Gender Recognition using Smartphone Usage Pattern

BY
PIYAL DEY
ID: 182-15-11720

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Mr. Ahmed Al Marouf
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Shah Md. Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

SEPTEMBER 2022

APPROVAL

This Project titled "Gender Recognition using Smartphone Usage Pattern", submitted by Name: Piyal Dey, ID No: 182-15-11720 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on Tuesday, 13 September 2022.

BOARD OF EXAMINERS

21/09/22

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman

Nazmun Nessa Moon (NNM)
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Faisal Imran

Mr. Faisal Imran (FI)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Farid

Dr. Dewan Md Farid
Professor
Department of Computer Science and Engineering
United International University

External Examiner

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Ahmed Al Marouf, Lecturer, Department of Computer Science Engineering** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Ahmed Al Marouf
Lecturer
Department of Computer Science Engineering
Daffodil International University

Co-Supervised by:

T Siddique

Shah Md. Tanvir Siddiquee
Assistant Professor
Department of Computer Science Engineering
Daffodil International University

Submitted by:

Piyal

Piyal Dey
ID: 182-15-11720
Department of Computer Science Engineering
Daffodil International University

ACKNOWLEDGEMENT

First and important my heartfelt and sincere gratitude goes to Almighty Allah who has empowered us to practically whole our thesis. We really grateful and wish our profound our indebtedness to **Ahmed Al Marouf, Lecturer** and **Shah Md. Tanvir Siddiquee, Assistant Professor,** Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor and co-supervisor in the field of “Human Computer Interaction” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan, Head of Department of Computer Science and Engineering,** for giving us an opportunity to carry out the research work and also to other faculty members and the staff of the CSE department of Daffodil International University.

Thanks to Daffodil International University for the study opportunity and for the specialized help during the last period of completing this proposal for this thesis.

Finally, I must express my very profound gratitude to my parents and to my friend and for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of master and this thesis. This achievement would not have been conceivable without them.

ABSTRACT

In today's world, mobile phones can quickly establish long-distance interconnections and perform important tasks. For this reason, the number of smartphone users is increasing rapidly. Smartphones are used not only as an aid, but also to identify and track users based on their behavior. In this survey, we identified the gender of this user based on smartphone usage pattern. The prefer paper work inspect 429 instances of male and female of this specific age range (18-more than 30 years) and collects them in a research that includes the participants and their gender in order to build a classify dataset. In order to improve approximation and experimentation, some different machine learning algorithms are applying. In those algorisms I can choose 5 algorithms which is perform better compare to other algorithms and 3 algorithms including RandomCommittee, IBK and Kstar are compared to other is perform best. Five distinct categories of classification algorithms are compared, classifiers based on trees, Bayes, and function-based classifiers. One of the available algorithms RandomCommittee, improves performance with an accuracy of approximately 83.50 %. Thanks to this survey, we are aware of the characteristics related to smartphones, and the use of smartphones is important for determining the gender of users. Knowing these features will help you think about security, biometrics, and privacy issues.

TABLE OF CONTENTS

LIST OF TABLES

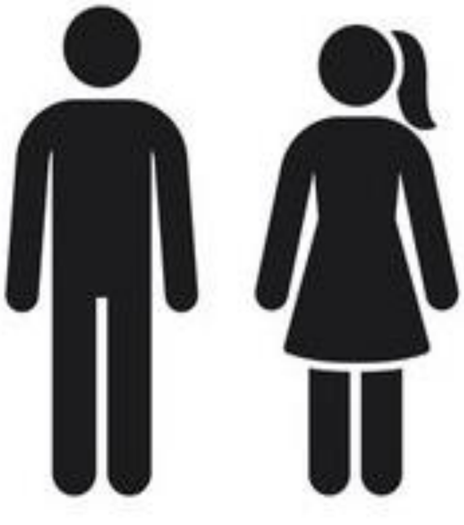
TABLES **PAGE**

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Most of the people in the world spend a lot of time on mobile phones, which are essential to today's world and the environment. Advances in digitalization in almost every aspect of our lives have led to the pinnacle of artificial intelligence. People use smartphones as a tool. With multiple Android applications available that can perform the same functions as a desktop or laptop computer, today's users own more mobile phones than at any point in the past. Due to this frequency, the use of certain mobile applications leaves a modern impression. For example, chat / messenger apps, gaming apps, social networking apps, etc. One group uses smartphones for this reason, and another group uses smartphones for different purposes. It is based on various variables such as gender, age, generation, and personality. Gender is one of the most important factors in determining how different mobile phones are used. Gender may refer to a man, woman, or another person in relation to a person's self-image, society, or the identification of a person known to society. The difficult reality is that certain mobile phones cannot automatically determine how to use them. In a stacked technique for gender recognition through voice. In 2018 was researcher Gupta, P., Goel, S. and Purwar was made this easier (e.g. [13]). In this study, an attempt was made to automatically detect gender using machine learning techniques. This may be a new addition to today's world.



1.2 Motivation

In today's world, the number of people using smartphones is steadily increasing all over the world. However, men and girls always use smartphones differently. When we buy a smartphone, smartphone makers usually don't want to know our gender. The smartphone should have been specially designed for this gender if you understand the usage history of boys or girls. This would have greatly benefited the user. For ease of understanding, we collect information about use by boys or girls for this purpose.

After reviewing the literature, no new findings were found in this area. There is an urgent need for research here. Therefore, this study is conducted using widely recognized machine learning techniques apply for this thesis gender recognition using smartphone usage pattern. But initially, body language, face shape, and facial features are used to identify gender (e.g [4], [5], [6], [7]). A person's voice may now be used to establish their gender. The two types of face-based gender identification are as follows: Information on geometry and texturing (e.g. [2], [8], [12]).

1.3 Rationale of the study

As already shown, from a Bangladeshi perspective, gender identity based on smartphone usage patterns is less important. For this reason, we strongly recommend using machine learning techniques and smartphone usage patterns for gender detection.

The study of probabilistic, optimization, and statistical methods that allow computers to "learn" from past experiences and patterns is known as machine learning, a subfield of artificial intelligence. This technique finds similarities between large, noisy, difficult, and complex datasets. Machine learning approaches are currently used in the areas of detection, identification, and classification. Machine learning is currently used to address several classified and known problems.

Machine learning is important in this area and is often applied by researchers with excellent results. For this reason, I decided to use machine learning in my research.

1.4 Research Questions

Acceptable research topics for developing research papers, projects, or treatise recommendations. Give your efforts a clear focus and purpose by identifying exactly what you need to find.

- How do smartphone usage patterns help predict gender perception?
- How can I continue to use this feature with those data?
- How much data do we collect?
- How much of our train and test dataset are there?
- Is the information collected and the machine learning methods used accurate and consistent?
- Do I need to create a new model or apply a well-known and widely used machine learning technique?

1.5 Expected Outcome

The purpose of our work is to predict gender perception based on smartphone usage trends. Recently, the use of smartphones is expanding rapidly all over the world. However, men and girls rarely have the same tendency to use smartphones. In most cases, smartphone makers are not interested in learning about our gender.

1.6 Report Layout

The following are the contents of this research project:

- Chapter 1 summarizes research motivations, rationale, research questions, and expected outcomes.
- Chapter 2 The scale of the issue, the difficulties we encounter, and a synopsis of pertinent prior research are all included in.
- Chapter 3 The workflow flow charts, data collecting procedure, preprocessing of the data, statistical analysis, and feature implementation for this project are all described in.
- Chapter 4 The experimental analysis, several relevant research, a summary of study accuracy, and study outcomes are all included in.
- Chapter 5 The findings of this society's research are presented in.
- Chapter 6 An summary, restrictions, and future work on this topic are provided in.

CHAPTER 2
BACKGROUND STUDY

2.1 Introduction

This section describes previous work comparable to this. The size of the problem and the challenges we faced. See the Linked Work section for a list of research studies, related projects, their usage, classifiers, and work accuracy. For the sake of clarity, we have created a summary of each paper in the "Research Summary" section and summarized it in a table. Discuss how to contribute or continue with this effort, depending on the severity of the problem. The Issues section also contains solutions to problems and failures found during the investigation.

2.2 Related Works

This section focused on work on research goals similar to related topics. Researchers in different disciplines are trying to detect gender in different ways. Some systems tried to use text data, while others used the face image feature. In an effort to establish sex. Some studies have used voice modulation to localize the same. We studied and monitored their methods and made further developments and publications of their research.

An important and exciting field of study in the science of human-computer interaction is gender detection. Several studies have used different approaches, models, and procedures to determine gender. In addition to using facial photographs to determine gender, researchers also used voice recordings in many studies was published in 2011 (e.g [1]). Gender recognition by voice using an improved and gender recognition using machine learning approach was publish by Livieris, I.E., Pintelas, E. and Pintelas this the most valuable research (e.g. [17], [18]).

