Smoking Predictor

BY

Sk. Asif Hayder

ID: 151-15-5447

Peyas Chandra Das

ID: 151-15-4854

This report is presented in partial fulfillment of the requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Md. Tarek Habib

Assistant Professor

Department of Computer Science and Engineering

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

NOVEMBER 2018

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We are really grateful and wish our profound our indebtedness to **Md. Tarek Habib, Assistant Professor**, Department of CSE, Daffodil International University, Dhaka, for the deep knowledge & keen interest of our supervisor in the field of "*Artificial Intelligence*" to carry out this project. His endless patience ,scholarly guidance ,continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to the Almighty Allah and **Dr. Syed Akhter Hossain, Professor and Head,** Department of CSE, for his kind help to finish our project and also to other faculty members and the staffs of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parent

.

ABSTRACT

Data Mining can be defined as the use of complex tools of data analysis to discover previously unknown relationships and patterns in large datasets. Therefore, data mining comprises techniques that enable more process than data collection and management, including data analysis and prediction. Healthcare databases have huge amounts of data, and with effective analysis, a great deal of hidden knowledge can be discovered. Therefore, predictive analytics can be particularly useful for analyzing and extracting hidden knowledge in large amounts of data obtained from smokers. Predictive Analytics with data mining has found an application in the medical sector. Data mining in healthcare organizations can transform the raw data held by the organization into useful knowledge with minimal intervention by the user. It also can help to discover new healthcare knowledge for clinical and administrative decision making, as well as producing scientific hypotheses from large sets of experimental data and clinical databases. In the analysis of smoking characteristics, there are a limited number of cases where prediction by data mining has been well utilized. A review on the subject reveals that there is an apparent lack of theoretical and empirical systems that address how this research can be used to understand a person's smoking possibility in the near future. This study aims to build a self-developing system for human health using a data mining technique to predict smoking and help to determine the humans to stay safe. The system is based on a continuous acquisition of data, thereby improving its results regularly.

TABLE OF CONTENTS

List of Figures

CHAPTER 1

Introduction

1.1 Introduction

In the present age smoking is a chronic recurring disorder. Most clinical trials specify, as their endpoint, a single terminal outcome. This sacrifices information regarding the recurring nature about change in smoking behavior that requires new methods to address. For this reason "Smoking Predictor" is a project based on our research which aims to protect human beings around us from going into the worse result of smoking. Smoking is very much injurious to human health. For this worse habit of human being many dies every year. Not only that, this is an influence to our young generations too for selecting a bad life. "Smoking Predictor" can be a first step solution to this problem. It takes the information of a person's surrounding in the age from 18 to 30 and gives a possible result for that person's smoking possibility in the near future. This is the age where many human finds himself facing various smoking caused problems. "Smoking Predictor" gives a possible result regarding a person's smoking possibility. By this, any health care organizations or person can take necessary steps to turn their younger generation to a good future. Thus Human beings can lead a better life apart from smoking related diseases.

1.2 Motivation

Every Year a huge number of people die in various circumstances regarding smoking cautions. In a review top diseases caused by smoking are [1] –

- Lung Cancer
- Heart Disease.
- Diabetes.
- Liver Cancer.
- Erectile Dysfunction.
- Ectopic Pregnancy.

- Vision Loss.
- Tuberculosis.
- Rheumatoid Arthritis.
- Colorectal Cancer.
- Stroke.
- Destroy Immune System.
- Psoriasis.
- Menopause.
- Sudden Infant Death Syndromes.

For these diseases are increasing day by day and more news are coming out in public but they are not preventing the first step or to give support to their young generation, the young generation are going for smoking as it is a trend. In order to get rid of this smoking, people aged from 18 to 30 are the main targets to this smoking caused diseases as per the review. To prevent human beings from this from "Smoking Predictor" give a possible result to determine who can smoke or who can't smoke. By this the younger generation can be turned into a safe future to lead human history.

