International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V68I12P111

Volume 68 Issue 12, 47-63, December 2020

Review Article

Predictive Analysis As a Solution To Branding

Hari Purnapatre[1],Satvik Shukla[2] ,Srishti Pareek[3]

[1,2] *Pune Institute of computer technology, Survey No. 27, Dhankawadi, Pune, Maharashtra 411043*
[3] *S. No. 44/1, Off. Sinhgad Road, Vadgaon Budruk, Pune, Maharashtra 411041.*

Received Date: 14 October 2020
Revised Date: 22 November 2020
Accepted Date: 24 November 2020

Abstract *- Personal branding offers promises of increased success in the business world. Personal branding is how a brand presents itself, so people remember it. Chatbots are expected to become as common as business phone numbers and with increasing intelligence. So, they will replace human assistance entirely. Moving towards a more conversational way of doing things. Chatbot, which will answer all the queries related to startup or building connections around. Hence, this project will help develop a startup or a freelancer's overall development in whether its suggesting strategies, optimizing own image, or building a network around.*

Keywords *- Predictive analysis, data mining, machine learning, chatbot.*

I. INTRODUCTION

Personal branding is the discovery, understanding, and marketing of an individual's unique attribute. Personal branding is needed to provide greater focus and direction in career/business, help differentiate from competitors, help stand out from the crowd. In this project, we plan to make a chatbot that will engage customers in a meaningful business to business conversation. Chatbot will answer all the queries related to startup or building connections around. Chatbot can be used by a person to find a suitable domain company of his/her interest. Chatbots are algorithmic conversational agents that companies are coming up with to interact with their customers. Siri or Google assistant's limited functionality has been a big barrier for chatbots, and Machine Learning is an attempt to target this. With experience, the chatbot learns responses to different queries and hence increases its functionality. The use of machine learning thus seems very natural for chatbots. Machine Learning is a branch of AI based on the idea that machines should be able to learn and adapt through experience. It is a method of data analysis that automates analytical model building. In simple words, it ensures that computers can learn through experience. This simplicity has attracted different experts, and wonderful work is being witnessed in this direction. The fact that human behavior is largely repetitive makes it possible to invest and work on the concept of machine learning for chatbots. Machine Learning has enabled the chatbots to converse intelligently and with the added advantage of Natural Language Understanding & Processing, making it easier for chatbots to understand human language flow. They have been really good in suggesting information that a human requires, and the magic of NLP combined with ML has been evident in the development of chatbots. It uses machine learning algorithms, namely random forest, decision tree, K Nearest Neighbour, Naive Bayes. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to learn and improve from experience without being explicitly programmed automatically. Machine learning focuses on developing computer programs that can access data and use them to learn for themselves. Machine learning algorithms are often categorized as supervised or unsupervised.

II. RELATED WORK

For this project, an extensive literature survey of the problem was made to ensure that features (variables), which is the most important part of the machine learning task, are appropriate and discriminative. Datasets with various company domains, social presence, annual revenue, operation area were studied, and a single questionnaire was created. This questionnaire follows the original research questions used to find the best practices to develop and maintain a successful personal brand for its growth.

A. Data Collection :

Data collection for this study was made keeping the following questions in mind

- How would you, as an expert in your field, define and describe a personal brand?
- Please give an example of a successful personal brand.
- What is the most important aspect of a personal brand? How can they be used to improve one's career?

- How does one develop a personal brand? Please provide examples of tools and tactics that can be utilized in creating a personal brand.
- In what capacities would a personal brand be used? How can one use personal branding in a professional environment?
- What are the various methods to improve social presence online?

B. Data Analysis :

Due to the high usage rate, social networking sites have become a commercial area for many firms to use for data analysis. Machine learning approaches are useful for analyzing data to make meaningful results. Facebook data was used in the study. The data set given analyzed using machine learning approaches with Logistic Regression (LR), Random Forest (RF), and Adaboost (AB) algorithms.

C. Machine Learning Approaches :

This section provides LR, RF, and AB information from machine learning approaches used to analyze social networking data.

D. Logistic Regression :

LR is a classification approach that is effective for situations where variables do not always consist of quantitative values. It is preferred for binary and multiple classification approaches. The main goal is to determine the probability of obtaining another dependent variable by using independent variables. Whatever the values of the variables in the LR classifier, the result is between 0 and 1. For simplicity, it presents a practical approach, especially when analyzing big data.

E. Random Forest :

The Random Forest algorithm has been developed as a community learning method using decision trees. Each decision tree is constructed by applying a bootstrap sampling of the data. The feature size used for each node is randomly selected from among all features. To classify the data, each sample vector is decided on each tree of the forest. It is one of the preferred classifiers for classifying big data because of the simplicity and random choice.

F. Adaboost :

The AB performs the classification process by combining the weights of the feature classifiers created by training data. It is used with different classifiers. At the beginning of the training, equal weight is applied to each sample. It is multiplied by the weights given and the responses given by each classifier to all samples. In the end, the least faulty attribute classifier is chosen. The sample weights are updated by increasing the weights of the samples of the selected attribute classifier. With new weights, the new performance of each attribute classifier is determined. The main classifier is added by selecting the best performing classifier. This process continues as long as the error rate falls.

III. BRANDING MODEL

The datasets are collected from various sources such as dataworld.com and data.gov.uk. The datasets used in this project are startup datasets that include all the startups' data according to domain and area. A linked-in dataset includes information about employees and blogs dataset, which will have links to various blogs that suggest strategies.

A. Functional Requirements

1) Business Rules- A system that gives any government initiative feedback based upon the users' review in run time.
2) Transaction corrections, adjustments, and cancellations- Collect twitter data only related to the particular initiative or program asked by the user. Tweets have to be cleaned and pre-processed so that they can be directly fed to the system.
3) Administrative functions- Collecting tweets, Processing the raw data, Displaying the results to the user.
4) Authentication- Login ids for every registered user, Unique Twitter, Facebook Ids.
5) Authorization levels-
    - Admin
    - The user whom feedback will be given
    - Any unregistered user

B. Model Split-up

1) Data collection Module :

In this Module, we are going to collect data that will be needed in further processing. It includes gathering parameter rich data from various data sources. Performing extensive literature survey of the problem to ensure that features (or variables or predictors), which is the most important part of the machine learning task, are the relevant(or appropriate or useful or discriminative) ones. Datasets include online sources like dataworld.com, government websites, the company mostly startups, and employee datasets.

2) Chatbot Module :

The semantic analysis is performed with the help of linear models on vectorized texts such as multiclass

SVM. It is used to classify the user's input into slots and intent. The name-entity recognition slots interference task, slot-filling Conditional Random Field. (CRF).

3) Classification Module :

This Module includes various classification algorithms like naive Bayes studied to find which one best fits the dataset. Affinios AI-powered consumer intelligence platform is used for the further selection process. The accuracy of the algorithm is tested using weka classification. This Module is used to create a network with potential employees.

4) Network with Potential Employees :

The list of potential employees, according to the requirements, is specified by matching the intent and slot, which are basically parameters such as domain name and area matching rows and columns of the database. The output is the name and expected salary of the potential employees. The accuracy is achieved by WEKA classification, which uses naive Bayes classification.

5) Controlling Online Presence :

Various strategies will be suggested to gain more recognition in the market. An interaction matrix or rating matrix is used for this purpose. Along with that, algorithms like user-based KNN is used to find K-most similar users.

6) Business Growth :

This helps in achieving growth in the proposed business. Competitors based on locality are identified by algorithms, namely, Decision Tree and Random Forest. Revenue growth and marketing strategies are used to gaining recognition. The intent is the area where competitors need to figure out and the domain. The market share and revenue of each company will be present. This Module facilitates the startups to find their competitors and their respective market share so that startups can find their major competitors in the market and build strategies accordingly.