Horever Meena, T. and Sarawadekar publish an article (e.g [3]) Gender recognition using in-built inertial sensors of smartphone. Research publications on emotion recognition via smartphones are exciting. As a result, our plan will add a new dimension to the man-made area. Robot and human-computer interaction (HCI) are two examples. Robotics, artificial intelligence, HCI. Less effort is required to obtain gender identity. Most articles describe how mobile phones affect gender detection. There are several studies and articles on interesting topics of gender recognition using mobile phones and personality. Our project will be a new component to bring a new dimension to the areas of artificial intelligence, human-computer interaction, and robotics.

2.3 Comparative Analysis and Summary

Previously, some work was done using machine learning algorithms and data mining techniques to detect gender in smartphone usage patterns. Machine learning techniques are being used more and more frequently in today's gender identity. This section contrasts these similar works.

Table 2.1: SUMMARY OF SIMILAR RESEARCH WORKS

Source	Technique	Data Collection	Result
[19] Jayasankar, T., Vinothkumar, K. and Vijayaselvi, A., 2017	Genetic Algorithm	80 Data	90% accuracy
[9] Lemley, J., Abdul-Wahid, S., Banik, D. and Andonie, R., 2016	SVM Classification	11338 images	89% accuracy
[11] Buyukyilmaz, M. and Cibikdiken, A.O., 2016	Nadam optimization algorithm	1268 data	96% accuracy
[14] Chola, C., Benifa, J.V., Guru, D.S., Muaad, A.Y., Hanumanthappa, J., Al-Antari, M.A., AlSalman, H. and Gumaei, A.H., 2022.	KNN Classification	50 images	90% accuracy
[15] Gauswami, M.H. and Trivedi, K.R., 2018	CNN on raspberry Pi platform	1788data	94% accuracy
[16] Ertam, F., 2019	LSTM networks	586 data	98.4% accuracy
[20] Gupta, S., 2015	Functional Trees	1162 imgae	93.82% accuracy

Machine learning, deep learning, and AI have recently been used in all areas of data science for prediction, classification, and detection models. The detection model uses many well-known techniques such as CNN, ANN, SVM, kNN, and logistic regression. Based on literature reviews, we can conclude that kNN, naive Bayesian, SVM, random forest, CNN, and decision tree algorithms are particularly effective and popular for prediction, prediction, and detection models. I recognize in my study I endeavor to contraption the RandomCommittee, IBK, Kstar, DecisionTable and RandomizableFilteredClassifier classifier algorithm to predict Gender Recognition using Smartphone usage pattern and I find 83.55% accuracy in RandomCommittee.

2.4 Scope of the Problem

Basically, in our research, we examine the data and use machine learning techniques to build the model. You can use the proposed model to predict gender. Our main task is to determine gender based on the usage behavior of mobile phone users. We must evolve to catch up with this modern world. In addition, the number of mobile users around the world is increasing every day. When I try to buy a mobile phone now, I can't find the model I want. Due to the fact that usage remains the same regardless of gender. However, this issue is not taken into account by mobile device manufacturers. This gives consumers access to multiple features that go beyond what they really need. This algorithm was developed to predict gender detection based on smartphone usage trends. That’s why when we recognize gender using smartphone pattern then we can understand that the user is boy or girl and it’s through smartphone build up company made their phone requirement of this pattern that is use a boy or girl. That is the main scope of my paper to recognize gender using smartphone pattern.

2.5 Challenges

Efforts based on this research have many obstacles that need to be addressed. To implement it, you need to overcome each of these. Like as-

- When checking data, choose the appropriate questions.
- Dataset of various ages.
- Maintains the dataset.
- Figuring out and choosing the right value.
- Procedural difficulties.
- Pay attention to both masculine and female traits.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

This work will contribute to the development of a gender detection model for smartphones. Many machine learning techniques are used to create this model. RandomCommittee, IBK, Kstar, DecisionTable and RandomizableFilteredClassifier algorithm perform best and those are used to recognized gender using smartphone pattern. This algorithm is used to identify gender. There are a total of 20 characteristics that are directly or indirectly related to gender recognition. The dataset was processed as needed before the final embedding. To determine the best method for your model, evaluate the specificity, accuracy, sensitivity, recall, F1 score, and lock curve of each method. According to our research, RandomCommittee regression provides the highest accuracy. And I used this algorithm weka software to find best accuracy. On the other algorithm was perform pretty similar to the compare RandomCommittee algorithm.

3.2 Research Subject and Instrumentation

Subject: Gender Recognition using Smartphone Usage Pattern.

Instrumentation:

Deep learning, data mining, and machine learning algorithms have recently been widely accepted and endorsed by all types of prediction, detection, and detection. Apply some machine learning algorithms to the collected dataset to see which algorithm best performs our instructions. Multiple algorithm including RandomCommittee, IBK, Kstar, DecisionTable and RandomizableFilteredClassifier algorithm were apply to recognize gender using smartphone pattern. For deep analysis I can use SPSS software to analysis LIKERT-SCALE question in database. Recently, One of the most famous and popular programming languages is called "Python" and is used primarily by scholars for research purposes. I can use,

- Google Colab.
- Weka 3.9.6 Software.
- IBM SPSS Statistics 26 Software.

3.3 Data Collection

For the research purpose I gather the publicly accessible dataset for this study. Johora Akter Polin and Omayer Khan addressed gender identification using smartphone in their article (e.g. [10]). They used 429 samples for both male and female tests. This dataset was created using some queries. The published public data found has not been preprocessed.

The data was gathered based on the following 21 characteristics.

- Age
- Operating system
- In a day, how many hours do you use your mobile phone?
- Spent most time in- (Phone call & Messaging, Social Media, Gaming)?
- How much time you spent on mobile? (calls only)
- How much time do you use internet?
- What is your primary purpose for using internet on your mobile phone?
- How many social media applications you have in your phone?
- How many camera applications you have in your phone?
- How much time you spent on social media? (Facebook,Whatsapp,Instagram etc)
- Do you constantly check social media?
- Do you often think that your smartphone is ringing/ vibrating when it is not?
- When your phone rings buzzes, do you feel an intense urge to check?
- Do you look at your phone after you get up in the morning or before going to bed?
- Do you sleep with your smartphone on or under your pillow or next to your bed?
- Do you constantly check your phone if you did not have a data signal or WiFi?
- Do you feel a great deal of anxiety if your phone not working/you accidentally left it?
- Do you find yourself Always passing time in searching on google/ E-commerce sites?

- Do you spend more time texting, tweeting/emailing then talking to real-time people?
- When you eat meals, is your cell phone always part of the table place setting?
- Do you feel lonely if your smartphone doesn't ring for several hours?

3.4 Proposed Methodology

When I successfully gather data that is readily accessible, we may discover that certain data is missing some numbers. There are several kinds of data as well, including category and numerical data. For machine learning algorithms, this kind of data is unsuitable. Decide to modify the data as needed to make it compatible with the algorithm. After the data is collected, data processing can transform the data into the appropriate format. Best results can easily be achieved with certain types of processed data.

Start by cleaning up those data. Check the record for missing or null values. Allocator has solved the missing value issue. Instead of deleting the data, I tried to fill this null value with the most relevant value.

Then develop a layer that transforms each textual or category piece of information into numerical information. The final step in data conversion is normalization. The Min Max normalization strategy is utilized by the age data function. A correlated matrix displays the relationships between all of the actual data and all of the hypothetical data. Positive values indicate a strong connection between the data, whereas negative values indicate a weaker connection between the data and a decrease in significance.