1.3 Rationale of the Study

Research is an organized investment of a problem in which there is an attempt to gain solution to a problem. To get right solution of a right problem, clearly defined objectives are very important. Clearly defined objectives enlighten the way in which the researcher has to proceed. Research objectives are usually expressed in lay terms and are directed as much to the client as to the researcher. Research objectives may be linked with a hypothesis or used as a statement of purpose in a study that does not have a hypothesis. A researcher objective is clear, concise, declarative statement, which provides direction to investigate the variables. Generally research objective focuses on the way to measure the variable, such as identify or describe them. Sometime objectives are directed towards identifying the relationship difference between two variables. Research objective outline the specific goals the study plans to achieve when completed.

The research objectives are -

- To study data analysis, data mining and predictive analytics.
- To be aware of smoking.
- To develop human health for better purpose.
- To apply the result on the real life.
- To build an easy system for users.

1.4 Research Questions

In this research some questions may raise that-

- Is it really possible to predict the possibility from data?
- Can the results be used for other purpose?
- What are the main applications of this research?

For all those questions it can be said that almost every kind of diseases caused by smoking can be stopped in a very first step. This is possible to predict the possibility of a person's chances of smoking. On the other hand this research can be used in health analysis, decision making, taking necessary steps and so on.

1.5 Expected Output

We decided to work on predictive analytics by data mining as our project because we wanted to build a system which will be able to give a possible answer regarding the addiction of smoking. Building this system which will be able to analyze data for example the smoking status of a person's father/mother, surroundings such as friends, siblings is needed to predict the result. After the analysis, the system will provide a statistical overview result of the predicted situation of targeted people.

Though there are a lot of languages and tools in the market we decided to make our system by using “Python” programming language, where we will be able to predict our desired result to find the possibility of smoking.

1.6 Report Layout

Chapter 1: Introduction

This chapter briefs about what we have discussed the motivation, objectives and the expected outcome of the project.

Chapter 2: Background

This chapter briefs about the background circumstances of our project. We also talked about the related work, comparison to other related systems, the scopes of the problem and challenges of the project.

Chapter 3: Research Methodology

Research Subject and Instrumentation, Data Collection Procedure, Statistical Analysis, Experimental layout can be found in this chapter.

Chapter 4: Experimental Results and Discussion

Results and discussions, Experimental results and Descriptive Analysis are the topics explained here.

Chapter 5: Summary, Conclusion, Recommendation and Implication for future research

The topics are Summary, Conclusion and Future scope.

CHAPTER 2

Background

2.1 Introduction

”Smoking Predictor” is a computational process to predict the possibility of smoking regarding a person. It takes the information regarding the candidate person’s surroundings. It is possible to give a possible result. We have collected and analyzed a huge number of data from different people aged from 18 to 30 from a society and analyzed those data to predict a person’s possibility of smoking. As we have worked with smoking, so this paper contains different data related to smoking only.

Data mining is a computational process of discovering patterns, trends and behaviors, in large datasets using artificial intelligence, machine learning, statistics and database systems. The overall goal of the data mining process is to extract information from a dataset and transform it into an understandable structure for further use. Data mining is considered as a synonym for another popularly used term known as KDD, knowledge discovery in databases. Data mining is an essential step in the process of Predictive Analytics.[2]

Predictive Analytics is the branch of the advanced analytics which is used to make predictions about unknown future events. Predictive Analytics uses many techniques from data mining, statistics, modeling, machine learning and artificial intelligence to analyze current data to make predictions about future. It uses a number of techniques to make predictions about future.[3] Predictive Analytics process works in 7 steps –

- Define Project.
- Data Collection.
- Data Analysis.
- Statistics.

- Modeling.
- Deployment.
- Model Monitoring.

Define Project:

Define the project outcomes, deliverables, scoping of the effort, business objectives, identify the data sets which are going to be used.

Data Collection:

Data mining for predictive analytics prepares data from multiple sources for analysis. This provides a complete view of the customer interactions.