C. Analysis

Weka was exclusively built for machine learning for data mining and comprised various data preparation, classification, regression, clustering, association rules mining, and visualization. It is turning to be useful for data science researchers.

1) Decision Tree :

The decision tree is used in the third Module of the project, which is business growth. The user will enter the startup and area domain according to which subsets will be created, and further questions will be asked accordingly by the chatbot. In short, the chatbot drives the conversation to suggest the strategies or give the output such as data of competitors, etc.

2) Random Forest Algorithm :

It is used in the third Module, which is business growth, and it is like an extension of the decision tree algorithm, so this algorithm will be used to improve the accuracy of the data. Each decision tree's data is taken and the one with maximum votes from decision trees is selected because it produces output with the highest potential correct output.

3) Naive Bayes Algorithm :

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values. A list of probabilities is stored to file for a learned naive Bayes model. This includes:

- Class Probabilities: The probabilities of each class in the training dataset.
- Conditional Probabilities: The conditional probabilities of each input value given each class value.

It is used in the second Module of the project, which is Network with potential employees. In this Module, the dataset of candidates available for hire, e.g., LinkedIn dataset, will be stored at the back end and in the front end according to user inputs, the domain, location, salary, etc., potential employees will be shown. To improve and check the accuracy of WEKA. Classification software will be used.

4) K-Means Algorithm :

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data. The goal is to find groups in the data, with the number of groups represented by the variable K.It works iteratively to assign each data point to one of K groups based on the provided features.

The results of the K-means clustering algorithm are:

- The centroids of the K clusters, which can be used to label new data.
- Labels for the training data(each data point is assigned to a single cluster).

IV. RESULTS AND EVALUATIONS

Experiments

Till now, we have focused completely on the pre-processin

g of datasets. We tried various approaches to analyze data and determine the performance of each algorithm. The performance metrics used to analyze the data are precision,

recall, and F1 score metrics. The F1 Score metric is the harmonic mean of the precision and recall metrics.

