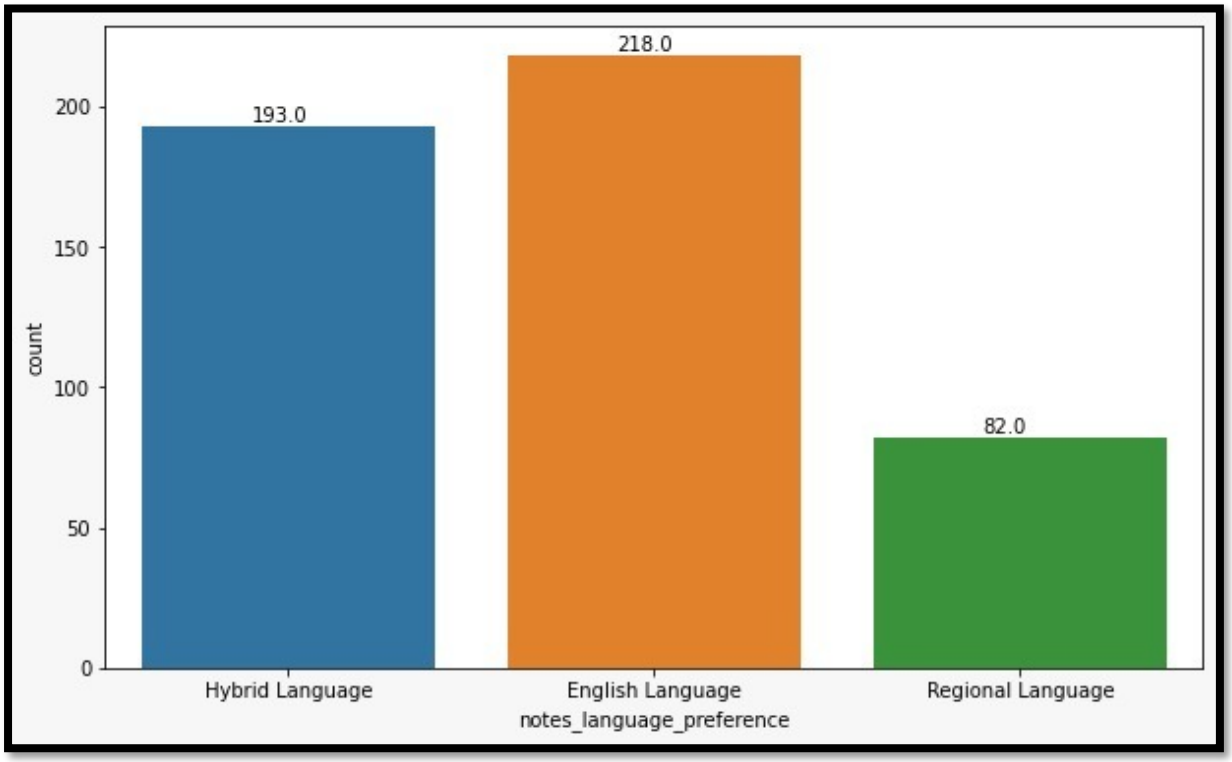

Figure 9.14: Notes Language Preference

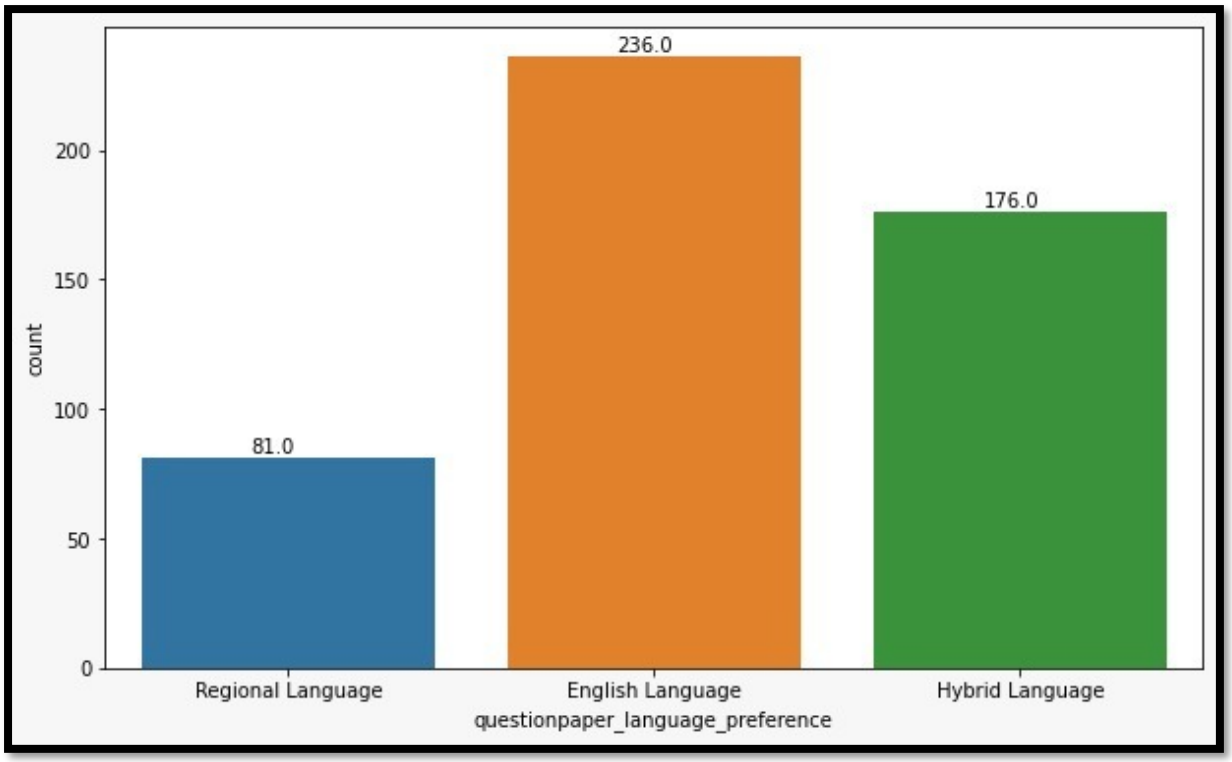

Figure 9.15: Question Paper Language Preference

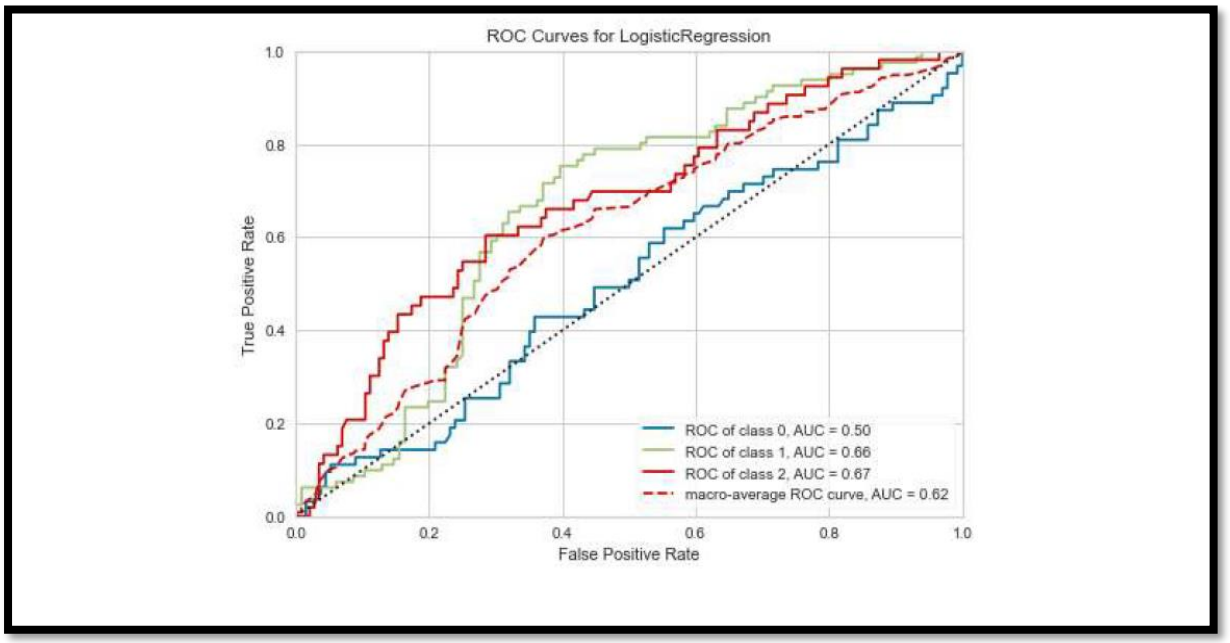

Figure 9.16: ROC Curve for Logistic Regression

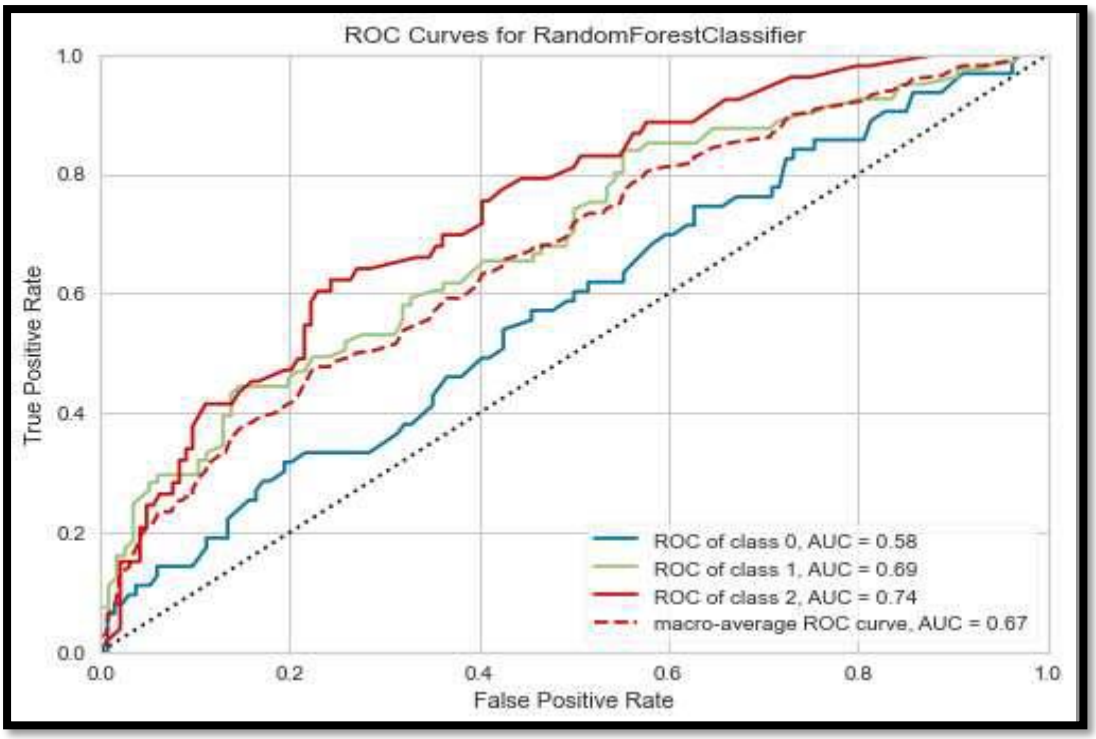

Figure 9.17: ROC Curve for Random Forest Classifier