Data Analysis:

It is the process of inspecting, cleaning, transforming and modeling data with the objective of discovering useful information, arriving at conclusions.

Statistics:

Statistical analysis enables to validate the assumptions, hypotheses and test them with using standard statistical models.

Modeling:

Predictive modeling provides the ability to automatically create accurate predictive models about future. There are also options to choose the best solution with multi model evaluation.

Deployment:

Predictive model deployment provides the option to deploy the analytical results in daily decision making process to get results, reports and output by automating the decisions based on the modeling.

Model Monitoring:

Models are managed and monitored to review the model performance to ensure that it is providing the results expected.

2.2 Related Works:

Predictive Analytics is widely being used all over the world nowadays using data mining techniques in various purposes on different sectors and platforms in order to reach desired targets in medical purposes. In order to lead a better human life this is becoming a one way solution to what can be the outcome. But a limited number of people have done their research regarding smoking related data but no one gave a possible solution to predict the possibility of smoking of a person. A few important of them are –

- Development and validation of the cigarette smoking consequences looming scale.[4]
  By- D. McDonald, A. Kaufmann, and D. A. F. Haaga.
- Correlates of attempting to quit smoking among adults in Bangladesh.[5]
  By- Shariful Hakim, Muhammad Abdul Baker Chowdhury, Md. Jamal Uddin.
- Association of exercise with smoking-related symptomatology, smoking behavior and impulsivity in men and women.[6]
  By- Tosun, Nicole L. Allen, Sharon S. Eberly, Lynn E. Yao, Meng Stoops, William W. Strickland, Justin C. Harrison, Katherine A. Al'Absi, Mustafa Carroll, Marilyn E.
- Smoking cessation strategies in pregnancy: Current concepts and controversies.[7]
  By- Ioakeimidis, Nikolaos Vlachopoulos, Charalambos Katsi, Vasiliki Tousoulis, Dimitrios.

- Predictors of reduced smoking quantity among recovering alcohol dependent men in a smoking cessation trial.[8]

  By- Worley, Matthew J. Isgro, Melodie Heffner, Jaimee L. Lee, Soo Yong Daniel, Belinda E. Anthenelli, Robert M.

2.3 Research Summary

As there is no work done in this predicting of a person's smoking possibility and by the research paper's we have gone through, it is possible to predict a possible answer to this solution. By this result anyone can understand a possible situation regarding a person. Moreover, this will help to increase more awareness among people too. As a result, we can save young generations to lead a better life. As many people did their research on data science they have created their willing projects using different algorithm and techniques. With the use of data mining, many people did played their spot roles for making the "Artificial Intelligence" as it is at the top of buzz words.

2.4 Scope of the Problem

Predicting the desired result is one of the most popular research works at present all over the world nowadays, because till now though a limited work has been done on smoking related data but still there is a lack of accuracy and understanding for the machine. Therefore many people are working hard on these sectors by different algorithms and various techniques. In this research we are using an algorithm based technique known as Naive Bayes algorithm. Our focus is on the improvement of the accuracy of possibility.

2.5 Challenges:

First challenge was to find a number of people to collect desired information. As a result we had to go through a limited number of people aged from 18 to 30 and in a society to collect information. But analyzing those data sometimes prediction can be wrong too. As artificial intelligence is getting strong day by day it needs time to fix this system to be more accurate for giving more accurate results. Again for being machine language, it is not to be said that this system will give a 100% possible prediction to everyone's possibility, therefore, it can give wrong answers too. A few problems we faced for this system are given below-

- Data preparation.
- Data cleansing.
- Identifying important columns.
- Recognizing correlations.
- Understanding how different algorithms work.
- Choosing the right algorithm for the right problem.
- Deciding the right properties for the algorithm.
- Ensuring the data format is correct.
- Understanding the output of the algorithm run.
- Re-training the algorithm with new data.
- Dealing with imbalanced data.
- Deploying/re-deploying the model.
- Predicting in real time/batch.
- Integrating with the primary application to build data insights into the application and initiate user action.