A. Decision Tree

1) Dataset 1: Employee dataset

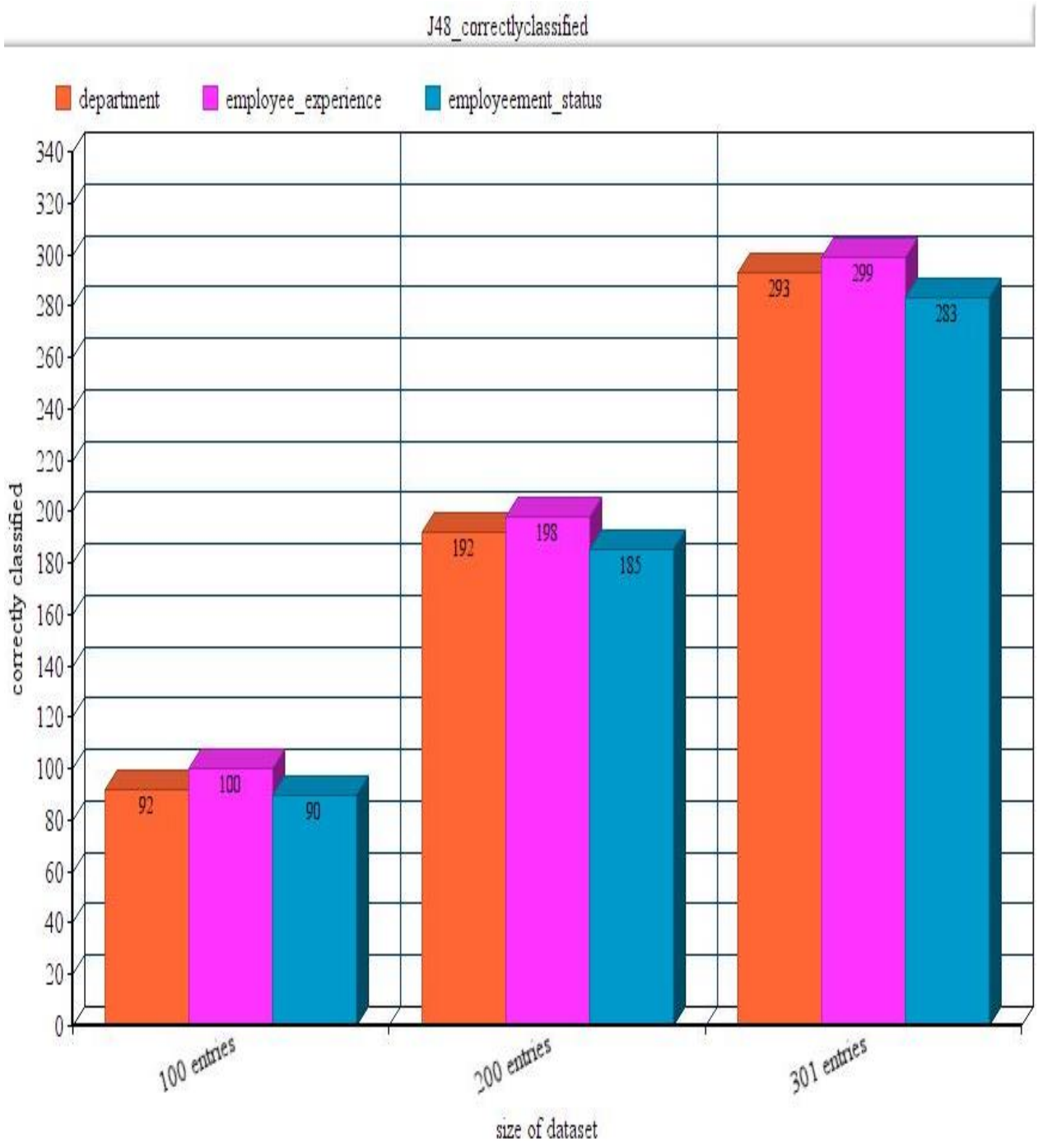

Fig1: Correctly classified instances

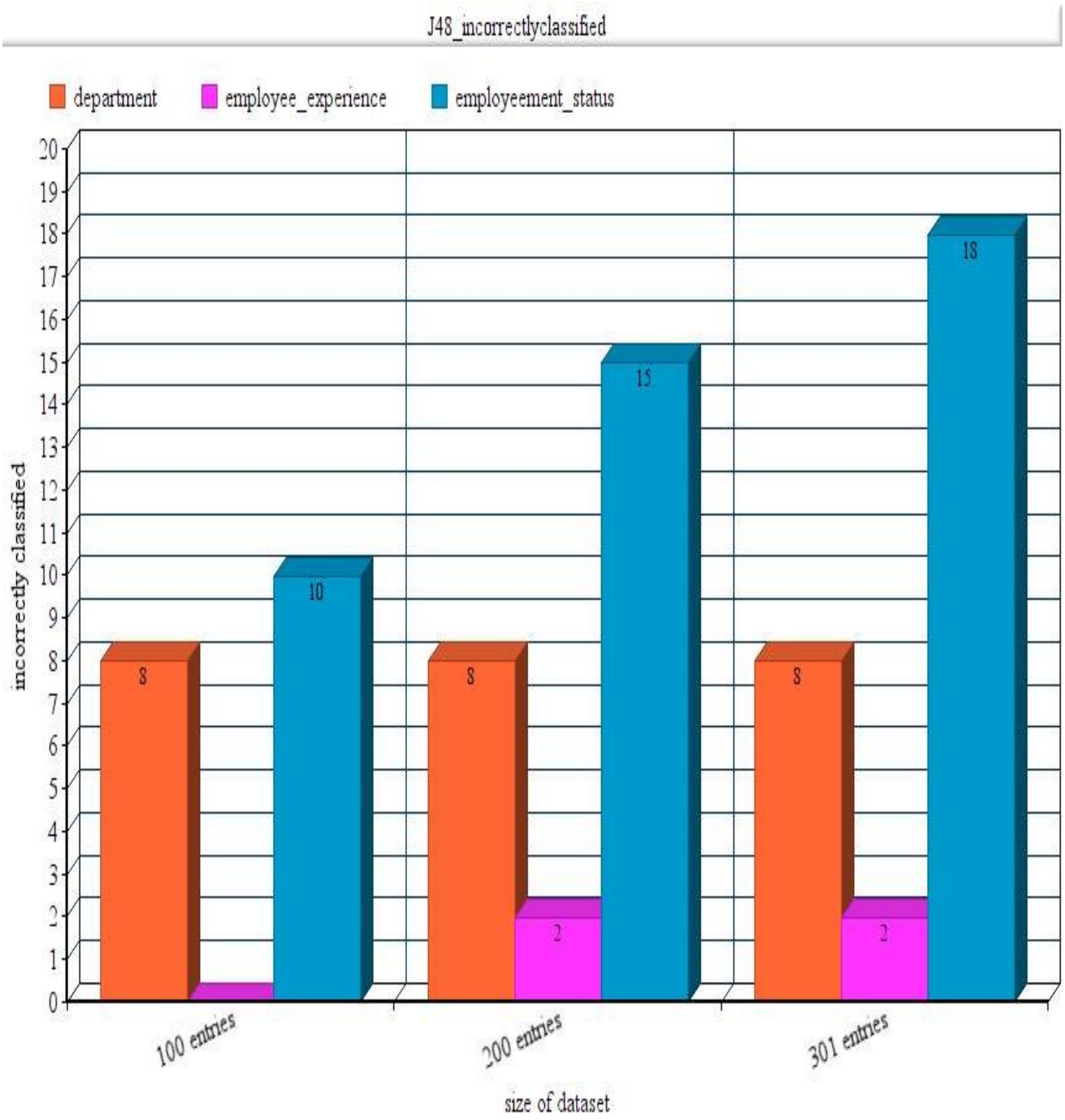

Fig 2: Incorrectly classified instances

This graph depicts the comparison between wrongly classified entries of different attributes (department,employee_expirience, employment_status) of the training data. The variation in the classification is observed by comparing a different number of entries for each instance to select a suitable attribute. The attribute employee experience gives minimum incorrect classification errors and is the most suitable.

2) Dataset 2: Social Network

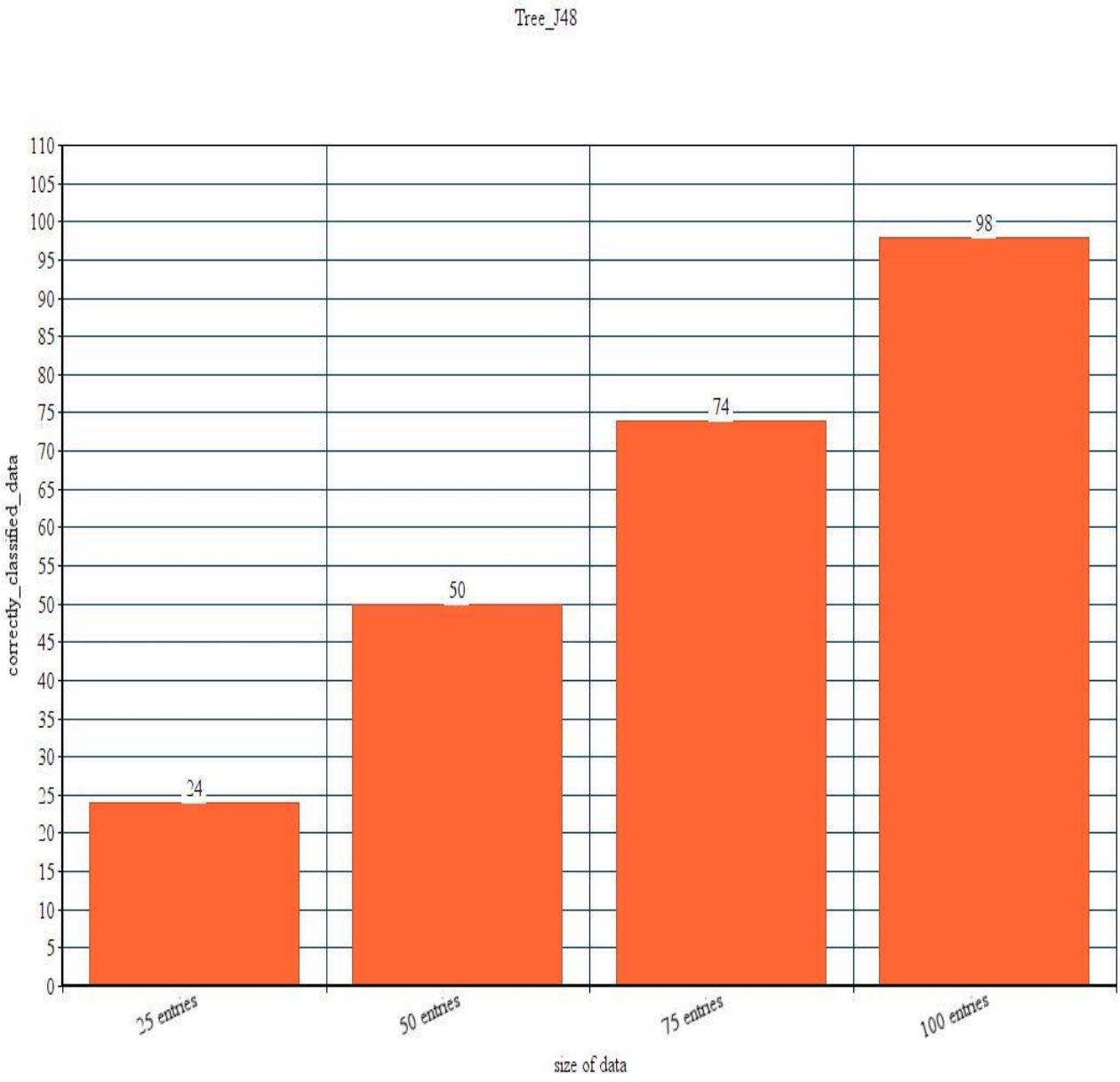

Fig 3: Correctly Classified instances

This graph depicts the comparison between correctly classified entries of different sizes of the training data. The variation in the classification is observed by comparing different dataset sizes.