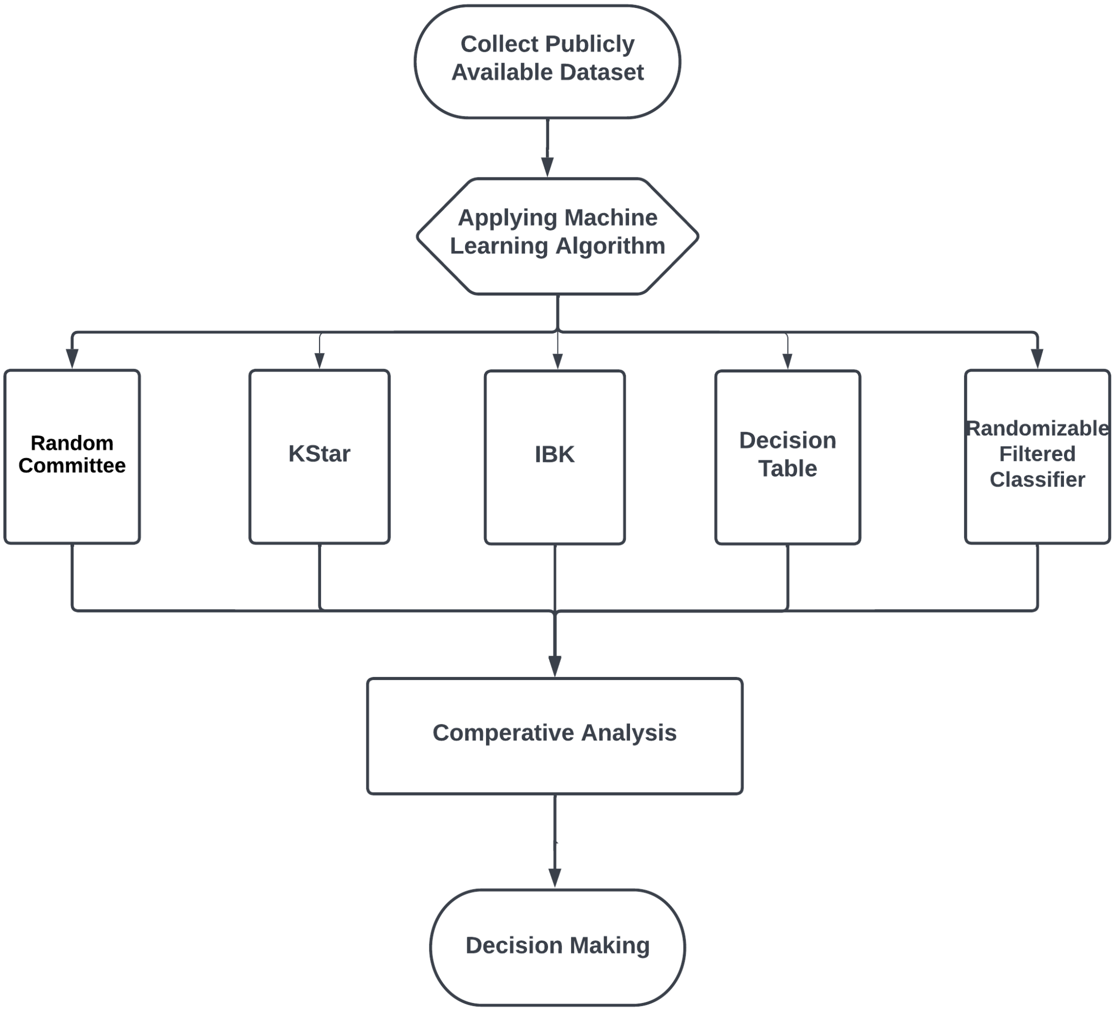

Collect Publicly Available Dataset
Applying Machine Learning Algorithm
Random Committee
KStar
IBK
Decision Table
Randomizable Filtered Classifier
Comperative Analysis
Decision Making

Figure 3.1: Steps of proposed methodology

3.5 Data Preprocessing Process

When I successfully collect publicly available dataset then I was mainly focus to cleaning the dataset and I was check for null value this dataset and find some null values. Then handling the null values which was top answer in this single data column. For deep analysis this value replaces for numerical values. After data gathering, data handling has the ability to transform data into the proper forms. A certain form of processed data makes it simple to get the best results.

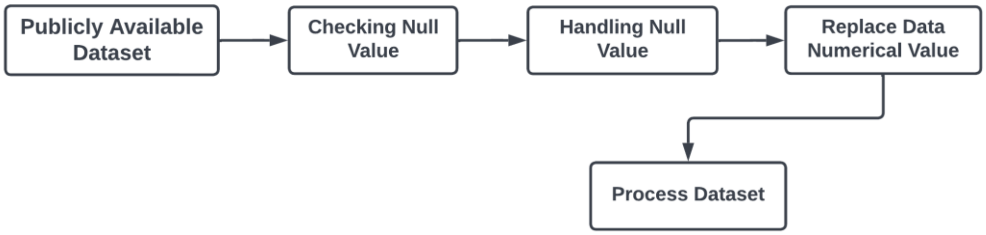


Figure 3.2: Data pre-processing process

We eventually receive the final processed dataset that we desired as a result. Utilizing "Google Colab," the entire data processing procedure was completed.

3.6 Data Analysis Process

Data analysis is the methodical application of logical and statistical approaches to explain and demonstrate, summarize and assess, and assess data. For data analysis I can use SPSS Statistic 26 software. Data analysis for descriptive and bivariate statistics, numerical result forecasts, and predictions for group identification are all provided by SPSS. Additionally, the program offers graphing, direct marketing, and data processing functions. In its main view, the software interface shows open data in a manner akin to a spreadsheet.

Can an instrument be consistently read across multiple situations is the main goal of reliability? The measurement of a research tool's reliability is whether it consistently produces the same results. Data came from a publicly accessible dataset.

Table 3.1: Validity and Reliability

Scale	Number of items	Cronbach’s Alpha
Likert scale	10	.702
Total	10	>.700

Internal consistencies obtained by Cronbach's alpha were higher than the minimal value of.700 necessary for satisfactory dependability.

The Other data table displays data along with the dependability coefficient for removed items. Cronbach's alpha coefficient continues to be steady.

Table 3.2: Reliability Item-total statistics

Item-Total Statistics				
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@11.Doyouoftenthinkthatyoursmartphoneisringingvibrating	17.05	10.087	.310	.688
@12.Whenyourphoneringsbuzzesdoyoufeelanintenseurgeto	17.21	10.021	.335	.684
@13.Doyoulookatyourphoneafteryougetupinthemorningor	17.71	9.880	.359	.681
@14.Doyousleepwithyoursmartphoneonorunderyourpillowor	17.54	9.815	.272	.697
@15.Doyouconstantlycheckyourphoneifyoudidnothaveadata	17.15	9.344	.441	.666
@16.Doyoufeelagreatdealofanxietyifyourphonenotworking	17.28	9.585	.382	.676
@17.DoyoufindyourselfAlwayspassingtimeinsearchingongoog	17.16	10.385	.203	.705
@18.Doyouspendmoretimetextingtweetingemailingthentalkin	17.08	9.421	.415	.670
@19.Whenyoueatmealsisyourcellphonealwayspartofthetab	16.93	9.492	.382	.676
@20.Doyoufeellonelyifyoursmartphonedoesntringforsevera	16.82	8.777	.510	.650

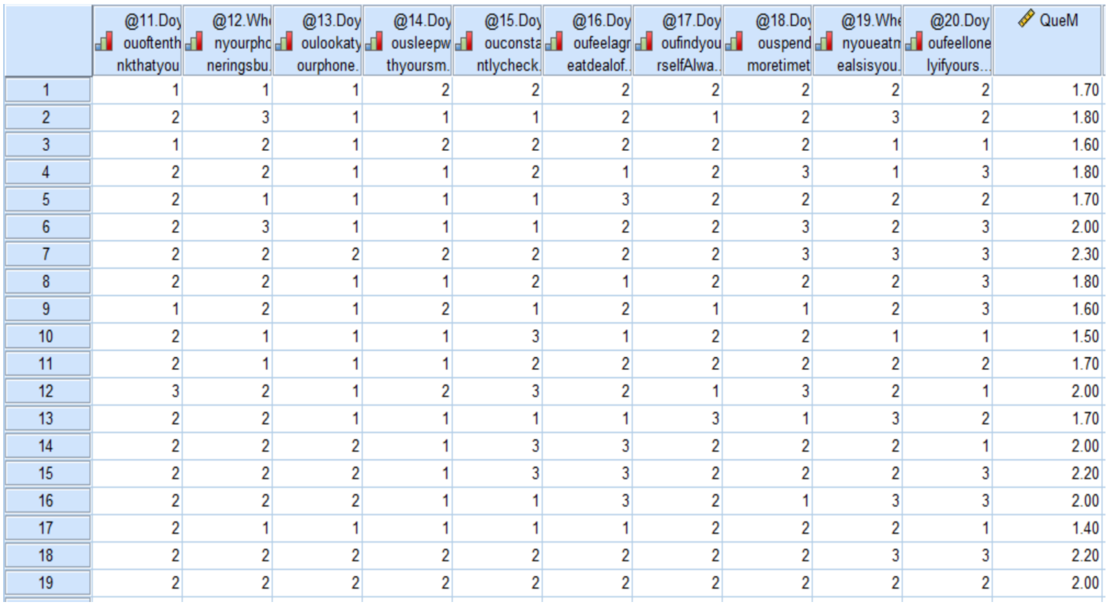
The elements in this table have been sorted based on the means or mean scores. And as we can see, item number 20 is the one that is most usually mentioned or agreed upon.

Table 3.3: Descriptive Statistics

Descriptive Statistics				
	N	Sum	Mean	Std. Deviation
@13.Doyoulookatyourphoneafteryougetupinthemorningor	429	598	1.39	.593
@14.Doyousleepwithyoursmartphoneonorunderyourpillowor	429	671	1.56	.723
@16.Doyoufeelagreatdealofanxietyifyourphonenotworking	429	782	1.82	.657
@12.Whenyourphoneringsbuzzesdoyoufeelanintenseurgeto	429	812	1.89	.574
@17.DoyoufindyourselfAlwayspassingtimeinsearchingongoog	429	832	1.94	.616
@15.Doyouconstantlycheckyourphoneifyoudidnothaveadata	429	837	1.95	.663
@18.Doyouspendmoretimetextingtweetingemailingthentalkin	429	867	2.02	.668
@11.Doyouoftenthinkthatyoursmartphoneisringingvibrating	429	882	2.06	.581
@19.Whenyoueatmealsisyourcellphonealwayspartofthetab	429	933	2.17	.683
@20.Doyoufeellonelyifyoursmartphonedoesntringforsevera	429	981	2.29	.742
Valid N (listwise)	429			

If the items are challenging, give up trying to understand their significance. When we determine the total number of respondents, we may understand this table based on the mean score. The standard deviation, on the other hand, indicates how each respondent differs from one another, which is why it is also known as the measure of dispersion. As a result, this scale serves to display the descriptive data.

