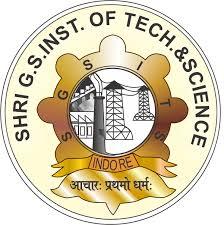
**Wikipedia Participation Prediction**



A project report submitted to

Rajiv Gandhi Proudyogiki Vishwavidhyalaya, Bhopal

Towards the partial fulfillment of

The degree of Bachelor of Engineering In

Computer Engineering

**Guided by Submitted by**

Dr.Vandan Tewari Pooja Choudhary(0801CS151053)

Roshni Choungad(0801CS151063)

Shruti Nagraj(0801CS151071)

**Co.Guided by** Shweta Kerketta(0801CS151076)

Prof. Shweta Dubey Nikita Varfa(0801EI151037)

**Computer Engineering Department**

**Shri. G. S. Institute of Technology and science**

**Indore(M.P)**

**2018-2019**

**Shri. G. S. Institute of Technology and science**

**Indore(M.P)**

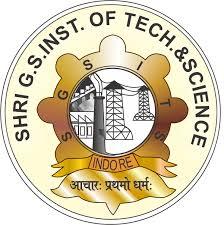
**CERTIFICATE**

This is to certify that the project report entitled “ Interactive Tutor For Data Structures” submitted by Pooja Choudhary(0801cs151053),Roshni choungad(0801cs151063), Shruti Nagraj(0801cs151071) ,Shweta kerketta(0801cs151076), Nikita varfa(0801EI151037) students of B.E. final year of this institute in the session 2018-2019, towards the partial fulfillment for the award of degree of **Bachelor of Engineering in Computer Engineering** of Ragiv Gandhi Proudyogiki Vishwavidyalaya,Bhopal,is a satisfactory account of their work based on syllabus and is approved for the award of the degree.

**Internal Examiner External Examiner**

**Shri. G. S. Institute of Technology and science**

**Indore(M.P)**



**DECLARATION**

We , Pooja Choudhary,Roshni choungad, Shruti Nagraj ,Shweta kerketta, Nikita varfa hereby declare that the work presented in BE project report has been carried out by us.We further declare that the best of our knowledge the report doesn’t contain any part of the work, which has been submitted for the award of any degree either in this university,without proper citation.

**Date: 19/11/2018** Pooja Choudhary

Roshni Choungad

Shruti Nagraj

Shweta Kerketta

Nikita Varfa

**ACKNOWLEDGEMENT**

It’s a privilege to have been a student of the **Computer Engineering** stream in Shri G.S. Institute of Technology and science, Indore.

We would like to express our greatest gratitude and respect to our project guide **Dr.Vandan Tiwari** and co. Guide **Prof.Shweta Dubey**,for his excellent guidance, valuable suggestions and endless support. Without his guidance, this work wouldn’t have been possible.

We also thank to **Prof. K.P. Singh** for his guidance and support to complete this work.

We sincerely wish to express our gratefulness to all members of staff of Computer Engineering Department who have extended their cooperation at all times and have contributed in their own way in developing the project.

The successful completion of project work is not an individual effort. It is an outcome to the cumulative effort of a number of people, each having their own importance to tha objective. We express love and respect towards our parents and all family members who are our strength in everything we do.

With a blend of gratitude, pleasure and great satisfaction we convey our indebtedness to all those who have directly or indirectly contributed to the successful completion of our project work.

Pooja Choudhary

Roshni Choungad

Shruti Nagraj

Shweta Kerketta

Nikita Varfa

**Chapter 1**

**INTRODUCTION**

This chapter gives a basic outline of the project describing the problem statement, the objectives, the need of our project, a brief description of our solution approach and the details about the organisation of the project report.