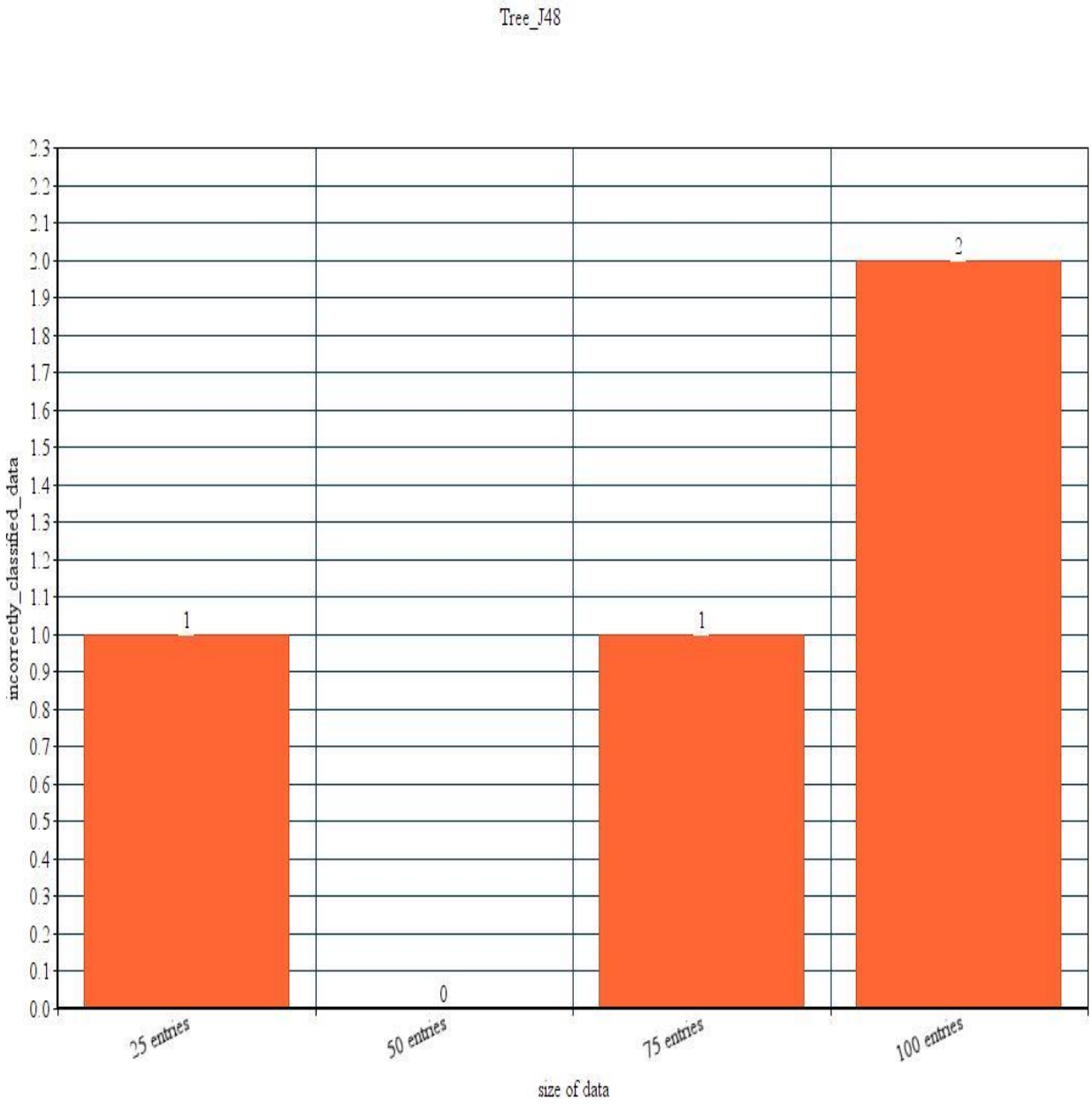

Fig 4: Incorrectly classified instances

This graph depicts the comparison between wrongly classified entries of different sizes of the training data. The variation in the classification is observed by comparing a different number of entries for each instance to select a suitable attribute.