In the dataset at first I generate a compute variable for analyze mean of all questioners that’s why I create a column QueM where analyze mean. In figure 3.3 it will be show that.



	@11.Doy ouoftenth nkthatyou	@12.Wh nyourphc neringsbu.	@13.Doy oulookaty ourphone.	@14.Doy ousleepw thyoursm.	@15.Doy ouconsta ntlycheck.	@16.Doy oufeelagr eatdealof..	@17.Doy oufindyou rselfAlwa..	@18.Doy ouspend moretimet	@19.Whe nyoueatn ealsisyou.	@20.Doy oufeellone lyifyours...	QueM
1	1	1	1	2	2	2	2	2	2	2	1.70
2	2	3	1	1	1	2	1	2	3	2	1.80
3	1	2	1	2	2	2	2	2	1	1	1.60
4	2	2	1	1	2	1	2	3	1	3	1.80
5	2	1	1	1	1	3	2	2	2	2	1.70
6	2	3	1	1	1	2	2	3	2	3	2.00
7	2	2	2	2	2	2	2	3	3	3	2.30
8	2	2	1	1	2	1	2	2	2	3	1.80
9	1	2	1	2	1	2	1	1	2	3	1.60
10	2	1	1	1	3	1	2	2	1	1	1.50
11	2	1	1	1	2	2	2	2	2	2	1.70
12	3	2	1	2	3	2	1	3	2	1	2.00
13	2	2	1	1	1	1	3	1	3	2	1.70
14	2	2	2	1	3	3	2	2	2	1	2.00
15	2	2	2	1	3	3	2	2	2	3	2.20
16	2	2	2	1	1	3	2	1	3	3	2.00
17	2	1	1	1	1	1	2	2	2	1	1.40
18	2	2	2	2	2	2	2	2	3	3	2.20
19	2	2	2	2	2	2	2	2	2	2	2.00

Figure 3.3: QueM Mean Column

The number of valid values is displayed in the table 3.4, the Case Processing Summary. Since pairwise missing data handling was chosen, the analysis makes use of all available data for each variable.

Table 3.4: Case Processing Summary

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
QueM	429	100.0%	0	0.0%	429	100.0%

The next table is the descriptive one. Each of the quantitative variables, including kurtosis and skewness, provides comprehensive univariate descriptive statistics.

Table 3.5: Descriptive analysis

Descriptive				
			Statistic	Std. Error
QueM	Mean		1.9103	.01642
	95% Confidence Interval for Mean	Lower Bound	1.8780	
		Upper Bound	1.9425	
	5% Trimmed Mean		1.9170	
	Median		1.9000	
	Variance		.116	
	Std. Deviation		.34013	
	Minimum		1.00	
	Maximum		2.90	
	Range		1.90	
	Interquartile Range		.40	
	Skewness		-.282	.118
	Kurtosis		.028	.235

The sampling skewness and kurtosis of the variables can be interpreted in terms of the classical normal distribution with skewness = 0 and kurtosis = 0. For height, the kurtosis is 0.113 and the skewness is 0.23 (slightly right skewed) (slightly stronger tails than normal, but not too much). For the weights, the kurtosis is 1.5 and the skewness is about 1 (heavier tail than normal distribution). These numbers by themselves are not particularly reliable predictors of deviation from normality, but they can support graphs and tests of normality.

Kolmogorov-Smirnov and Shapiro-Wilk tests are provided for two variables in the Normality Tests table.

Table 3.6: Tests of Normality

Tests of Normality						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
QueM	.094	429	.000	.984	429	.000

a. Lilliefors Significance Correction

The p-values for both the K-S and Shapiro-Wilk tests for body weight are very low (p 0.001) and the decision to reject is fairly clear. The height result is less certain, with a K-S p-value of 0.049 (just below the significance level of 0.05) and a Shapiro-Wilk p-value of 0.070. The results of these tests point in different directions. The Shapiro-Wilk test indicates normality and the KS test indicates nonnormality.

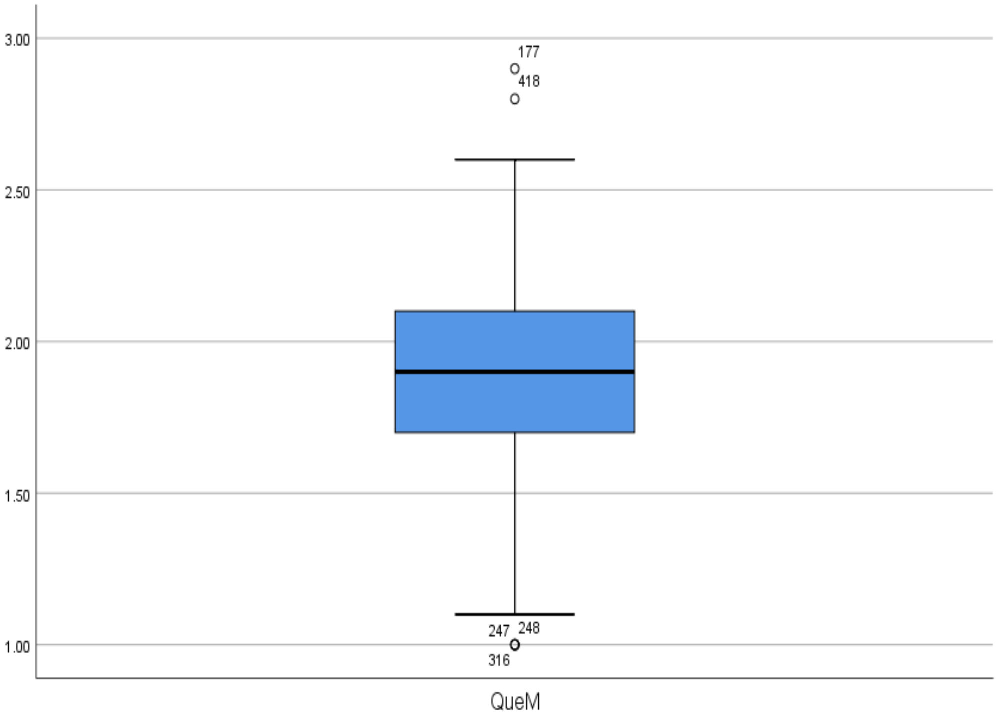


Figure 3.4: BoxPlot QueM

The length of the left and right tails appears to be comparable. This is not quite as dramatic as it was for the weights, despite a few high-end outliers and a median that is somewhat to the left of the center. In general, the distribution`s center seems to be the point at which the heights are symmetrically dispersed.

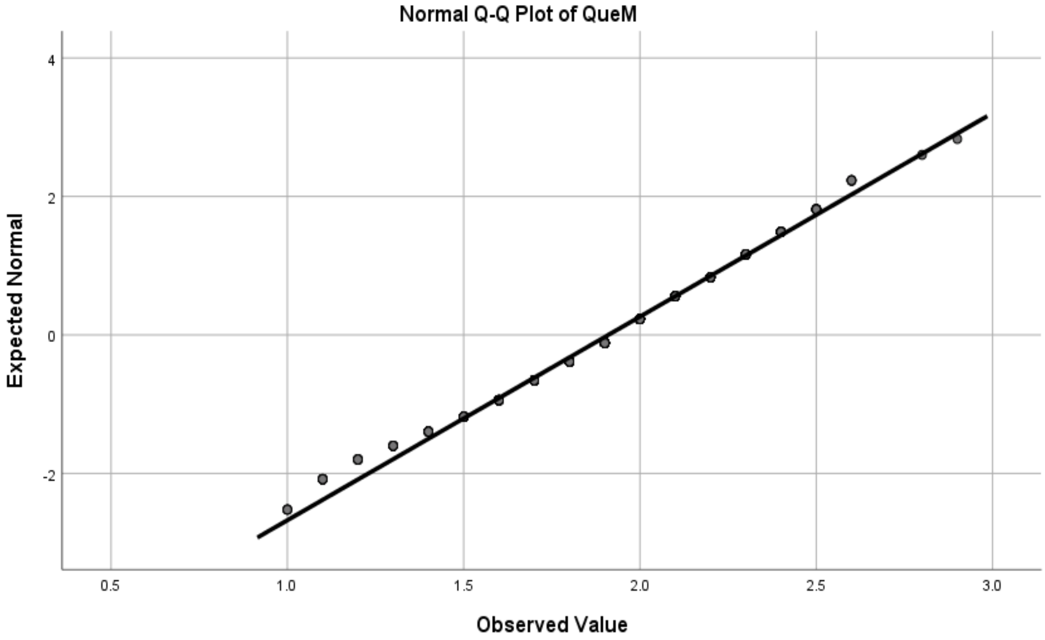


Figure 3.5: Normal Q-Q plot of QueM

A horizontal line illustrating what would be anticipated for that value if the data were normally distributed is shown in the detrended normal Q-Q figure on the right. Any numbers below or above indicate how much the value differs from what would be anticipated if the data were regularly distributed, accordingly.