**1.1 Preamble**

This project is intended to build a predictive model which can predict the number of editors that will stop editing in the future. In this project, likelihood of future outcomes based on historical data will be identified. In addition, this model can also demonstrate visualization of editing behavior of editors. Hence the project provides following functionality:

* Provide complete visualization of number of editors who edits frequently.
* Provide quantitative understanding of the factors that determine editing behaviour of the editor.
* Provide visualisation of number of editors who will stop editing, on the basis of their editing behaviour.

**1.2 Need of the Project**

WIKIPEDIA has come a long way since it started in 2001. With around 70,000 volunteers editing in over 100 languages, it is by far the world’s most popular reference site. Its future is also uncertain. One of the biggest threats it faces is the rise of smartphones as the dominant personal computing device. This is a challenge for Wikipedia, which has always depended on contributors hunched over keyboards searching references, discussing changes and writing articles using a special markup code. Even before smartphones were widespread, studies consistently showed that these are daunting tasks for newcomers. The difficulty of bringing on new volunteers has resulted in seven straight years of declining editor participation.

In 2005, during Wikipedia’s peak years, there were months when more than 60 editors were made administrator — a position with special privileges in editing the English-language edition. For the past year, it has sometimes struggled to promote even one per month.

The need of the proposed system is to predict the number of editors who edits frequently. The main focus of this project is to understand the editing behaviour of the editors which leads to reduction in editing, which helps to analyze the future of one of the world's biggest websites. Moreover, Wikipedia's freedom will grow only more important as the content grows in breadth, depth, and quality. Thus, identifying frequent editors will help to promote more administrators in order to achieve depth and quality of the content.

**1.3 Problem Statement**

Wikipedia is losing editors, while the number of internet users continues to grow.People stop editing after the initial phase of Wikipedia editing. The study led by Aaron Halfaker of the University of Minnesota found that the number of "[collaborators](https://phys.org/tags/collaborators/)" or volunteer [editors](https://phys.org/tags/editors/) has been on the decline from around 56,000 in 2007 to some 35,000 at the end of 2012.

So there is need to predict number of editors who will stop contributing which will help in overcoing problem

**1.4 Objectives**

* The main objective of Wikipedia Participation Prediction is to build a predictive model which can predict the future outcome of the editors who will stop contributng to Wikipedia. The prediction will be based on historical data. The goal is to go beyond knowing what has happened to providing a best assessment of what will happen in the future.
* The project’s objective is to quantitatively understand the factors which determines editing behaviour of the editor. The most common factor by which editor’s behaviour can be determined is edit count. The other factors such as number of bytes and edit history will also help in measuring the work of the editor.
* On the basis of editors’ editing behaviour, prediction of number of editors that will stop contributing can be made.
* The predictive model must be built which can operate on the input dataset and make the output process automated.

**1.5 Solution Approach**

**1.5.1 Quantitatively understand what factors determine editing behavior of the editor**.

It is important to know the factors which will determine editing behavior of the editor. It will predict about number of editors who will stop contributing. Factors determining the behaviour of the editor are:

a. Number of total edits of user till date

b. Number of bytes edited.

c. Frequency of editing

**1.5.2 Data preprocessing**

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.As data is main concern of project so it is important to clean it and extract important information which will be needed for the project.

Steps followed in preprocessing are

**Step 1 :** Import the libraries

Some of important libraries which will be needed for project.

**Pandas**is fordata manipulation and analysis*.* Pandasis an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the python programming language.

**Matplotlib** is a Python 2D plotting library which produces publication quality figures in a variety of hard copy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells ,the Jupyter notebook.