CHAPTER 3

Research Methodology

3.1 Introduction

This chapter briefs of the majority of the practical work. In order to start the process we need some data to work on. At first we had to collect data as for and after that we had to build a data set. For collecting data we used our own formats. This format had to made contact with our system the server and get data by analyzing them into application's database. Once the data has been collected, it was preserved as unprocessed raw format. In order to get somewhat accurate result from our program we were in need to pre-process the data and make it usable for our purpose. Once pre-process is completed, we can then run our program on the data set to determine results based on several parameters. Last part of this step is concerned about the experiment layout. This experiment layout will later serve as the model for our final product.

3.2 Research Subject and Instruments

As we are using predictive analytics smoking related data as our project, we needed to collect data by hand to hand, going to different people with different ages, aged from18 to 30. By collecting hand to hand data from people, therefore, most of them were unstructured data. So, before starting our research work and implementation we were to consider this as our challenge for predictive analytics.

There are different platforms for data mining such as Python, Orange, Weka, Rattle GUI, Apache Mahout, Hadoop, UIMA, Sentic Net API, Natural Language Toolkit etc. But we selected "Python" as our data mining platform. We selected python because of its user interface and strong library for data processing, data mining and output visualization. Python language is mainly used for data analysis and it is a high level programing language. We created a system app for data collection and using python language. This application can connect to the database and collect data and analyze them.

3.3 Data Collection Procedure

We collected a number of data by hand to hand reaching many people, which gave us access to the information which we will be able to use them for this project. After that, those data were needed to be categorized to build up data sets. Then, those raw data were needed to be stored in the database by categorizing the data sets. This data collection was held on a selected society and the people of aged from 18 to 30. Selecting the age limit was to ensure the successive intensions for this aged limit people are often faces smoking related problems.

3.4 Statistical Analysis:

After all the data are collected, the main task was sorting data. But for processing the data we need to work in two steps. They are:

i. Data pre-processing.

ii. Data analyze.

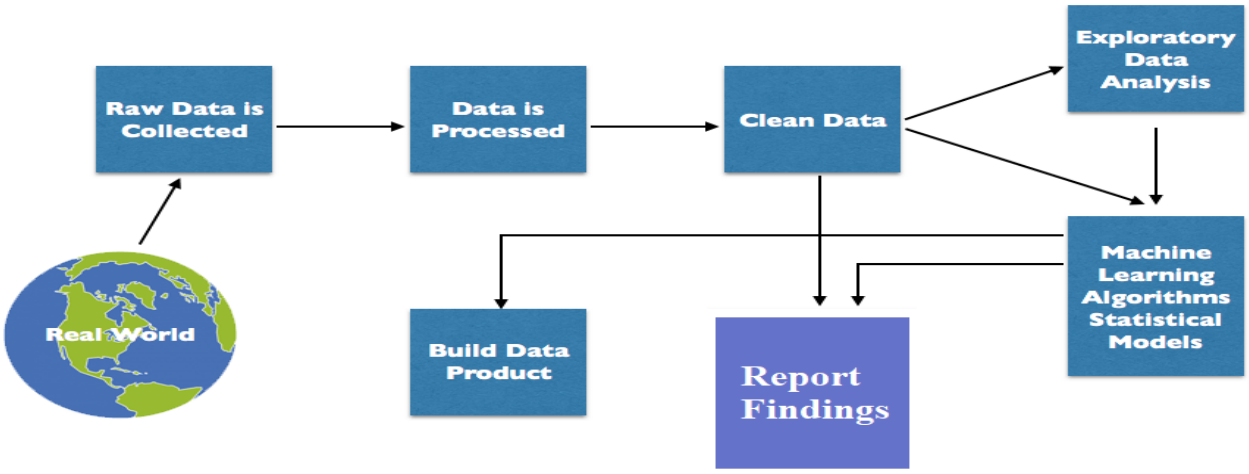

Figure 3.1: Flowchart.