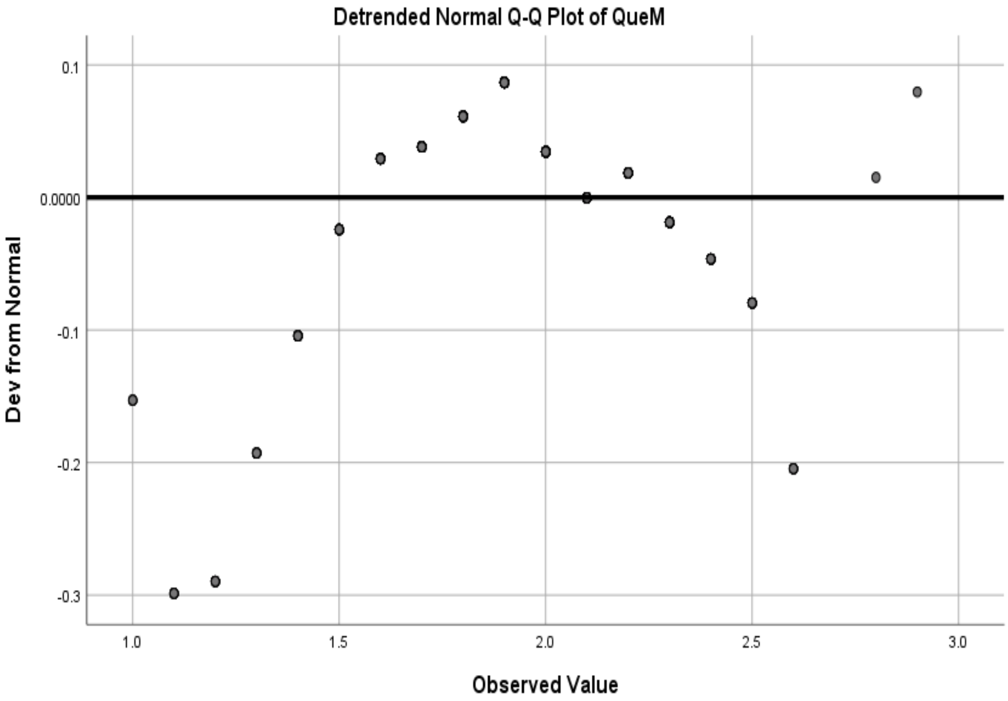


Figure 3.6: Detrended Normal Q-Q Plot of QueM

The normal Q-Q plot demonstrates that nearly all of the measured height values are in line with what we would anticipate if the data were normally distributed. The tails seem to be where the discrepancies are most prevalent. This is magnified by the detrended normal Q-Q plot, which enables us to determine the magnitude of the current deviations: The y-axis of this figure reveals that the standard deviations fall between -0.2 and 0.6. The discrepancies don't appear to follow any clear pattern as the weights did.

This figure3.7 was show the histogram of the column in QueM. That’s histogram was frequency explore for the dataset std dev.

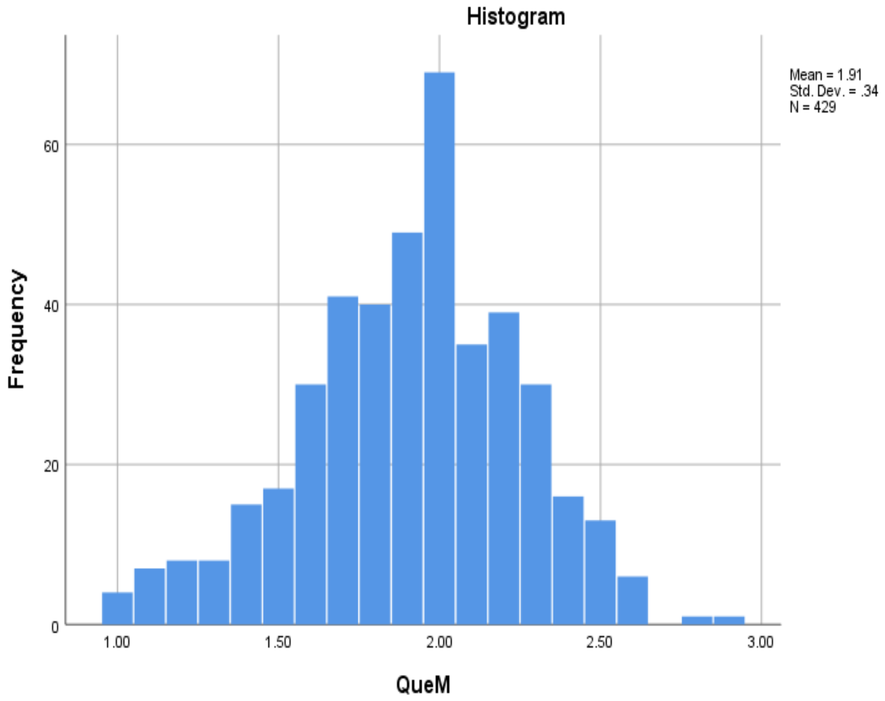


Figure 3.7: Histogram of QueM

In QueM histogram was show highest value is 68% and std. dev is .34. This is the perfect value for measurement descriptive analyze. We would anticipate that the height of the bars would coincide with a normal overlay if the data were exactly normally distributed. Even though this graph cannot determine if the English test results are normally distributed, it does show that the data is fairly symmetrically distributed around the mean and that there do not seem to be many significant departures from the normal distribution.

3.7 Implementation Requirements

Weka 3.9.6, IBM SPSS Statistics 26 software was principally utilized to implement the implementation experiment. The experiment then needed a set of algorithms. collecting information from participants using a question form. Data must be organized after collection in order to remove those that are irrelevant to the issue at hand. Finally, several methods are employed to assess the data.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

I aim to forecast the most accurate result for genders that have been identified. As described in the (Research Methodology) section, we selected a few of the top algorithms for this. Our major objective was to create a dataset on which we could execute computations and maintain records. We were able to understand the barriers and possible results thanks to this trial. This section discusses the results of certain algorithms that were applied on the processed dataset. I used RandomCommittee, IBK, Kstar, DecisionTable and RandomizableFilteredClassifier algorithm. You can use the results to determine which algorithm has the highest accuracy. We used a two-step process to determine the correctness. Accuracy is tested both before applying engineering techniques to raw data and after applying preprocessing and feature engineering techniques to processed data. Follow these three steps to calculate with the highest possible accuracy. I have a gender detection dataset.

4.2 Experimental Results & Analysis

Five machine learning methods have been applied. They then calculated their sensitivity, accuracy, confusion matrix, F1 score, accuracy, recall, and specificity and compared them to each method. Our dataset has 19 features. This part describes the algorithms used to measure higher accuracy. Applying machine learning to selected qualities is one of the greatest achievements of our research. We evaluated the features used in this phase of creating annotated datasets. This is explained below.

The predictions for each basic classifier are calculated after building a set of basic classifiers with different random seeds on a random board. The final prediction is obtained by averaging the classification predictions of several basic classifiers.

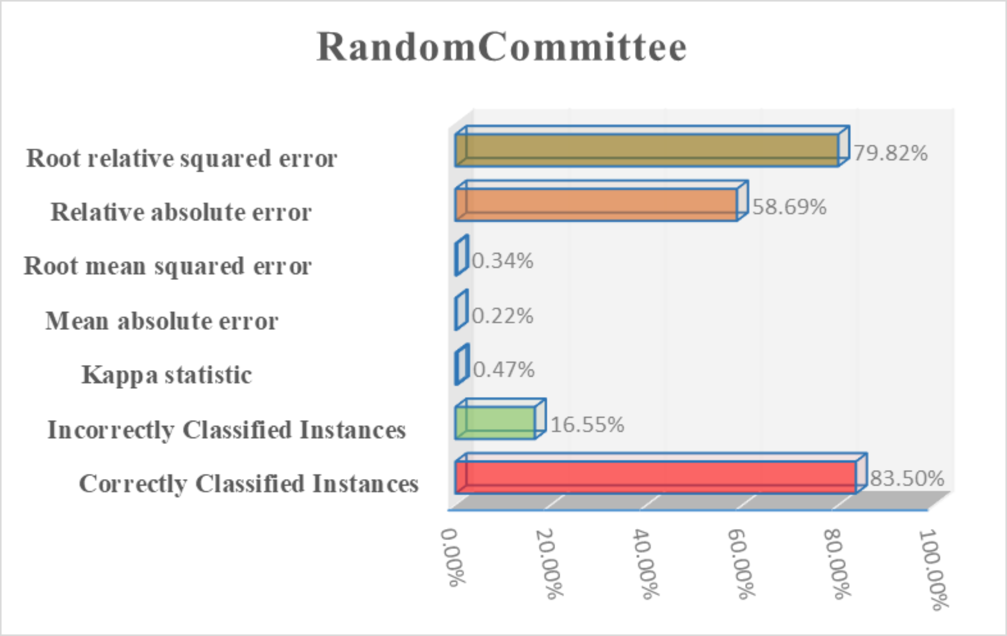


Figure 4.1: RandomCommittee Algorithm

Applying the Random Committee Algorithm, I can find that the accuracy is (83.50%) and this is showing the highest accuracy these all algorithm which was I apply. And Cases That Were Wrongly Classified (16.55%).