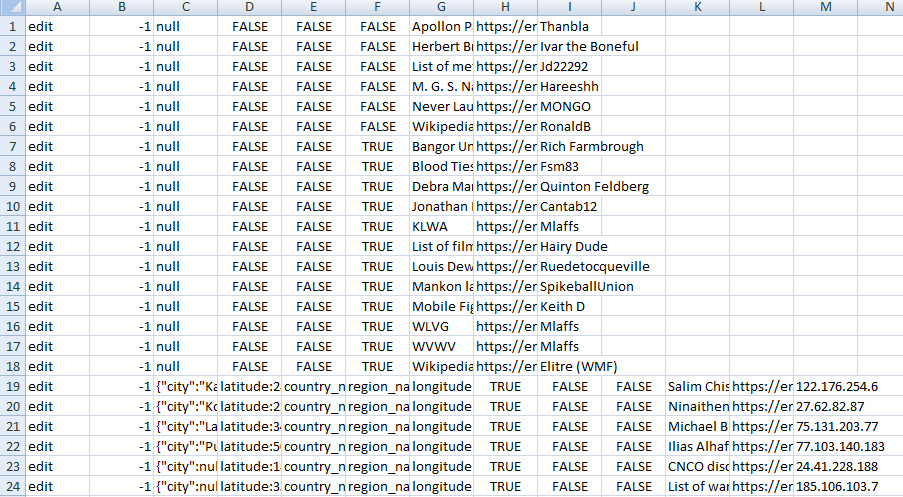
**Seaborn**is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphic

**Step 2 :** Import the data-set

By using Pandas we import our data-set and the file I used here is .csv file. However, to access and to use fastly CSV files are used because of their light weights. After importing the dataset head function ( This function returns the first n rows for the object based on position. It is useful for quickly testing if your object has the right type of data in it. By default it returns 5 rows. ) is used.

**Step 3 :** Check out the missing values

The concept of missing values is important to understand in order to successfully manage data. If the missing values are not handled properly by the researcher, then he/she may end up drawing an inaccurate inference about the data. Either we can delete whole row containing null value or we can calculate mean, median and mode and replace it with missing values



**1.5.3 Build a suitable model**

Data mining technique will be used for this project. As it is required to precisely predict the count of editors who will stop contributing in future, so specifically predictive modeling will be used.

**Predictive Modelling**

Predictive modeling is a process that uses  and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results. Once data has been collected for relevant predictors, a statistical model is formulated. The model may employ a simple linear equation, or it may be a complex neural network, mapped out by sophisticated software. As additional data becomes available, the statistical analysis model is validated or revised.

**Modelling Methods**

Once the sample data has been gathered, the right model must be selected. Different categories of predictive models are:

1.Linear Regression

 Linear regressions are among the simplest types of predictive models. Linear models essentially take two variables that are correlated -- one independent and the other dependent -- and plot one on the x-axis and one on the y-axis. The model applies a best fit line to the resulting data points. Generally this is used to predict future occurrences of the dependent variable.

2.Decision Trees

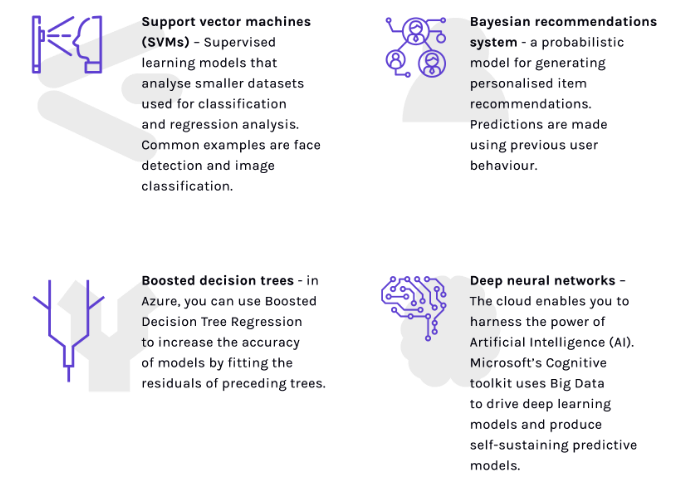
A decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, its also widely used in machine learning.