Information Form :

Name	Present Age	Starting Age	Starting Influence	Amount Per Day	Parents Smoking Status	Siblings Smoking Status	Friend's Smoking Status
1. Mahir Khan	26	16	Friend	25	Father	No	Yes
2. Istiak Hasan	25	16	Friend	25	Father	No	Yes
3. Benzir Ahmed	25	17	Friend	30	Father	No	Yes
4. Arabi Imon	26	15	Friend	15	Father	No	Yes
5. Zihad Hasan	25	No	No	0	Father	No	Yes
6. Mahtab Alam	24	16	Friend	20	No	Yes	Yes
7. Mehedi Hasan	26	15	Friend	20	No	No	Yes
8. Arif Bhuiyan	24	17	Friend	15	Father	Yes	Yes
9. Akib Jabed	26	15	Friend	30	Father	Yes	Yes
10. Atul Alam	24	No	No	0	No	Yes	Yes
11. Mazedul Alam	25	16	Friend	20	No	No	Yes
12. Zahid Hasan	24	16	Friend	15	No	No	Yes
13. Mazedul Haque	24	17	Friend	20	Yes	No	Yes

Figure 3.2: Database.

3.5 Implementation Requirements

To implement this research, few things must be done at first. Firstly, the system needs to be installed in the device and based on what we are going to work with, required libraries needed to be installed in python platform. Then the programming part, a database has to be created. Then the raw data which was collected by hand to hand survey, is to be given input on the database. Later from that database, those data needed to be sorted out by categorized portions. There will be an option to give input of a person's surroundings information which the system needs to predict the possible result. If we just follow the instructions as we discussed before, we will get see some user interface options showing different categories of data inputs. After giving input for the data inputs, those data will be stored. After storing the given input data, the system will process those data and analyze them. When the analyzing comes to an end then the system will match given input data with the database and match out all the possible functions regarding. Then the system will prepare an answer for the result. After finishing up the predicted answer, the result is to shown on screen. The result comes by a predictive analytical process to give a possible solution to our need. The system aims to give accurate prediction as it is trained.

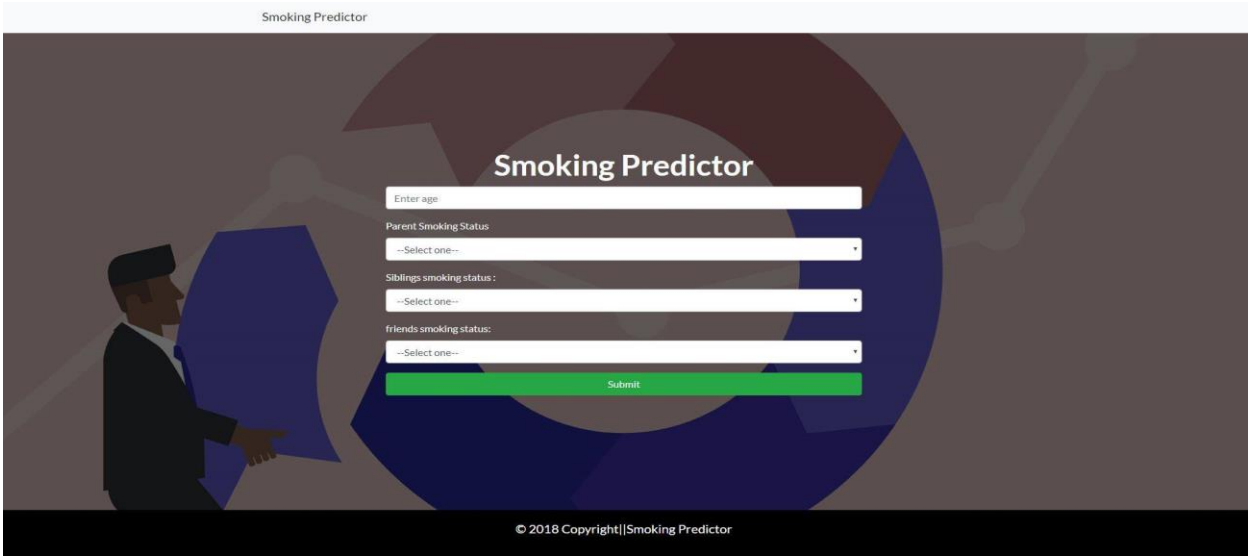

Figure 3.3: Homepage.