A heuristic search method is used to find the k shortest routes. The k shortest pathways between two nodes in a directed weighted network are found using the K * directed search method, which is discussed on this page.

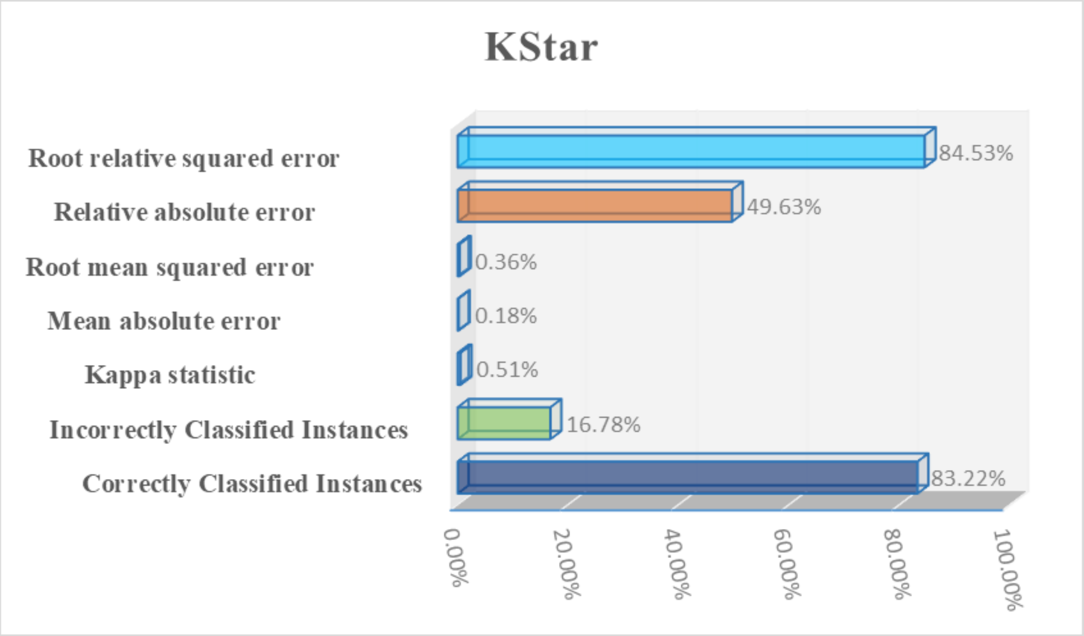


Figure 4.2: KStar Algorithm

Applying the KStar Algorithm, I can find that the accuracy is (83.22%). And cases That were wrongly classified (16.78%).

Instead of building a model, the IBk algorithm creates a real-time prediction of the test situation. For each test case of training data, the IBk method uses distance metrics to find k "near" instances and make predictions based on those instances.

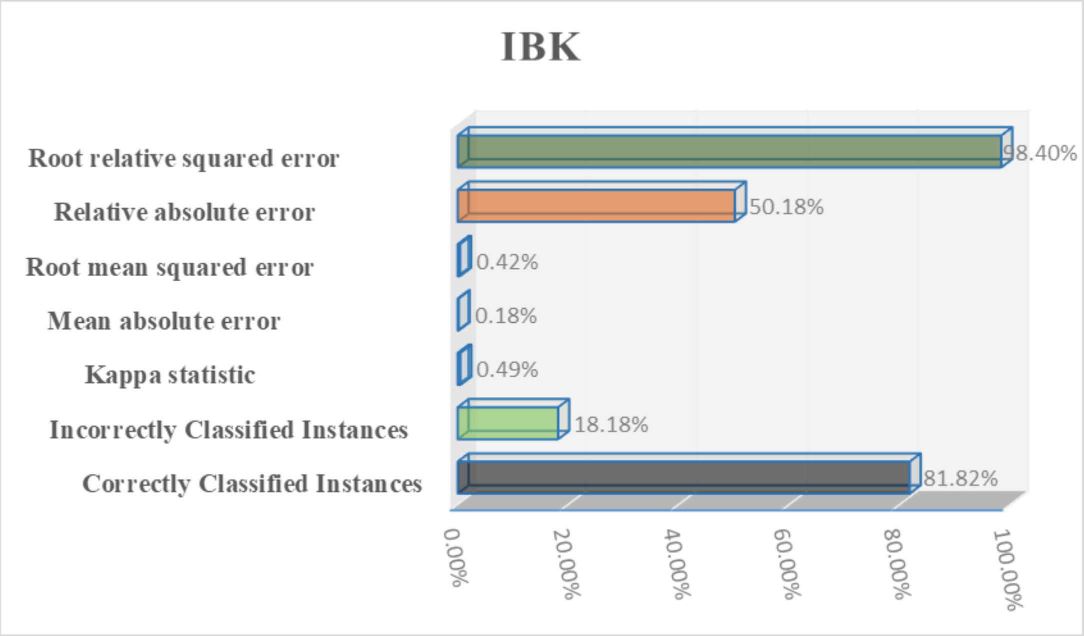


Figure 4.3: IBK Algorithm

Applying the IBK Algorithm, I can find that the accuracy is (81.82%). And cases That were wrongly classified (18.18%).

A decision table is a planned rule logic entry that is displayed as a table and contains conditions (represented by the row and column titles) and actions (represented by the crossing points of the conditional cases in the table), both of which are referred to as conditional cases. The use of decision tables is ideal for business rules with several criteria.

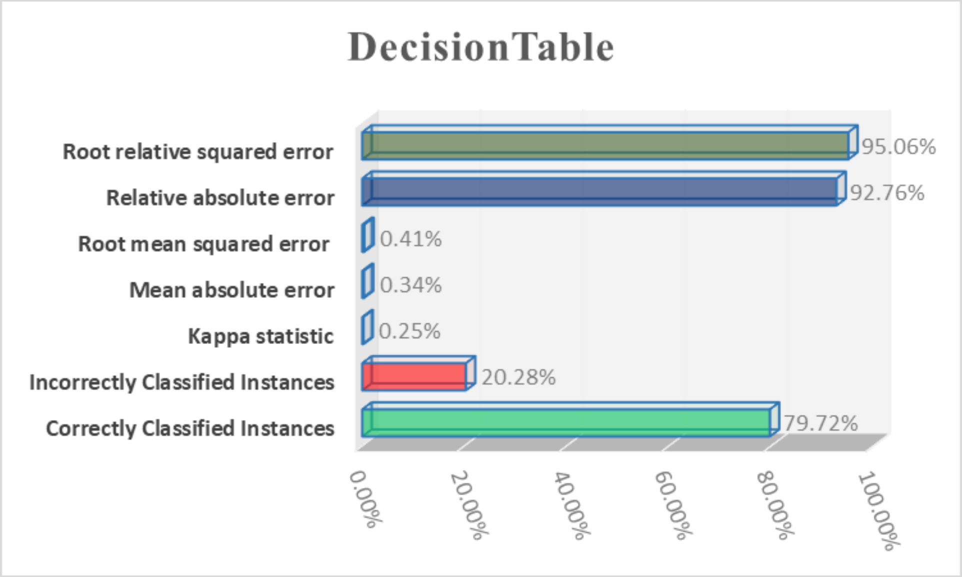


Figure 4.4: DecisionTable Algorithm

Applying the decision table Algorithm, I can find that the accuracy is (79.72%). And cases That were wrongly classified (20.28%).

Randomizable Classifier With Filtering. This approach used an arbitrary classifier to data that had been routed via an arbitrary filter. Like the classifier, the filter's structure is completely dependent on training data, and test instances are handled without altering that structure in any way.