3.Bayesian Classification

Bayesian classification is based on Bayes' Theorem. Bayesian classifiers are the statistical classifiers and can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

4. Neural Network

The most complex area of predictive modeling is the neural network. This type of machine learning model independently reviews large volumes of labeled data in search of correlations between variables in the data. It can detect even subtle correlations that only emerge after reviewing millions of data points. The algorithm can then make inferences about unlabeled data files that are similar in type to the data set it trained on.



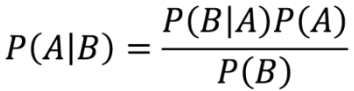
Among these four predictive models, Bayesian Classifiers are to be used in this project.

**Bayesian Classifiers**

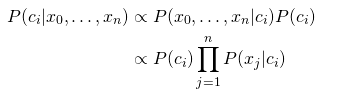
Naïve Bayes is a simple, yet effective and commonly-used, machine learning classifier. It is a probabilistic classifier that makes classifications using the Maximum A Posteriori decision rule in a Bayesian setting. It can also be represented using a very simple Bayesian network. Naive Bayes classifiers have been especially popular for text classification, and are a traditional solution for problems such as spam detection.

The goal of any probabilistic classifier is, with features x\_0 through x\_n and classes c\_0 through c\_k, to determine the probability of the features occurring in each class, and to return the most likely class. Therefore, for each class, we want to be able to calculate P(c\_i | x\_0, …, x\_n).

In order to do this, we use Bayes rule, Recall that Bayes rule is the following:

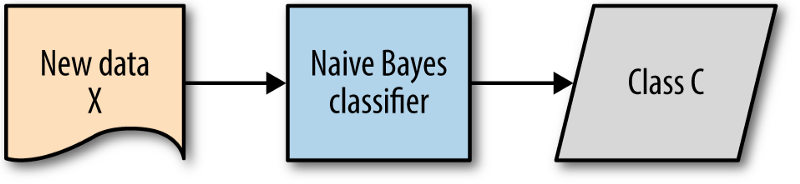


In the context of classification, you can replace A with a class, c\_i, and B with our set of features, x\_0 through x\_n. Since P(B) serves as normalization, and we are usually unable to calculate P(x\_0, …, x\_n), we can simply ignore that term, and instead just state that P(c\_i | x\_0, …, x\_n) ∝ P(x\_0, …, x\_n | c\_i) \* P(c\_i), where ∝ means “is proportional to”. P(c\_i) is simple to calculate; it is just the proportion of the data-set that falls in class i. P(x\_0, …, x\_n | c\_i) is more difficult to compute. In order to simplify its computation, we make the assumption that x\_0 through x\_n are conditionally independent given c\_i, which allows us to say that P(x\_0, …, x\_n | c\_i) = P(x\_0 | c\_i) \* P(x\_1 | c\_i) \* … \* P(x\_n | c\_i). This assumption is most likely not true — hence the name *naive* Bayes classifier, but the classifier nonetheless performs well in most situations. Therefore, our final representation of class probability is the following:

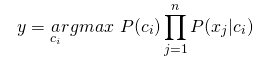


Calculating the individual P(x\_j | c\_i) terms will depend on what distribution your features follow. In the context of text classification, where features may be word counts, features may follow a multinomial distribution. In other cases, where features are continuous, they may follow a Gaussian distribution.

Note that there is very little explicit training in Naive Bayes compared to other common classification methods. The only work that must be done before prediction is finding the parameters for the features’ individual probability distributions, which can typically be done quickly and deterministically. This means that Naive Bayes classifiers can perform well even with high-dimensional data points and/or a large number of data points.



Now that we have a way to estimate the probability of a given data point falling in a certain class, we need to be able to use this to produce classifications. Naive Bayes handles this in a very simple manner; simply pick the c\_i that has the largest probability given the data point’s features.



This is referred to as the Maximum A Posteriori decision rule. This is because, referring back to our formulation of Bayes rule, we only use the P(B|A) and P(A) terms, which are the likelihood and prior terms, respectively. If we only used P(B|A), the likelihood, we would be using a Maximum Likelihooddecision rule.