B. Naive Bayes

1) Dataset 1: Employee dataset

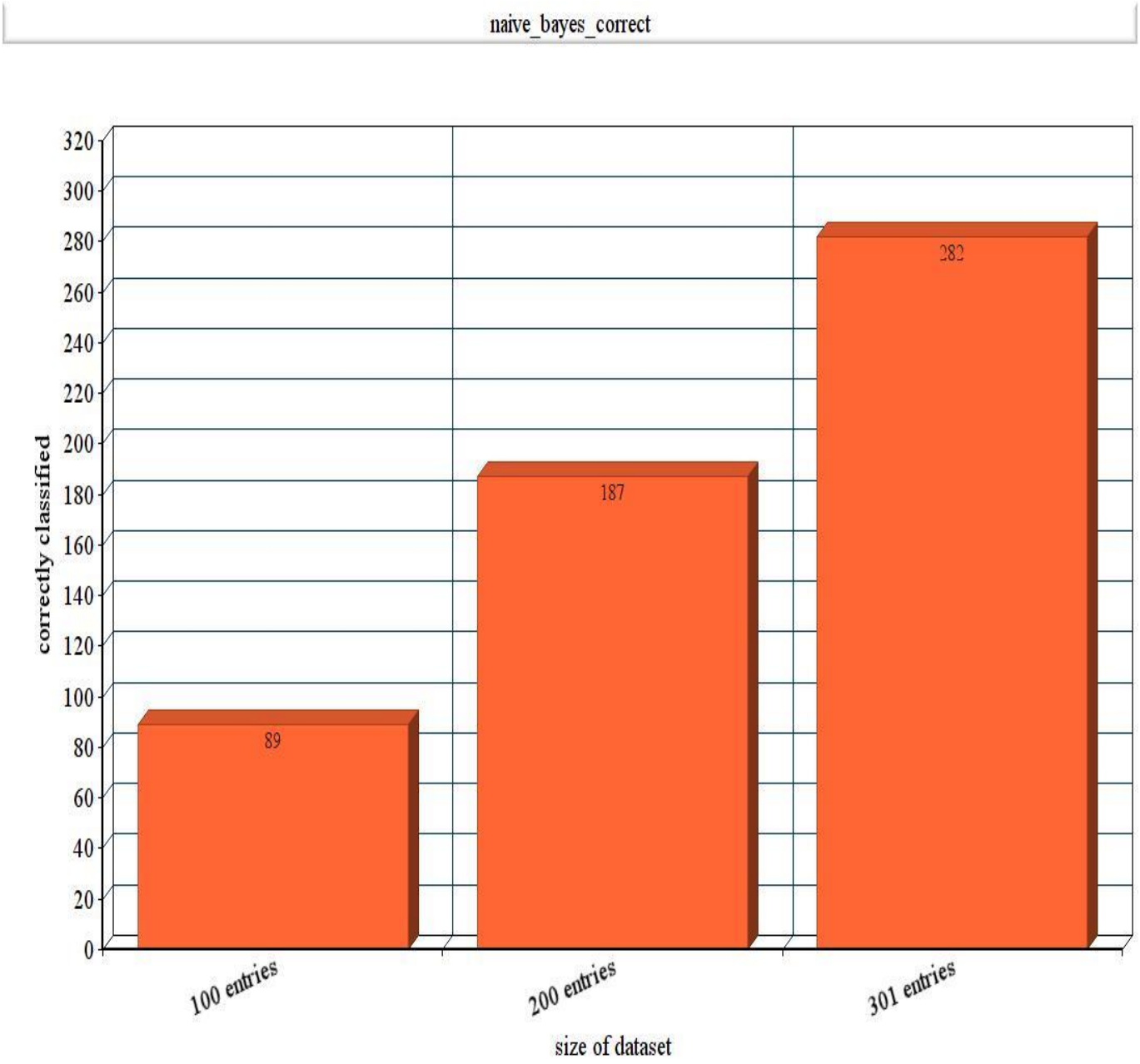

Fig: 5 Correctly classified data

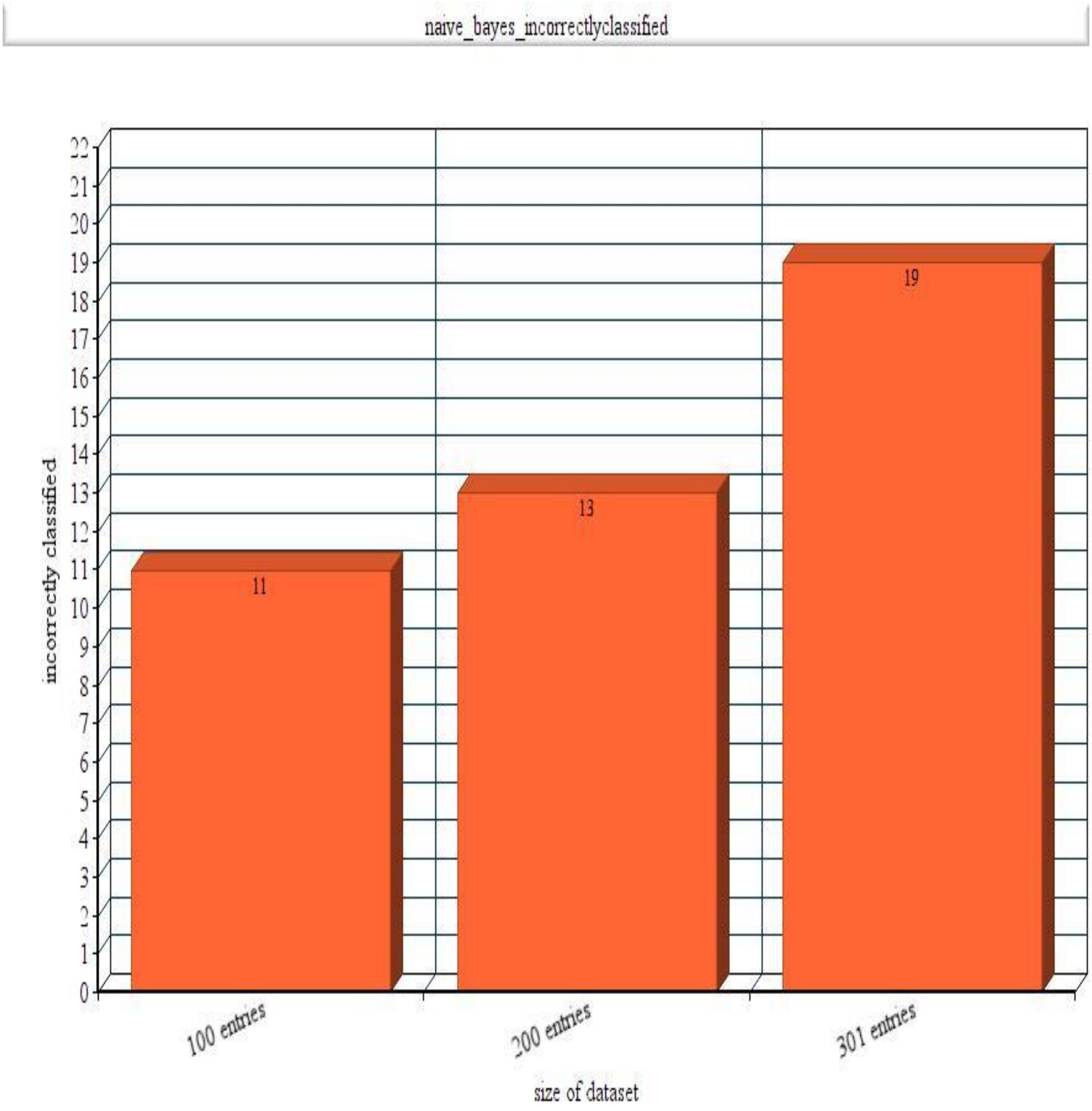

Fig 6: Incorrectly classified instances

The error ratio for 100 entries is 0.11,for 200 entries is 0.065 and for 301 entries is 0.063. The dataset with 300 entries gives us the least error ratio.

2) Dataset 2: Social Network

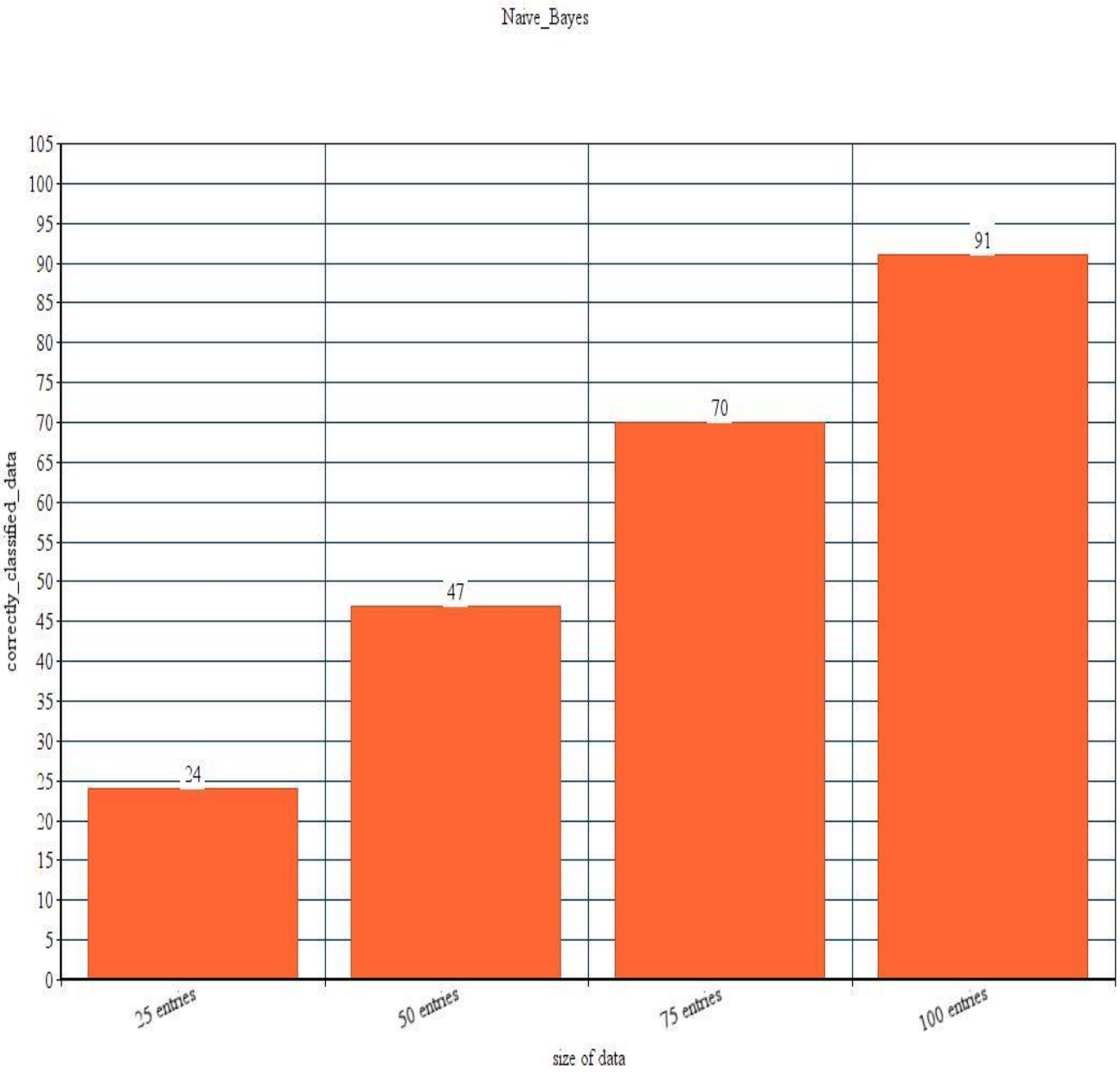

Fig 7: Correctly classified instances

For the dataset with 25 entries, 24 instances are correctly classified. For the dataset with 50 entries, 47 are classified correctly. For the dataset with 75 entries, 70 are classified correctly, and with 100 entries, 91 instances are classified correctly.