CHAPTER 4

Experiment Results and Discussion

4.1 Introduction

We created a user friendly interface for our system so that anyone can use them easily. We used Naive Bayes algorithm on the other hand at the backend and we got four types of data as input. The input data is to be the candidate's data. Before that, a huge number of data has to be given input for training the system to work. These data are used for the system to be trained. It is mainly the trained data which the system has to analyze to predict a possible result.

4.2 Experimental Result

We created this system in a different way for people to use this system easily. Those who have no knowledge about any programming language can operate this system too. Because there is always some people who are interested about the internal structure and play with it and some other people who wants to work using a simple user interface.

So first, we created data sets and connected our data sets with python library, after that it collects data and store them in the working library. After storing the data, it has to be sorted and categorized, our system will retrieve the stored data and sort them as categorized way and it is ready for analysis.

The analysis will show us the complete data sets sorted and ready to use. We stored data of human beings aged from 18 to 30. Among them many people started their smoking before reaching 18 years.

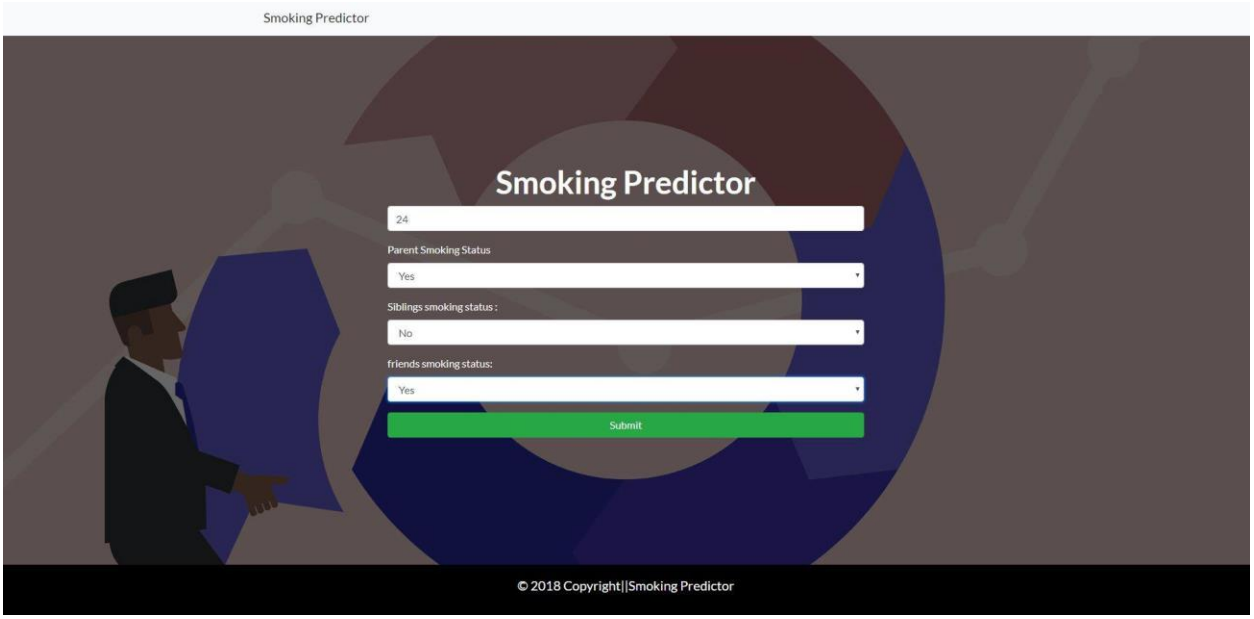

Figure 3.4: Input.