**1.6 Organization of the report**

The project report includes five chapters.

Chapter 1 provides the introduction about the project, need of the project, problem statement, objectives and solution approach.

Chapter 2 discusses the background of the project.

Chapter 3 equips with holistic analysis of the project which includes the detailed problem statement, requirement analysis i.e. functional and non-functional requirements followed by feasibility study.

Chapter 4 deals with design aspect of the project which includes the architectural diagram such as use case diagram.

Chapter 5 cites the conclusion deduced after the detailed study of the project work and also equips the information regarding future work.

**Chapter 2**

**BACKGROUND**

Wikipedia, one of the top ten most visited websites, is commonly viewed as the largest online reference for encyclopedic knowledge. Because of its open editing model -allowing anyone to enter and edit content.

**2.1 Literature Review**

The sixth most widely used website in the world is not run anything like the others in the top 10. It is not operated by a sophisticated corporation but by a leaderless collection of volunteers who generally work under pseudonyms and habitually bicker with each other. Anyone who visits the site can edit it, and this fact has encouraged contribution of a tremendous amount of content. Because there is no other free information source like it, many online services rely on Wikipedia.

When Wikipedia launched in 2001, it wasn’t intended to be an information source in its own right. Wales and Larry Sanger started the site to boost Nupedia, a free online encyclopedia started by Wales that relied on contributions from experts. After a year, Nupedia offered a strange collection of only 13 articles on such topics as Virgil and the Donegal fiddle tradition. Sanger and Wales hoped Wikipedia, where anyone could start or modify an entry, would rapidly generate new articles that experts could then finish up.

Newcomers to Wikipedia making their first, tentative edits—and the inevitable mistakes—became less likely to stick around. Being steamrollered by the newly efficient, impersonal editing machine was no fun. The number of active editors on the English-language Wikipedia peaked in 2007 at more than 51,000 and has been declining ever since as the supply of new ones got choked off. This past summer only 31,000 people could be considered active editors. Even though Wikipedia has far fewer active editors than it did in its heyday, the number and length of its articles continue to grow. This means the volunteers who remain have more to do. A 2011 survey by the Wikimedia Foundation suggested that being an active editor already required a significant time commitment. Of 5,200 Wikipedians from all language editions of the project, 50 percent contributed more than one hour a day, and 20 percent edited for three or more hours a day.

**2.2 Tools and Technology**

**2.2.1 Jupyter Notebook**

Jupyter Notebook is an open-source web application that allow to create and share codes and documents.

It provides an environment, where the code can be documented, run, look at the outcome, visualize data and see the results without leaving the environment. This makes it a handy tool for performing end to end data science workflows –data cleaning, statistical modeling, building and training machine learning models, visualizing data, and many, many other uses. The code is written in indepedent cells, which are executed individually. This allows the user to test a specific block of code in a project without having to execute the code from the start of the script.

**Chapter 3**

**Analysis**

This chapter describes the problem statement in detail followed by the functional and non-functional requirements identified during the analysis phase. It also contains the use-case diagram that aids in the proper implementation of the model.

**3.1 Detailed Problem statement**

Wikipedia, one of the world's biggest websites, is losing many of its English-language editors, crippling its ability to keep pace with its mission as a source of knowledge online, a study says.  
The study led by Aaron Halfaker of the University of Minnesota found that the number of "[collaborators](https://phys.org/tags/collaborators/)" or volunteer [editors](https://phys.org/tags/editors/) has been on the decline from around 56,000 in 2007 to some 35,000 at the end of 2012.

The researchers said there are a number of reasons, including the rise of automated programs or "bots," but also noted that some potential contributors are being discouraged by Wikipedia's structure.

"Several changes the Wikipedia community made to manage quality and consistency in the face of a massive growth in participation have ironically crippled the very growth they were designed to manage”,  the researchers wrote in American Behavioral Scientist.