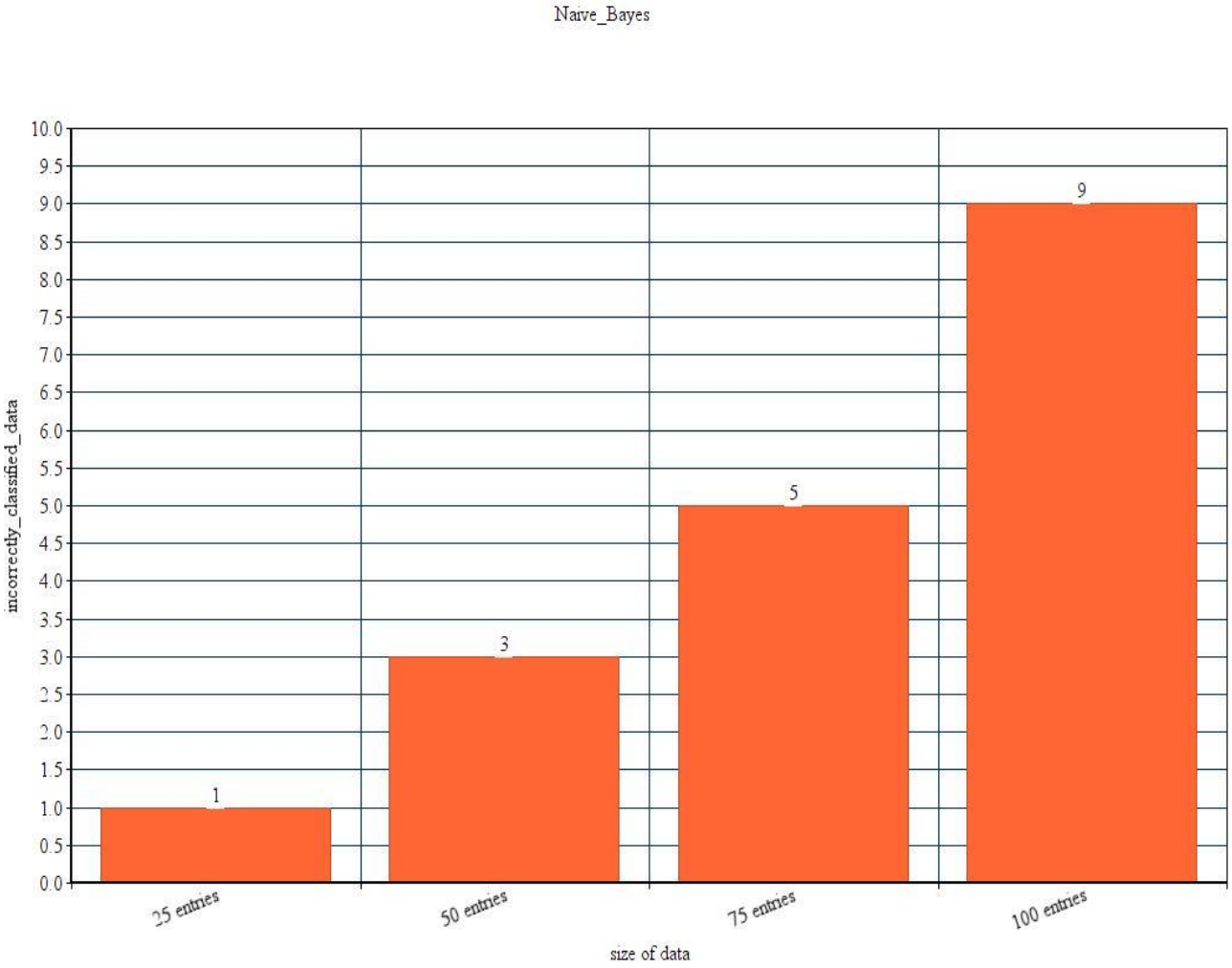

Fig 8: Incorrectly classified instances

For the dataset with 25 entries, only 1 instance is wrongly classified. For the dataset with 50 entries, 3 instances are wrongly classified. For the dataset size of 75 entries, 5 instances are wrongly classified, and for 100 entries, 9 instances are wrongly classified.

C. K-means

1) Dataset 1: Employee dataset

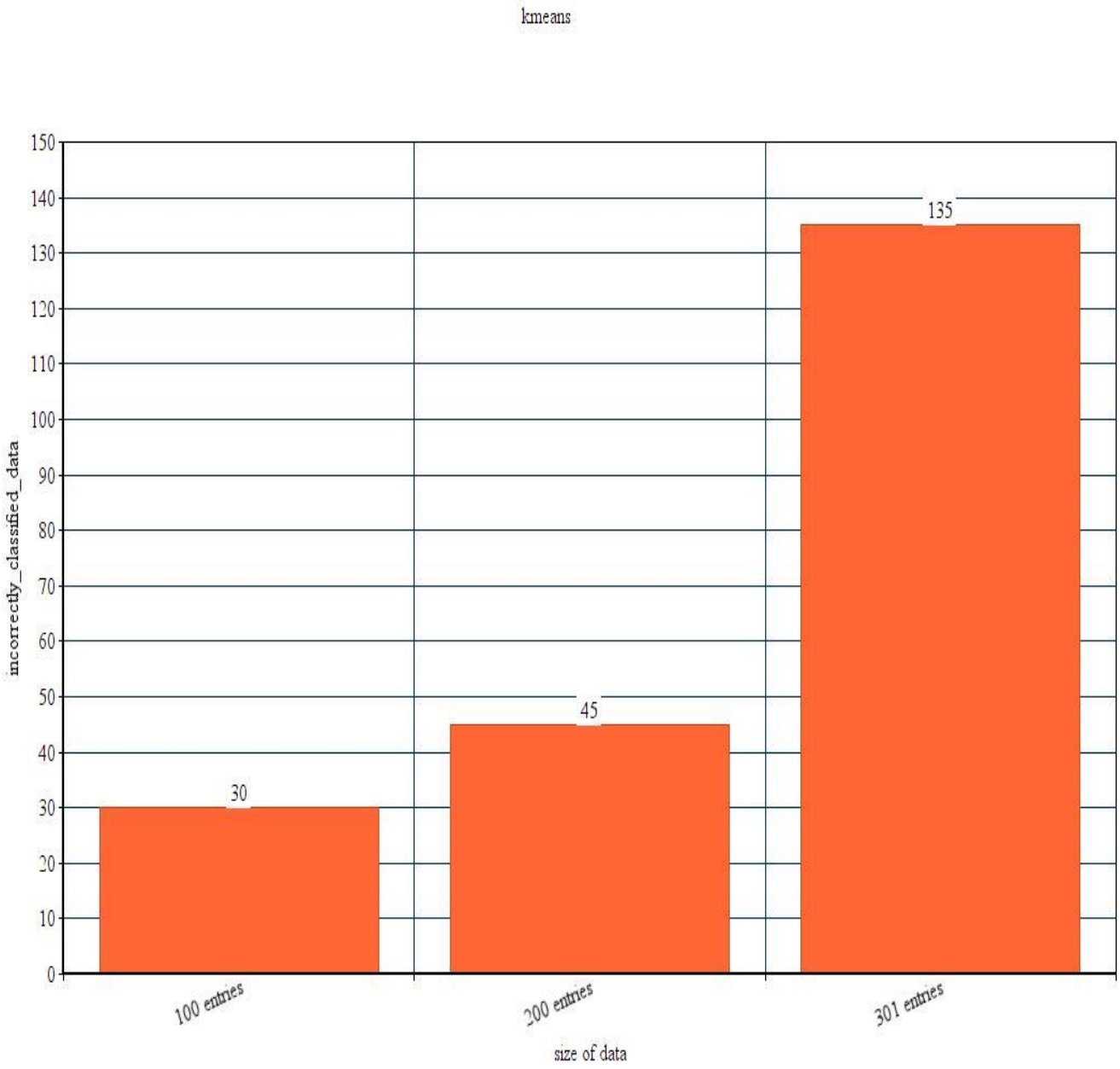

Fig 9: K-means incorrectly classified data

For the dataset with 100 entries, 30 are not correctly clustered instances. For the dataset with 200 entries, 45 are not correctly clustered instances. For the dataset with 301 entries, 135 are not correctly clustered instances.