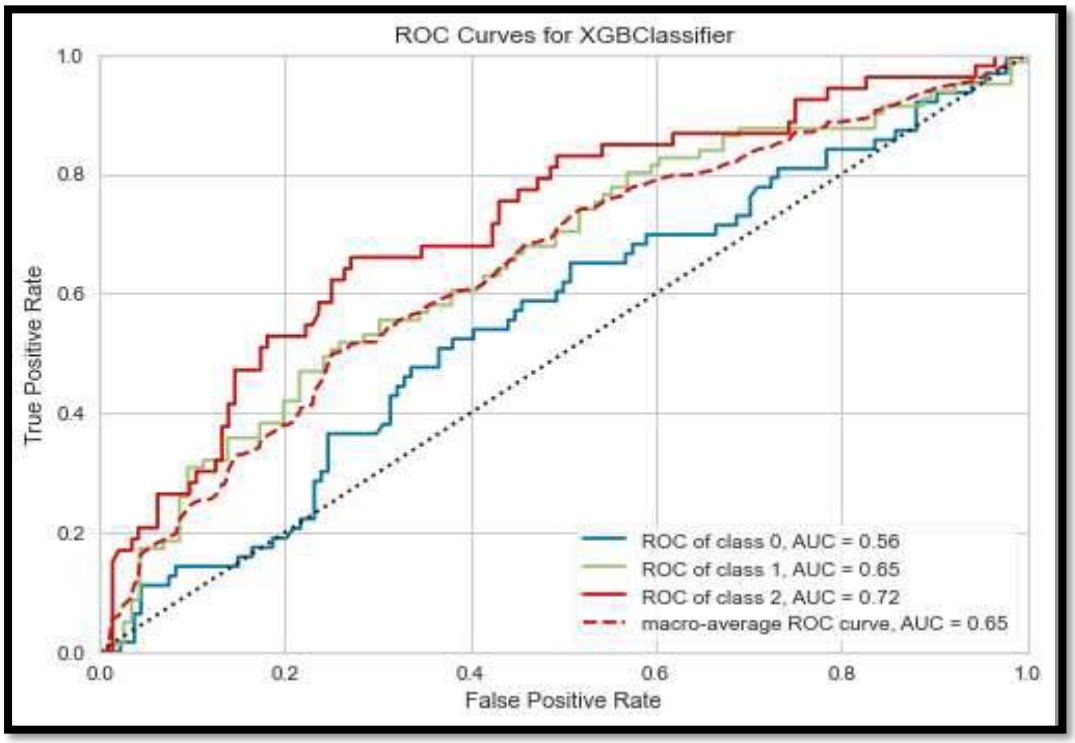

Figure 9.18: ROC Curve for XGB Classifier

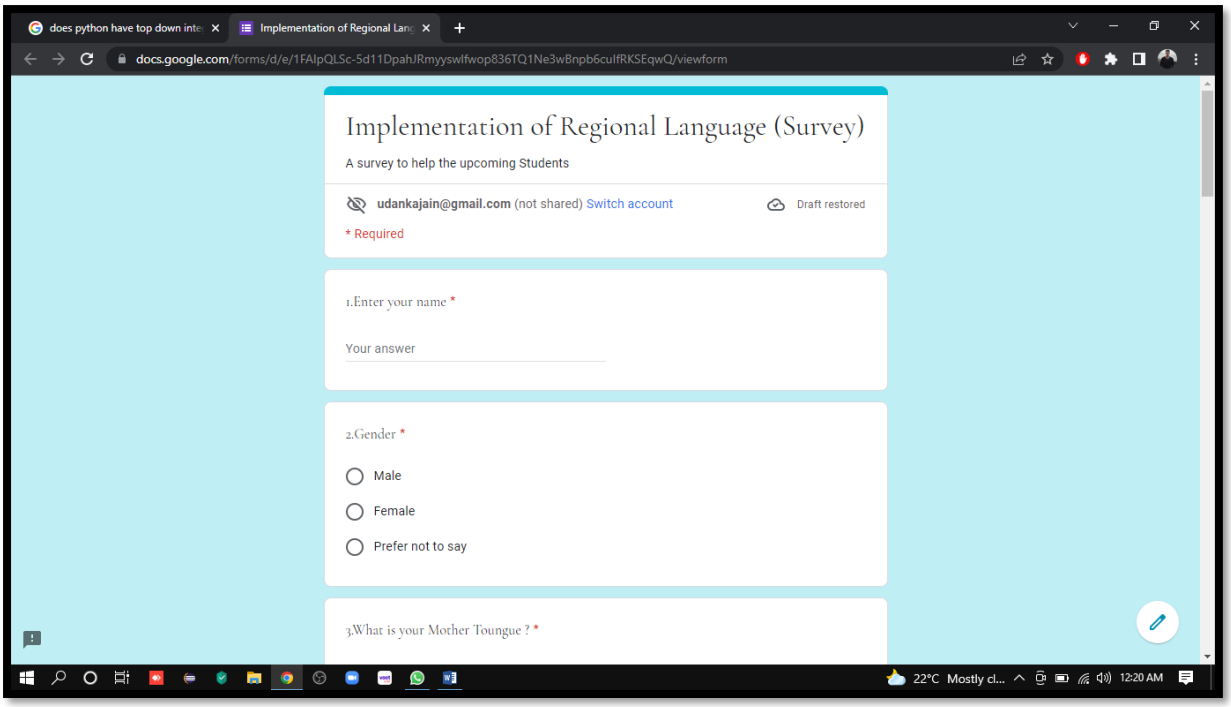

Figure 9.19: Snapshot 1 from Survey

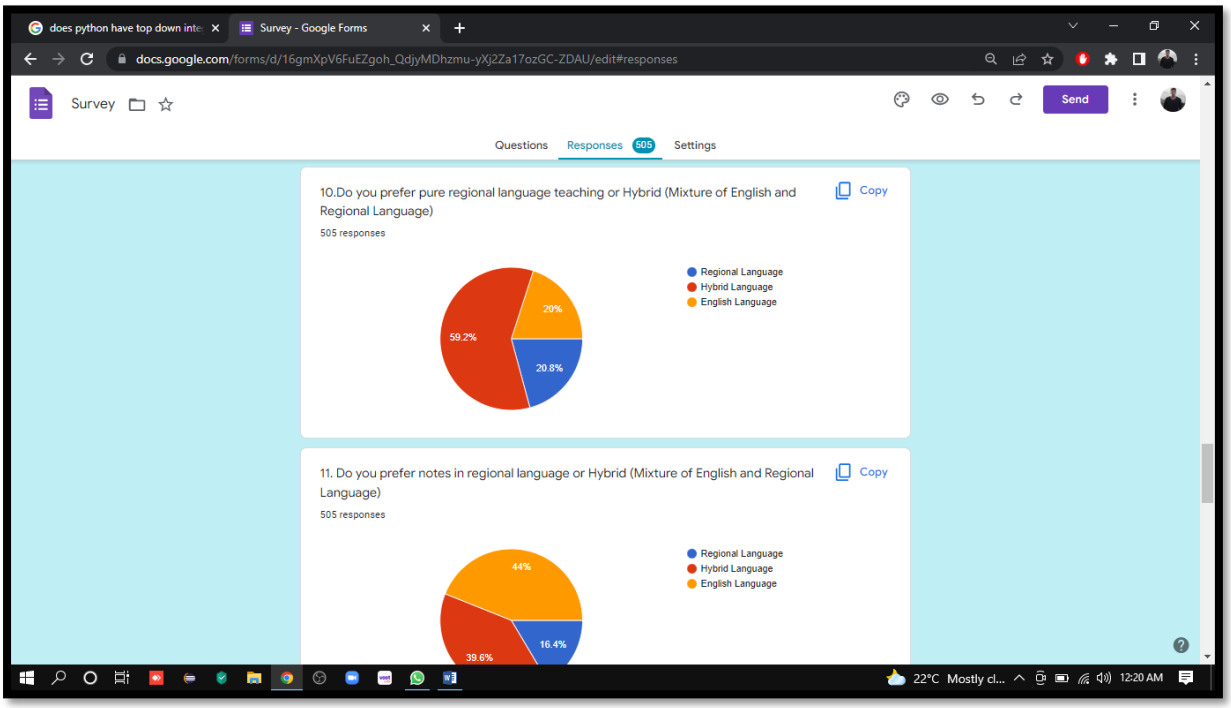

Figure 9.20: Snapshot 2 from Survey

CONCLUSION

Implementation of the project was done by the process of conducting a nationwide survey, this survey contained the set of questionnaires which was been sent to students, teachers and other important academic heads.

This questionnaire was nationwide so that we would get the pulse of all the types of people in the whole nation and hence we got the accurate result using these survey outputs. later the following data was collected and then the set of algorithms were written which included the following algorithm like K Means, KNN, Linear Regression, Random Forest which then was used to process the data and then this data was trained for the system and testing of the new data was prepared to get the following results.

Then the following processed dataset was taken in account of predicting the results and hence the data was then published for the user to concluded so that the following implementation to conclude if the survey is accountable one or not. By this way the project was concluded and the data can also be given to the following academic planners for planning out on giving an effective education and can further be planned by the officials on the following implementations.