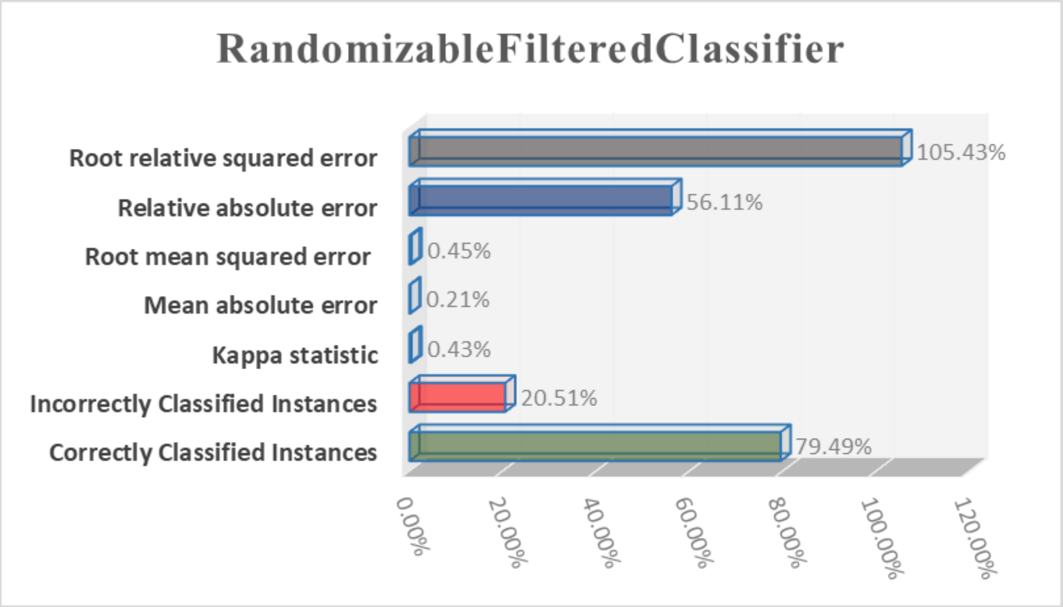


Figure 4.5: Randomizable Classifier algorithm

Applying the Randomizable Classifier Algorithm, I can find that the accuracy is (79.49%). And cases That were wrongly classified (20.51%).

4.2.1 Expressive Analysis

In addition to the Fscore, accuracy, recall, specificity, sensitivity, and confusion matrix for each approach, we also calculated the Fscore, accuracy, recall, and specificity, sensitivity, and confusion matrix. Before a model is selected, it must all be reviewed. To build a model, you must measure a certain categorization. The test dataset is used to make classification measurements for more sophisticated measures.

Sensitivity is the ability of a test to accurately identify the real positive rate.

Sensitivity = TP / (TP + FN) × 100% (i)

A test's specificity is its capacity to recognize the genuine negative rate with accuracy. Specificity =TN / (FP + TN) × 100% (ii)

Recall is the quantity of true positives discovered. Since this is the case, the ratio of genuine positive value to true positive value is true positive value to true positive value.

Recall = TP / (TP + FN) × 100 % (iii)

Precision is defined as the proportion of relevant elements that were positively classified. In the end, everything boils down to the ratio of actual positive to projected positive value. This may be used by a test by only reporting positive for the outcome that is most certain.
Precision = TP / (TP + FP) × 100% (iv)

The weighted average of Precision and Recall is known as the F1 Score. As a result, this score considers both false positives and false negatives.
F1 score = (2 x Precision x Recall) / (Precision + Recall) x 100% (v)

A technique for forecasting classification outcomes in a machine learning problem is the confusion matrix. This differs from the intended objective values predicted by the machine learning model. It gives an overview of how well our categorisation model is working. The kind of errors we make can also be identified. It is essential in figuring out how successful a classifier is.

Table 4.1 displays the confusion matrix for the techniques we employed. Each categorization is fully described in the table below.

Table 4.1: Confusion Matrix of all Classifiers.

Algorithms	Confusion Matrix			Algorithms	Confusion Matrix		
Random Committee	True Class			KStar	True Class		
		Male	Female			Male	Female
	Male	310	15		Male	303	22
	Female	61	42		Female	47	56
	Predict Class				Predict Class		
IBK	True Class			Decision Table	True Class		
		Male	Female			Male	Female
	Male	296	29		Male	287	35
	Female	45	58		Female	54	29
	Predict Class				Predict Class		
Randomizable Filtered Classifier	True Class						
		Male	Female				
	Male	288	37				
	Female	46	57				
	Predict Class						

Table 4.2 displays each method's performance. The optimum approach for our model will be chosen based on the algorithms' accuracy as well as their performance. RandomCommittee is without a doubt the best based on its accuracy, specificity, and precision.

Table 4.2: Classifier Performance Table.

Algorithm Name	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Recall (%)	F1 Score (%)
Random Committee	83.50	88.6	55.3	86.2	88.6	87.30
KStar	82.94	98.1	34.9	82.6	98.2	89.72
IBK	82.24	95.3	40.7	83.6	95.4	89.11
Decision Table	81.72	94.4	38.8	84.3	96.7	88.12
Randomizable Filtered Classifier	80.37	88.9	53.3	85.8	88.9	87.32

4.2.2 Bar graph for Accuracy

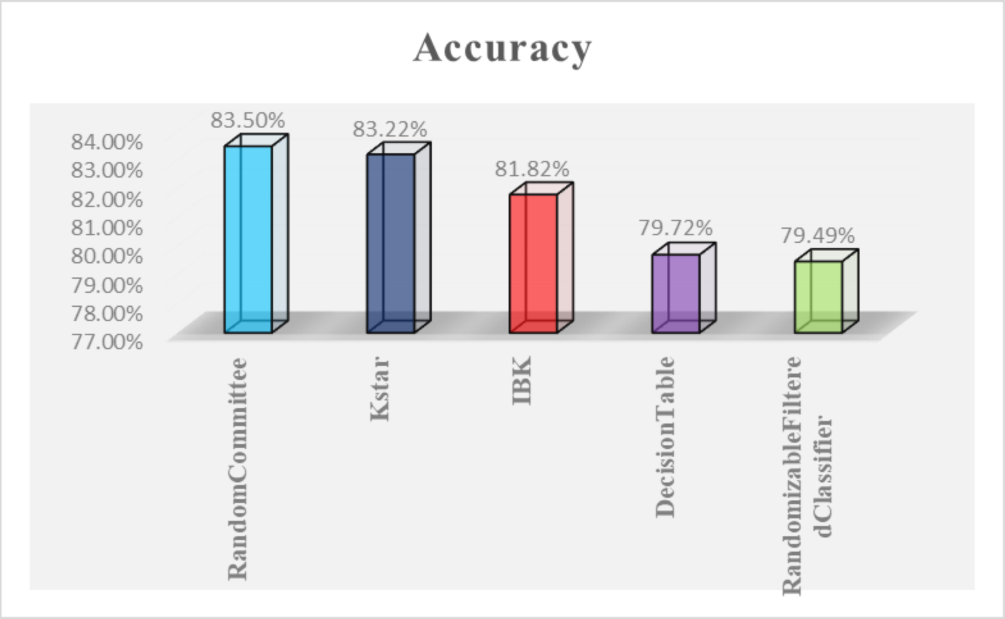


Figure 4.2.1: Bar Chart for Accuracy

4.3 Summary

The IT sector has seen an unparalleled shift during the past 20 years. Smartphone advancements over the past 20 years have improved lifestyles. Our research on gender recognition was excellent. We also attempted it without telling anyone who we were. It made accurate forecasts the majority of the time. When a gender uses a smartphone, the gender may be accurately predicted. The distribution mechanism for smartphones will benefit greatly from it.

CHAPTER 5
SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

5.1 Summary of the Study

Regarding the use of cell phones, everyone has a different perspective. In developing nations, women use mobile phones less often than men do, yet in middle- and low-income nations, 80 percent of women possess smartphones. Women utilize the internet in greater than half of all users. Women are 57 percent less likely to use the internet in North Asia than males are, and they are 27 percent less likely to use it. The gender gap in mobile phone usage is especially obvious in low- and middle-income countries. A smartphone is the most essential form of communication in the world today. Additionally, it's the easiest approach to overcome challenges and learn new things. A segment of our society will fall behind in picking up new skills if the gender gap in cell phone use continues, disconnected from the outside world. Additionally, it affects the economy and GDP. The gender pay gap differs by area. Therefore, before to making any choices, investors must have a thorough knowledge. As a result of the research, we are better able to understand the barriers and causes of women's lower access to mobile devices and mobile internet. The obstacles that cause the gender gap must be removed if investors are to succeed. By fully rewarding users and eradicating gender inequities, the mobile phone revolution may be extended.

5.2 Conclusions

In this thesis, we sought to determine gender using smartphone usage patterns. For this purpose, we choose the most respectable methods for gender recognition. As a consequence, eleven distinct algorithms— RandomCommittee, IBK, Kstar, DecisionTable and RandomizableFilteredClassifier algorithm —are used to analyze this data. All of the datasets that were gathered were unique from one another. It leads to better outcomes for us. The Random Forest method yields the greatest results when compared to other algorithms.

5.3 Recommendations

I recommend,

- The requirements of a user must be understood. One choice is online consulting.
- The design and application of gender-specific policies should be improved.
- Improving the accuracy of gender-related statistics, coming up with initiatives, and monitoring results.
- Increase awareness of the benefits of internet and mobile phone use. Consult customers, mostly women, when developing mobile design and execution methods. Create a user-friendly mobile environment.
- Extend the reach of your distribution and marketing strategies. Make affordability more affordable.
- Extend the reach of your distribution and marketing strategies. Make affordability more affordable.

5.4 Implication for Further Study

- Business opportunity: Many smartphone operators may be motivated by closing the gender gap. If operators can close these gaps by 2023, there will be a revenue of 140 billion dollars.
- Potential for economic growth: Bridging the gender gap is essential for future economic growth. These nations may see an increase in GDP of $700 billion over the following four years by eliminating gender disparities in smartphone internet usage. North Asia has the highest probability and has the greatest gender disparity.