.

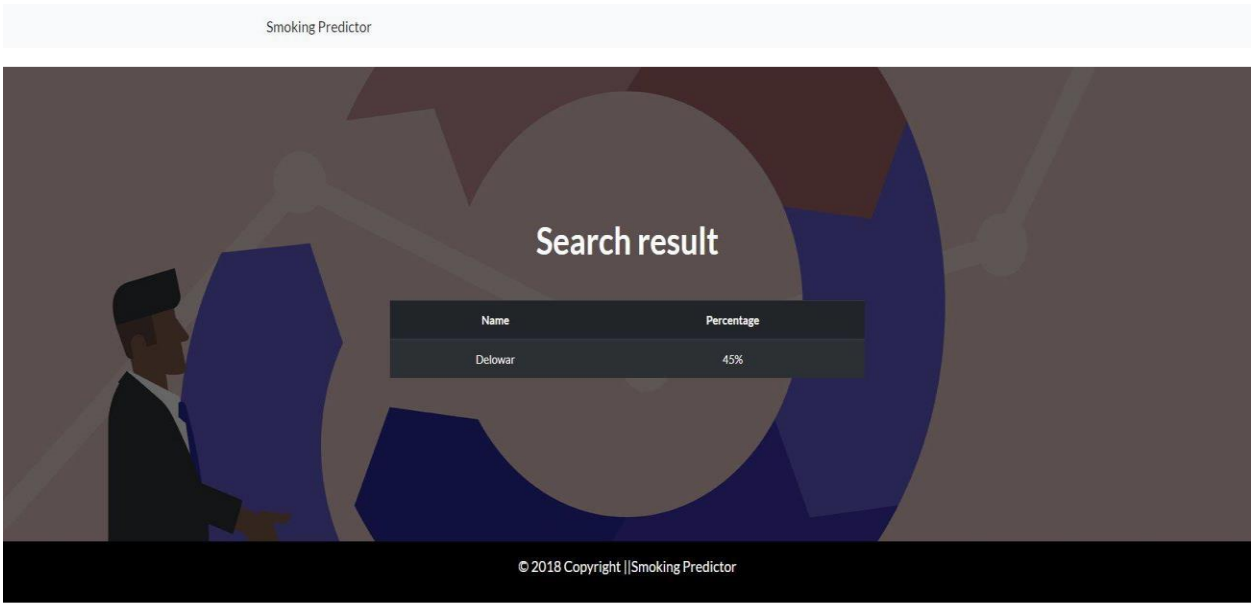

Figure 3.5: Output.

4.3 Descriptive Analysis

In this section, about all types of analysis are discussed briefly. For reminder all of those are analyzed using the naive bayes algorithm in the python platform and python was used because this is one of the best tools for statistical analysis.

4.4 Summary

Though this may seem complex to many people structure wise but it can create clear possible solution on the topic. On the other hand as all those charts are dependent on each other, so it is also a plus point for the user of this system. As if anyone can discover and see the corresponding word result. It will be easy to say about what users will feel after operating this system. If we consider another part, we will notice that using the Naïve Bayes algorithm, it is possible to predict the possibilities which are correct and giving us accurate answer.

CHAPTER 5

Conclusion and Future Scope

5.1 Summary of the Study

As we can see that the output of this project is giving us an overall statistical view for selected or given data. By that, we can understand the result given by the system. The system gives a result based on the pre-processed data sets and after matching with those data sets comes a possible result of the possibility of smoking for a person. Those outputs can be used for different sectors like education, research, medical healthcare systems. We are still working on it to get better result and use it in other types of specific work to help people to lead e planned life. But learning and applying our methods for the future development is our main challenge.