FUTURE ENHANCEMENT

With advancements in technology, surveying equipment and techniques are developing. Current advancements are making the science of surveying more valuable, accurate, and comprehensive than ever. Our project deals with the topic of implementation of regional languages in a professional degree course.

Here the survey was conducted to in order to understand the mindsets of people all around the nationwide, we took minimal amount of responses to show the implementation generally whereas this can be further used by the universities in order of collecting huge amount of sample of responses which would in order give them the idea about the mindset of people.

The project we executed was basically of a general purpose but in future universities can utilize them with a huge number of responses to know the people mind set weather they choose regional languages over English language or they choose hybrid languages. In the future, portal can be created where in the data can be processed in real time to give out results that whether the student can opt for Regional Language based the entries that they have entered.

REFERENCE

[1] **Analysis of computer science based on big data mining by Liu Xuan, Liu Chang in the year 2020. At Huazhong University of science and Technology library, Wuhan**

[2] **An overview on machine learning technologies and their use in E-learning by Ramzi Farhat, Yosra Mourli, Mohamed Jemni, Houchine Ezzedine in the year 2020. At Latice research laboratory university Tunis,Tunisia, University Polytechnique Hauts-de-France , Valencinnes , France**

[3] **Anomaly Detection by Using Streaming K-Means and Batch K-Means by Zhuo Wang, Yanghui Zhou, Gangmin Li in the year 2020 at 2020 5th IEEE International Conference on Big Data Analytics**

[4] **Improved random forest classification approach based on hybrid clustering selection by Dong Yuan, Jian Huang, Xu Yang, Jiarui Cui in the year 2020 at 2020 Chinese utomation Congress**

[5] **Approach to Determining the Boundaries of the Search Range for the Number of Trees in the Random Forest Algorithm by Liliya Demidova, Maria Ivkina in the year 2020 at 2020 9th MEDITERRANEAN CONFERENCE ON EMBEDDED COMPUTING, BUDVA, MONTENEGRO**

[6] **Optimization of regression algorithms using learning curve in wsn by Vivek Kumar Verma, Vinod Kumar in the year 2021 at ABES engineering college Ghaziabad, UP, India and SRM institute of science and technology**

[7] **http://scikitearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html**

[8] **https://www.scikit-yb.org/en/latest/**

[9] **https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5?gi=9aabf4f41399**

[10] **https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62**

[11] **https://xgboost.readthedocs.io/en/stable/python/python_api.html**

[12] Data Science from Scratch by Joel Grus

[13] https://www.sciencedirect.com/topics/computer-science/logistic-regression

[14] https://matplotlib.org/

[15] https://www.w3schools.com/python/matplotlib_pyplot.asp

[16] https://stackoverflow.com/questions/34093264/python-logistic-regression-beginner

[17] Hands-on ML with Scikit-Learn, Keras & TensorFlow
Author – Aurélien Géron , Edition – Second Edition, Publisher – O'Reilly Media, Inc.

[18] Introduction to Machine Learning with Python
Author – Andreas C. Müller, Sarah Guido, Edition – First Edition, Publisher – O'Reilly Media, Inc

[19] Python for Data Analysis
Author – Wes McKinney, Edition – Second Edition, Publisher – O'Reilly Media, Inc

[20] Machine Learning with Random Forests And Decision Trees: A Mostly Intuitive Guide, But Also Some Python
by Scott Hartshorn (Goodreads Author)

DECLARATION

We do here by solemnly affirm to declare that, the contents of the project report submitted (Like the idea, concept, diagram, figures, video, etc) to the department of CSE, B.E whichever is of City engineering college Bangalore, affiliated to VTU Belgaum, Approved AICTE-new Delhi, is in majority the courtesy by the referred "National/International journals, Text-books, papers presented and published" by the learned authors in the area of their research and also available in the public domain. We sincerely acknowledge all of them. This project report has been submitted by us to satisfy the academic requirement affiliating university for the award of Bachelor Degree in the discipline of our study.

KARTHIK A N 1CE18CS030

R LAKSHMI SAI CHETANA NATH 1CE18CS061

SURABHI G R 1CE18CS084

UDANKA AARUNJAIN 1CE18CS090