APPENDIX

Basically, in our research, we examine the data and use machine learning techniques to build the model. You can use the proposed model to predict gender. Our main task is to determine gender based on the usage behavior of mobile phone users. We must evolve to catch up with this modern world. In addition, the number of mobile users around the world is increasing every day. When I try to buy a mobile phone now, I can't find the model I want. Due to the fact that usage remains the same regardless of gender. However, this issue is not taken into account by mobile device manufacturers. This gives consumers access to multiple features that go beyond what they really need. This algorithm was developed to predict gender detection based on smartphone usage trends. That’s why when we recognize gender using smartphone pattern then we can understand that the user is boy or girl and it’s through smartphone build up company made their phone requirement of this pattern that is use a boy or girl. That is the main scope of my paper to recognize gender using smartphone pattern.

REFERENCES

[1] Davarci, E. and Anarim, E., 2021, August. Hybrid architecture for gender recognition using smartphone motion sensors. In 2021 29th European Signal Processing Conference (EUSIPCO) (pp. 801-805). IEEE.

[2] Jain, A. and Kanhangad, V., 2016, March. Investigating gender recognition in smartphones using accelerometer and gyroscope sensor readings. In 2016 international conference on computational techniques in information and communication technologies (ICCTICT) (pp. 597-602). IEEE.

[3] Meena, T. and Sarawadekar, K., 2020, November. Gender recognition using in-built inertial sensors of smartphone. In 2020 IEEE REGION 10 CONFERENCE (TENCON) (pp. 462-467). IEEE.

[4] Agneessens, A., Bisio, I., Lavagetto, F. and Marchese, M., 2010. Design and implementation of smartphone applications for speaker count and gender recognition. In The internet of things (pp. 187-194). Springer, New York, NY.

[5] Akbulut, Y., Şengür, A. and Ekici, S., 2017, September. Gender recognition from face images with deep learning. In 2017 International artificial intelligence and data processing symposium (IDAP) (pp. 1-4). IEEE.

[6] Cabra, J.L., Mendez, D. and Trujillo, L.C., 2018, May. Wide machine learning algorithms evaluation applied to ECG authentication and gender recognition. In Proceedings of the 2018 2nd International Conference on Biometric Engineering and Applications (pp. 58-64).

[7] Pias, T.S., Kabir, R., Eisenberg, D., Ahmed, N. and Islam, M.R., 2019, October. Gender recognition by monitoring walking patterns via smartwatch sensors. In 2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE) (pp. 220-223). IEEE.

[8] Tuncer, T., Ertam, F., Dogan, S. and Subasi, A., 2020. An automated daily sports activities and gender recognition method based on novel multikernel local diamond pattern using sensor signals. IEEE Transactions on Instrumentation and Measurement, 69(12), pp.9441-9448.

[9] Lemley, J., Abdul-Wahid, S., Banik, D. and Andonie, R., 2016. Comparison of Recent Machine Learning Techniques for Gender Recognition from Facial Images. MAICS, 10, pp.97-102.

[10] Johora Akter Polin;Omayer Khan;Ahmed Al Marouf; (2020). Utilizing Smartphone Usage Pattern for Gender Recognition applying Machine Learning Algorithms. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), (), –. doi:10.1109/iceca49313.2020.9297407

[11] Buyukyilmaz, M. and Cibikdiken, A.O., 2016, December. Voice gender recognition using deep learning. In 2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA2016) (pp. 409-411). Atlantis Press.

[12] Cao, L., Dikmen, M., Fu, Y. and Huang, T.S., 2008, October. Gender recognition from body. In Proceedings of the 16th ACM international conference on Multimedia (pp. 725-728).

[13] Gupta, P., Goel, S. and Purwar, A., 2018, August. A stacked technique for gender recognition through voice. In 2018 Eleventh International Conference on Contemporary Computing (IC3) (pp. 1-3). IEEE.

[14] Chola, C., Benifa, J.V., Guru, D.S., Muaad, A.Y., Hanumanthappa, J., Al-Antari, M.A., AlSalman, H. and Gumaei, A.H., 2022. Gender identification and classification of Drosophila melanogaster flies using machine learning techniques. Computational and Mathematical Methods in Medicine, 2022.

[15] Gauswami, M.H. and Trivedi, K.R., 2018, January. Implementation of machine learning for gender detection using CNN on raspberry Pi platform. In 2018 2nd International Conference on Inventive Systems and Control (ICISC) (pp. 608-613). IEEE.

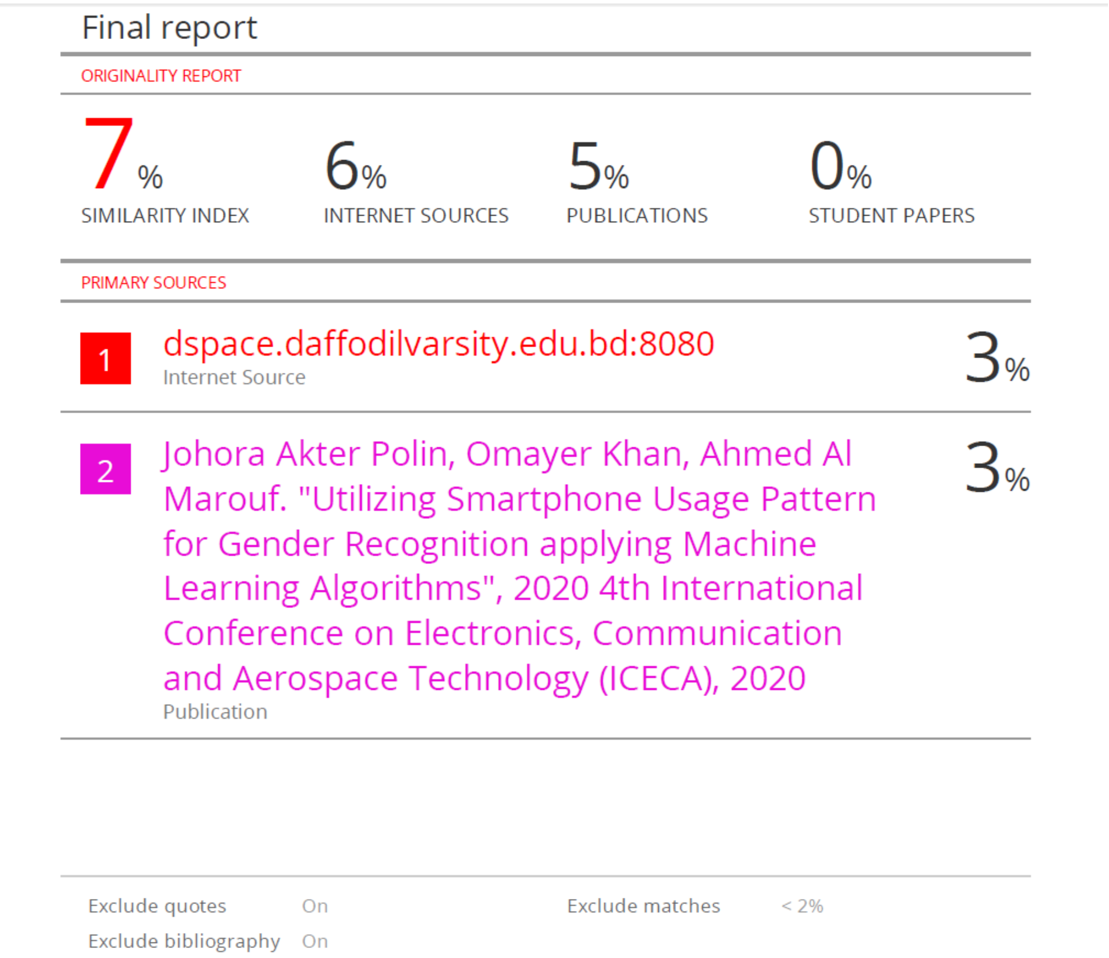
[16] Ertam, F., 2019. An effective gender recognition approach using voice data via deeper LSTM networks. Applied Acoustics, 156, pp.351-358.

[17] Livieris, I.E., Pintelas, E. and Pintelas, P., 2019. Gender recognition by voice using an improved self-labeled algorithm. Machine Learning and Knowledge Extraction, 1(1), pp.492-503.

[18] Li, X., Maybank, S.J., Yan, S., Tao, D. and Xu, D., 2008. Gait components and their application to gender recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(2), pp.145-155.

[19] Jayasankar, T., Vinothkumar, K. and Vijayaselvi, A., 2017. Automatic gender identification in speech recognition by genetic algorithm. Appl. Math. Inf. Sci, 11(3), pp.907-913.

[20] Gupta, S., 2015. Gender detection using machine learning techniques and Delaunay triangulation. International Journal of Computer Applications, 124(6).



Final report

ORIGINALITY REPORT

7%	6%	5%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	Johora Akter Polin, Omayer Khan, Ahmed Al Marouf. "Utilizing Smartphone Usage Pattern for Gender Recognition applying Machine Learning Algorithms", 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020 Publication	3%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		