2) Dataset 2:Social Network

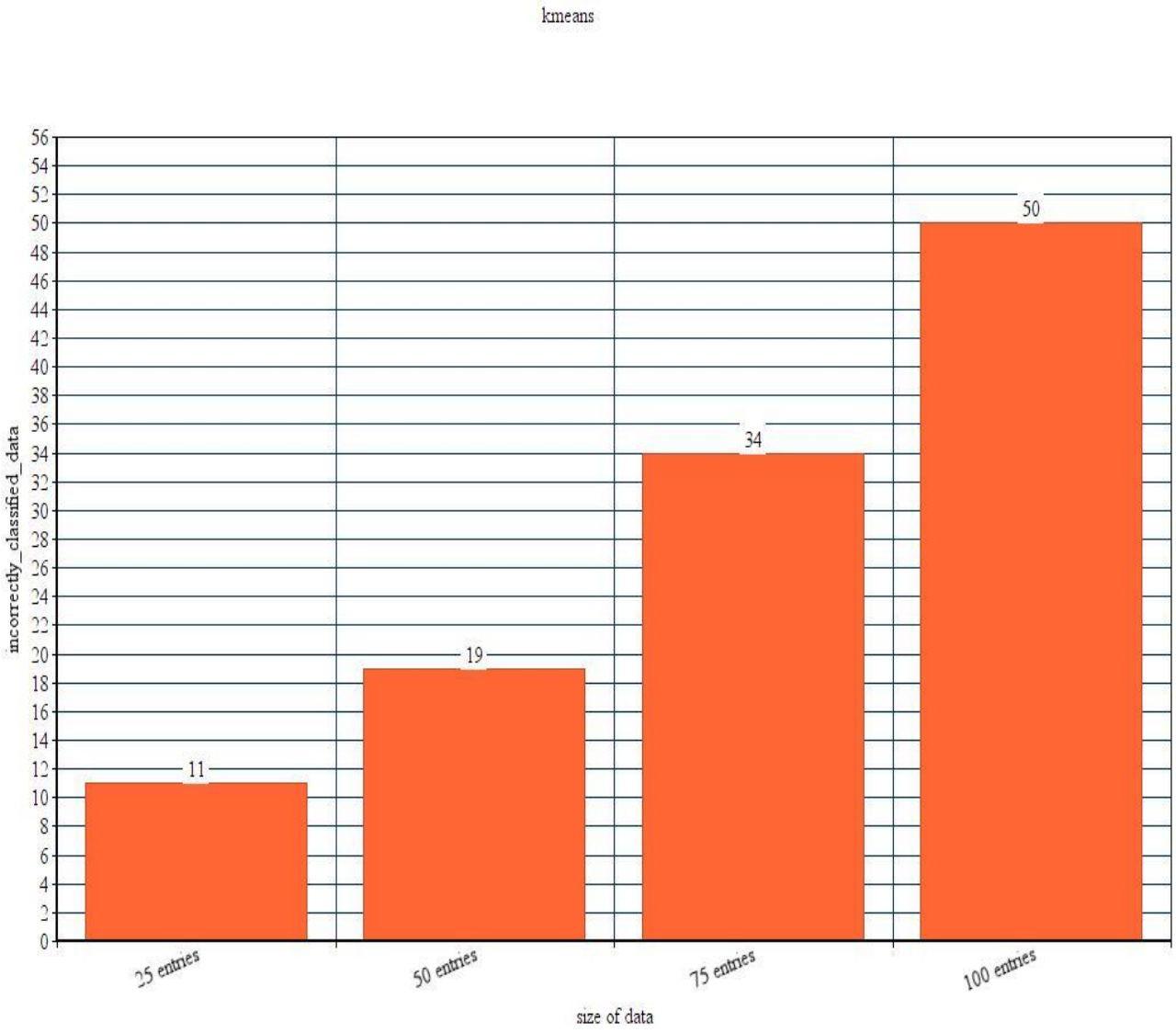

Fig 10: K-means incorrectly classified data

For the dataset with 25 entries, 11 are not correctly clustered instances. For the dataset with 50 entries, 19 are not correctly clustered instances. For the dataset with 75 entries, 34 are not correctly clustered instances, and for the dataset, with 100 entries, 50 are not correctly clustered instances.