"Specifically, the restrictiveness of the encyclopedia's primary quality control mechanism and the algorithmic tools used to reject contributions are implicated as key causes of decreased newcomer retention."  
They said that while Wikipedia has sought to root out less competent editors, its rules have also discouraged "desirable newcomers" who get discouraged when their contributions get deleted.

Wikipedia has editions in 285 languages. Its founder Jimmy Wales has cited the need to make Wikipedia more open to newcomers, to keep up with the vast amount of information it is trying to process.

But the authors of the study said Wikipedia is being crimped by its rules trying to improve quality.

**3.2 Requirement Analysis**

For the successful completion of any project implementation. Requirement analysis is must;

Much of the user’s satisfaction depends on this phase of the object deployement,if the requirement are gathered and analyzed properly , then final product will come up to the expectation of the users.

**3.2.1 Functional requirement**

3.2.1.1 Provide complete visualization of number of editors who edits frequently.

This feature will be determined with the help of factors like how much bytes the user has added or updated and the frequency of editing.

3.2.1.2 Provide quantitative understanding of the factors that determine editing behavior of the editor.

Factors determining the behavior of the editor are:

a.Number of total edits of user till date.

b.Number of bytes edited.

c. Frequency of editing.

3.2.1.3Provide visualization of number of editors who will stop editing on the basis of their editing behavior.

The editing behavior determined from the above three factors will be used to predict the count of editors that will stop contributing in future.

**3.2.2. Non-functional Requirements**

**i. Reliability-** The system should give the less or completely zero opportunities to have human errors, which itself will make system reliable.

**ii.** **Availability** – It will make sure that the system is available to the end user 24x7 and in the required form.

**iii. Maintainability** –The attribute will provide all kinds of required technical, documentation assistance to the end user so that the system can be easily maintained and work in a proper condition.

**iv. Performance**-Checking the fact that the system must perform as per the desirability. So in every action-response of the system, there are no immediate delays.

**3.3 Feasibility study**

**3.3.1. Economic feasibility**

a. the project only requires a computer to load the particular database.

b. also requires an internet connection if the model is accessible from particular network with the software. Hence model is economically feasible.

**3.3.2 Operational feasibility**

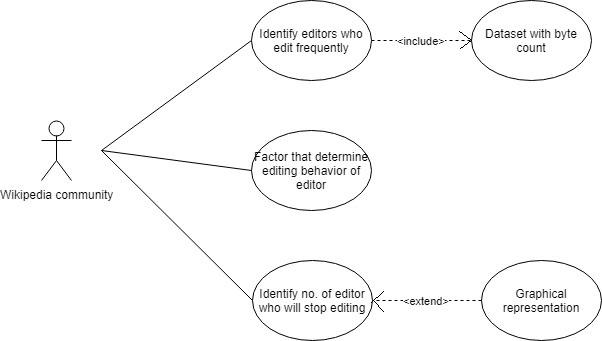
As the model is required to predict the precise count of editors that might leave contributing in future. So Operational feasibility is a measure of how well the model will predict the future outcomes.

**3.3.3 Technical feasibility**

It considers the technical requirements of the proposed model. This project needs the proper knowledge of working with datasets, Bayesian Classification and predictive modeling. Python will be used in this project for dealing with datasets.

**Chapter 4**

**USE CASE DIAGRAM**

****

Usecase Description

1. Identify editors who edit frequently

This identification will be done by analysing the byte count i.e. how many bytes a user is editing on Wikipedia and that too in what period of interval.

1. Factors that determine editing behavior of editor

These factors are total number of edits, byte count and frequency of editing. These will be determined by preprocessing of the dataset.

1. Identify number of editor who will stop editing

This count will be determined by proper analysis of the above three factors. If the editor’s frequency of editing is extremely less or the bytes he/she is editing in a single edit is very small or its total count of edits is less, by combining these three aspects the editors will be classified into categories.