5.2 Conclusion

In this project we tried to develop a complete project on predictive analytics of raw data using data mining techniques. Which will be able to match the data sets and analyze them and based on the analysis, it will provide statistical review of the result. Even it will be easy for non-technical people to use it and examine the output. While we were working this project we learnt a lot of things, also faced a lot of challenges too. From this project anyone can learn data mining and analysis and use the techniques in a better way. Now after having the real time data analysis experience, anyone who wished to work based on this topic can enjoy the work and can do more applications in various sectors.

On the other hand, the main challenge was to collect the resources. There are websites for learning data mining and predictive analytics. Even if anyone faces any problem it is not difficult to find solution for that. That's why anyone can utilize a lot of time in finding the usage of data mining and predictive analytics in various sectors.

Though it took a lot of time to build this system step by step and all those steps were very challenging. For the analysis of this system, python programming language and the related libraries are to be known. Different ideas about different techniques and how to work with them, can easily find its way through to create something new. We tried to not to make it much complex and obtain a high efficiency result from this project as it could.

5.3 Future Scope

Predictive Analytics is already evolving from general to much more complex or more granular and deep understanding. So the demand of predictive analytics on the field of "Artificial Intelligence" is increasing in many sides of medical research, medical improvement, healthcare business purposes and many more to come. Researchers are working on the accuracy of the algorithm and development of the predictive analytics in order to give a possible result. On the other hand, business companies are working on the market policy and customer satisfaction analysis to develop their business.

Out research work has potential of both to be used as commercial aspects or to do further research. For commercial aspect, business companies can find out their customer satisfaction based on these types of predictive projects. They can be able to change their healthcare policy to improve their benefits and attract new generations to do more improvement works on medical sectors connecting them with business purposes too. On the other hand, researchers can collect data sets for individuals to get the process going for another research if the data are related and also use the data for further research to have more accurate results. Even development of accuracy of algorithms can also be done by this research work. In future, we are planning to implement more algorithms in our research work to make it more accurate for predictive analytics to make this part of artificial intelligence open to the people to come forward and do more research works. Thus it can contribute more in this research field by keeping carry on study.

References

[1] "12 Diseases Caused by Smoking | Infographic." [Online]. Available: https://www.unitypoint.org/livewell/article.aspx?id=17ace3fc-fb01-45c3-8617-1beb81404fc4. [Accessed: 20-Nov-2018].

[2] "What is Data Mining? - Compare Reviews, Features, Pricing in 2019 - PAT RESEARCH: B2B Reviews, Buying Guides & Best Practices." [Online]. Available: https://www.predictiveanalyticstoday.com/what-is-data-mining/. [Accessed: 21-Nov-2018].

[3] "What is Predictive Analytics?" [Online]. Available: https://www.predictiveanalyticstoday.com/what-is-predictive-analytics/. [Accessed: 21-Nov-2018].

[4] D. McDonald, A. Kaufmann, and D. A. F. Haaga, "Development and validation of the cigarette smoking consequences looming scale," *Addict. Behav.*, vol. 87, pp. 238–243, Dec. 2018.

[5] S. Hakim, M. A. B. Chowdhury, and M. J. Uddin, "Correlates of attempting to quit smoking among adults in Bangladesh," *Addict. Behav. Reports*, vol. 8, pp. 1–7, Dec. 2018.

[6] N. L. Tosun *et al.*, "Association of exercise with smoking-related symptomatology, smoking behavior and impulsivity in men and women," *Drug Alcohol Depend.*, vol. 192, pp. 29–37, Nov. 2018.

[7] N. Ioakeimidis, C. Vlachopoulos, V. Katsi, and D. Tousoulis, "Smoking cessation strategies in pregnancy: Current concepts and controversies," *Hell. J. Cardiol.*, Oct. 2018.

[8] M. J. Worley, M. Isgro, J. L. Heffner, S. Y. Lee, B. E. Daniel, and R. M. Anthenelli, "Predictors of reduced smoking quantity among recovering alcohol dependent men in a smoking cessation trial," *Addict. Behav.*, vol. 84, pp. 263–270, Sep. 2018.