D. Comparison Chart

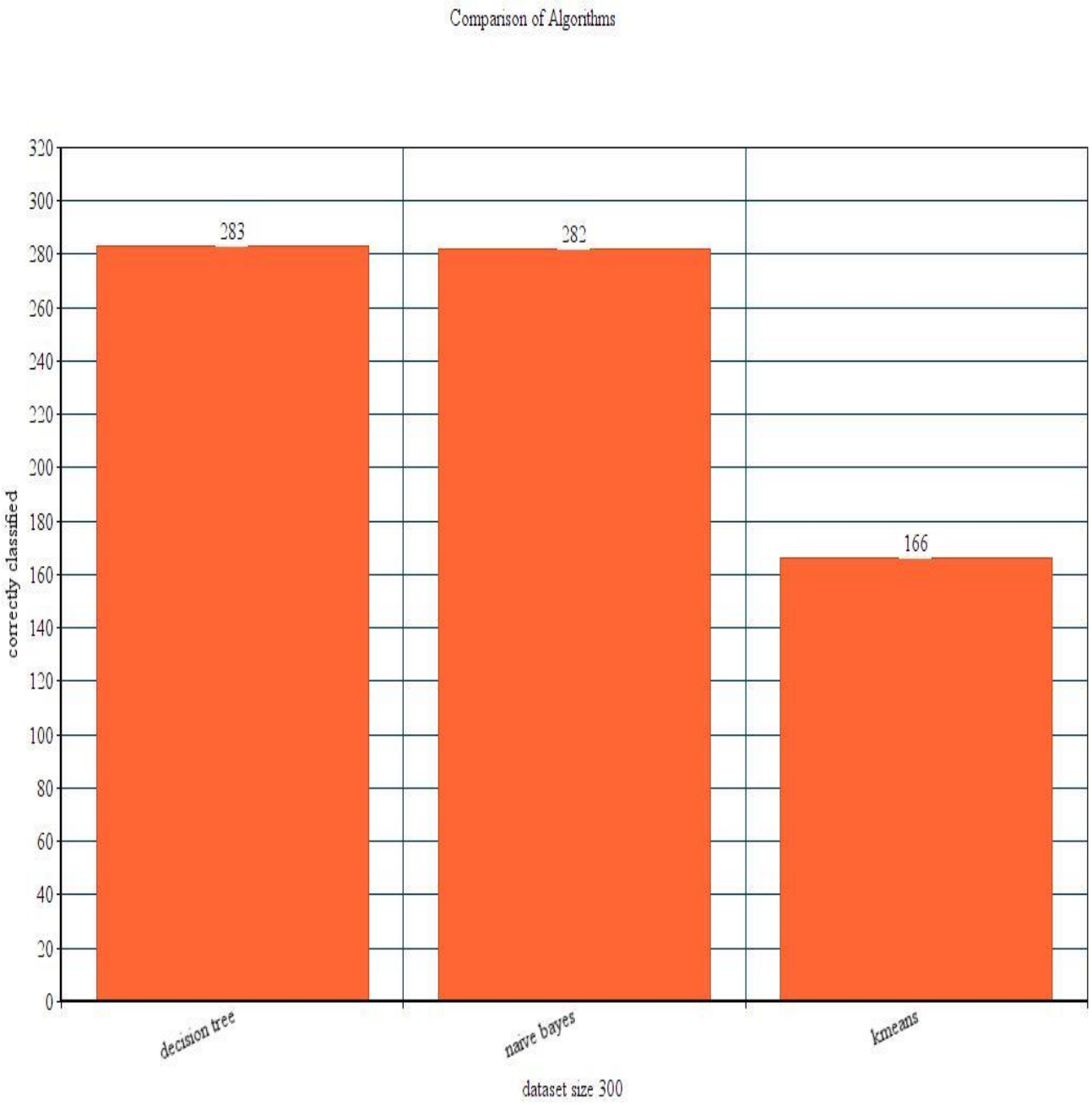

Fig 11: Comparison chart for employee dataset

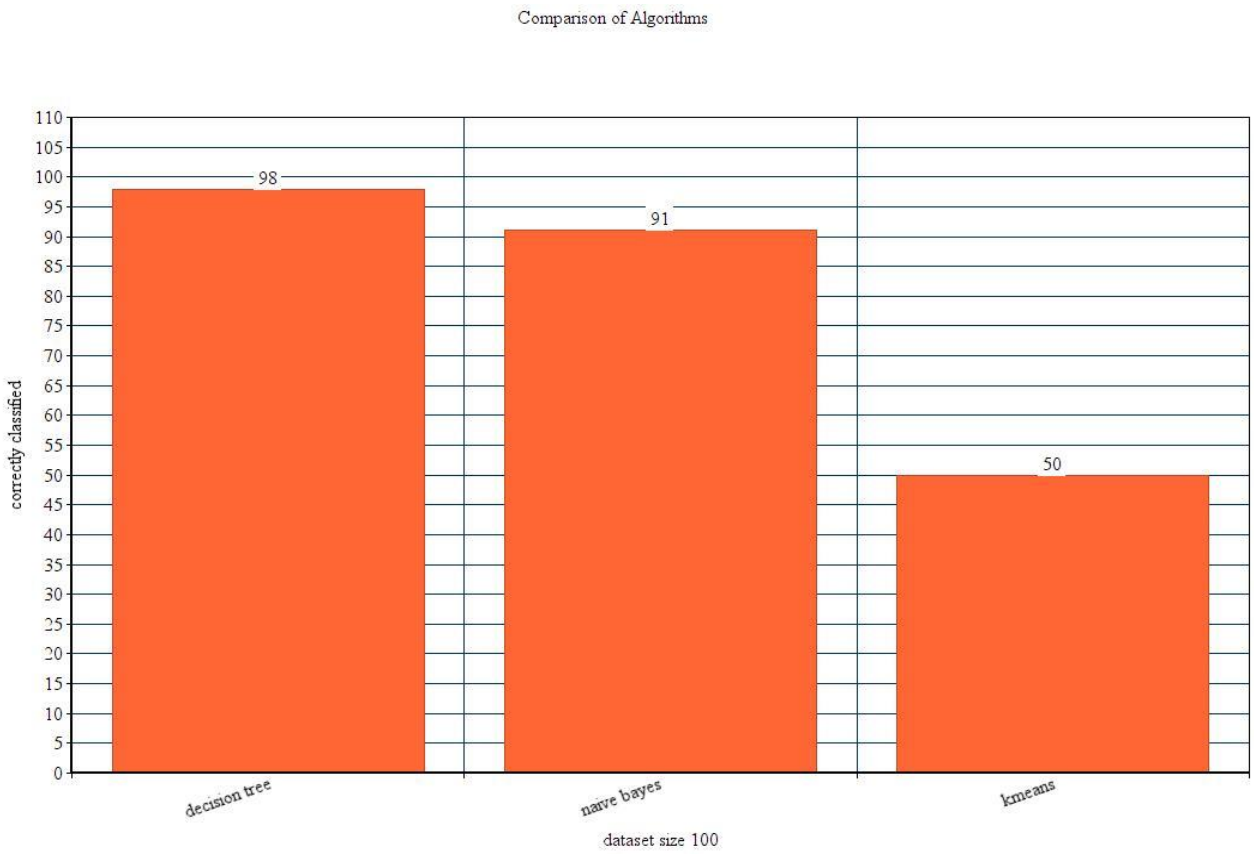

Fig 12: comparison chart for company dataset

E. Performance Metric :

The performance metrics used to analyze the data are precision, recall, and F1 score metrics. The F1 Score metric is the harmonic mean of the precision and recall metrics.

F. Precision :

Precision or positive predictive value is obtained by dividing the true positive (tp) value by the sum of true positive and false positive (fp).

G. Recall :

Recall or sensitivity is obtained by dividing the true positive value by the sum of true positive and false negative (fn).

H. F1 score :

The F1 Score metric is the harmonic mean of the precision and recall metrics.

I. Weka Results :

Algorithm	Precision	Recall	F-Measure
J48(Decision tree)	1.0	1.0	1.0
Naive Bayes	0.882	0.882	0.882

Table 1 Employee dataset(size 100 entries)

Algorithm	Precision	Recall	F-Measure
J48(Decision tree)	0.986	0.985	0.985
Naive Bayes	0.934	0.926	0.928

Table 2. Employee dataset(size 200 entries)

Algorithm	Precision	Recall	F-Measure
J48(Decision tree)	1.0	1.0	1.0
Naive Bayes	0.941	0.941	0.941

Table 3. Employee dataset(size 301 entries)

Algorithm	Precision	Recall	F-Measure
J48(Decision tree)	0.963	0.960	0.960
Naive Bayes	0.963	0.960	0.960

Table 4. Social Network dataset(size 25 entries)

Algorithm	Precision	Recall	F-Measure
J48(Decision tree)	1.000	1.000	1.000
Naive Bayes	0.940	0.940	0.939

Table 5. Social Network dataset(size 50 entries)

Algorithm	Precision	Recall	F-Measure
J48(Decision tree)	1.0	1.0	1.0
Naive Bayes	0.932	0.933	0.933

Table 6. Social Network dataset(size 75 entries)

Algorithm	Precision	Recall	F-Measure
J48(Decision tree)	0.980	0.980	0.980
Naive Bayes	0.912	0.910	0.911

Table 7. Social Network dataset(size 100 entries)

V. CONCLUSION

The proposed work involved extensive study of various machine learning algorithms, collection of a large number of datasets including startups and different aspects which support the development of a startup has been covered in this project which covers different aspects such as Networking with Potential Employee which helps an entrepreneur to find the potential employees in the domain and location specified, Controlling the Image Online which helps to create an image of the company that is different techniques as per domain which can be used by the entrepreneur to get the name of the company to appear in the first page of the Google Search concerning Search Engine Optimisation(SEO).

The third Module, which includes Business Growth, includes names of startups as per the Location and Domain, which helps the entrepreneur to find the key competitors in the domain he is approaching and the location where he is, according to which he can adopt strategies and techniques to be a leader in the market concerning key competitors. The Research part focuses on the analysis of different algorithms, which helps select attributes that should be undertaken to have maximum accuracy in the output displayed

VI. REFERENCES

[1] Halima Elaidi, Younes Elhaddar, Zahra Benabou, Hassan Abbar, An idea of a clustering algorithm using support vector machines based on a binary decision tree, International Conference on Intelligent Systems and Computer Vision (ISCV), (2018) 131-145.

[2] Fatih Ertam, Analysis of Data Using Machine Learning Approaches in Social Networks, International Conference on Computer Science and Engineering (UBMK), (2017) 112-136.

[3] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, TensorFlow: A System for Large-Scale

[4] Machine Learning, Usenix Association(Researchgate), 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI' 16). November 2–4, 2016, Savannah, GA, USA, (2016) 1-21.

[5] Cheng-Hung Tsai, Han-Wen Liu, Tsun Ku, Wu-Fan Chien, "Personal Preferences Analysis of User Interaction based on Social Networks," International Conference on Computing, Communication and Security (ICCCS), (2015) 273-297.

[6] Yoshihiro Kawano, Yuka Obu, Yorinori Kishimoto, Eiji Nunohiro, Tatsuhiro Yonekura, "A Personal Branding for University Students by Practical Use of Social Media," 15th International Conference on Network-Based Information Systems, 2012.

[7] Bruno Vicente Alves de Lima, Vinicius Ponte Machado, Machine Learning Algorithms applied in Automatic Classification of Social Network Users, Fourth International Conference on Computational Aspects of Social Networks (CASoN),2012.