**CONTENTS**

Title Page No

1. Introduction 1

1.1 Objectives of Research 1

1.2 Problem statement 1

2. Review of Literature 2

3. Data Collection 2

4. Methodology 2

4.1 Exploratory Data Analysis 2

4.1.1 Figures and Tables 3

4.2 Statistical techniques and data visualization 8

4.3 Data Modeling Using Supervised ML techniques 8

5. Finding and Suggestions 9

6. References 10

7. Conclusion 10

**ADVERTISING BASED ON USAGE**

1. **Introduction**

Artificial intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems. These processes include learning (the acquisition of information and rules for using the information), reasoning (using rules to reach approximate or definite conclusions) and self-correction. Particular applications of AI include expert systems, speech recognition and machine vision.

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

**1.1 Objectives of Research**

* The objective of the proposed study wants to shows that there is a relationship/impact

between Age, Daily Time Spent on Site, Area Income, Daily Internet Usage and Gender is directly or indirectly related with whether the person Clicked on Ad or not.

* The objective of this study is to know the whether the Ad is clicked or not.
* Through this study we know the what Ad's are opened and which age groups are visiting the most.
* This study also identifies the how much time people spend on internet and also the area income.

**1.2 Problem Statement**

From this we will predict whether the person clicked on the add or not. By using the dataset consisting of Daily Time Spent on Site, Age, Area Income, Daily Internet Usage and Gender as independent variables and Clicked on Ad as dependent variable.

1. **Review of Literature**

Literature related to Advertising Based on Usage . We included studies that reported on empirical data while leaving out non-empirical studies, such as legal studies. Furthermore, by definition, OBA involves tailoring advertising based on online behavior. Therefore, we excluded studies that addressed personalized advertising based on personal data that were not inferred from online behavior. If prediction is showed as 1 then it means that the user has opened that particular advertisement otherwise he/she didn't open it.

**3. Data Collection**

Data will be collected from secondary resources such as from people, city, country and their daily internet usage and also the time spent on the site. We collect a data from survey or diagnosis. Some tests are conducted through this we collect the data.

1. **Methodology**

For our model, we first looked at several different machine learning algorithms to see which ones to move forward with. Our first step was to split our data into training and test sets using train-test-split, which would allow us to cross-validate our results later. We also stratified the train-test-split, to ensure that the same proportion of our target variable was found in both our training and test sets.

# Stratify our train-test-split so that we have a balanced split  
 x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.2,random\_state=0)

**4.1 Exploratory Data Analysis**

Solutions In this project working with a advertising data set, indicating whether or not a particular internet user clicked on an Advertisement on a company website. We will try to create a model that will predict whether or not they will click on an ad based off the features of that user. This data set contains the following features:

'Daily Time Spent on Site': consumer time on site in minutes

'Age': customer age in years

'Area Income': Avg. Income of geographical area of consumer

'Daily Internet Usage': Avg. minutes a day consumer is on the internet

'Ad Topic Line': Headline of the advertisement

'City': City of consumer

'Male': Whether or not consumer was male

'Country': Country of consumer

'Timestamp': Time at which consumer clicked on Ad or closed window

'Clicked on Ad': 0 or 1 indicated clicking on

**4.1.1 Figures and Tables**

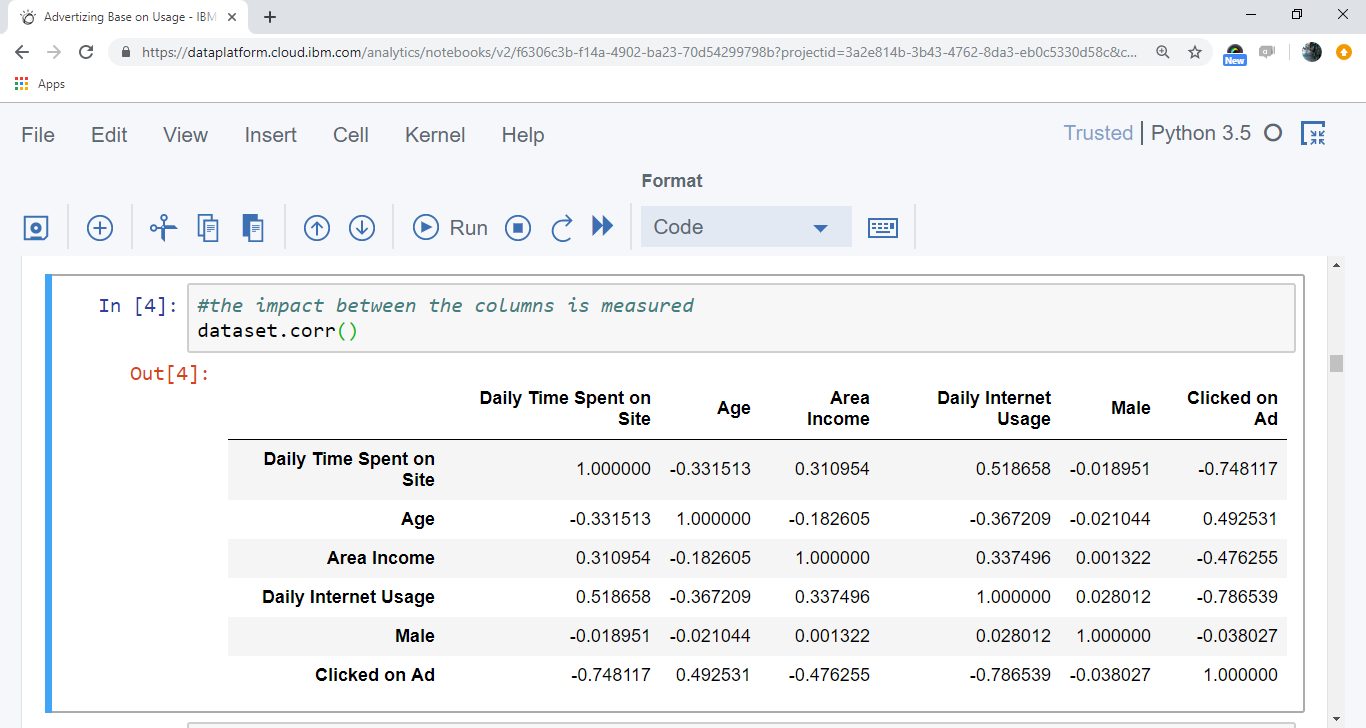


Fig 4.1.1.1 Correlation of the Dataset



Fig 4.1.1.2 AUC Graph after applying Logistic Regression on the Dataset



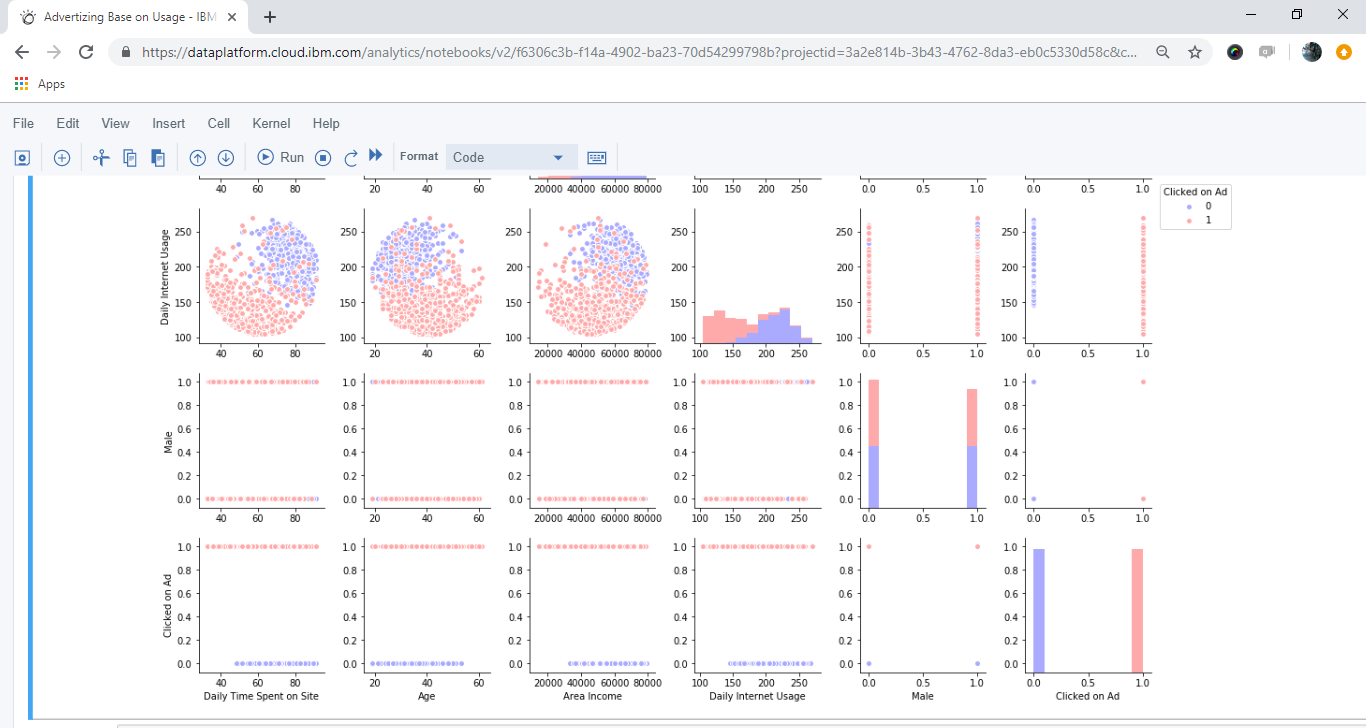


Fig 4.1.1.3 Pair-Plot of the Dataset

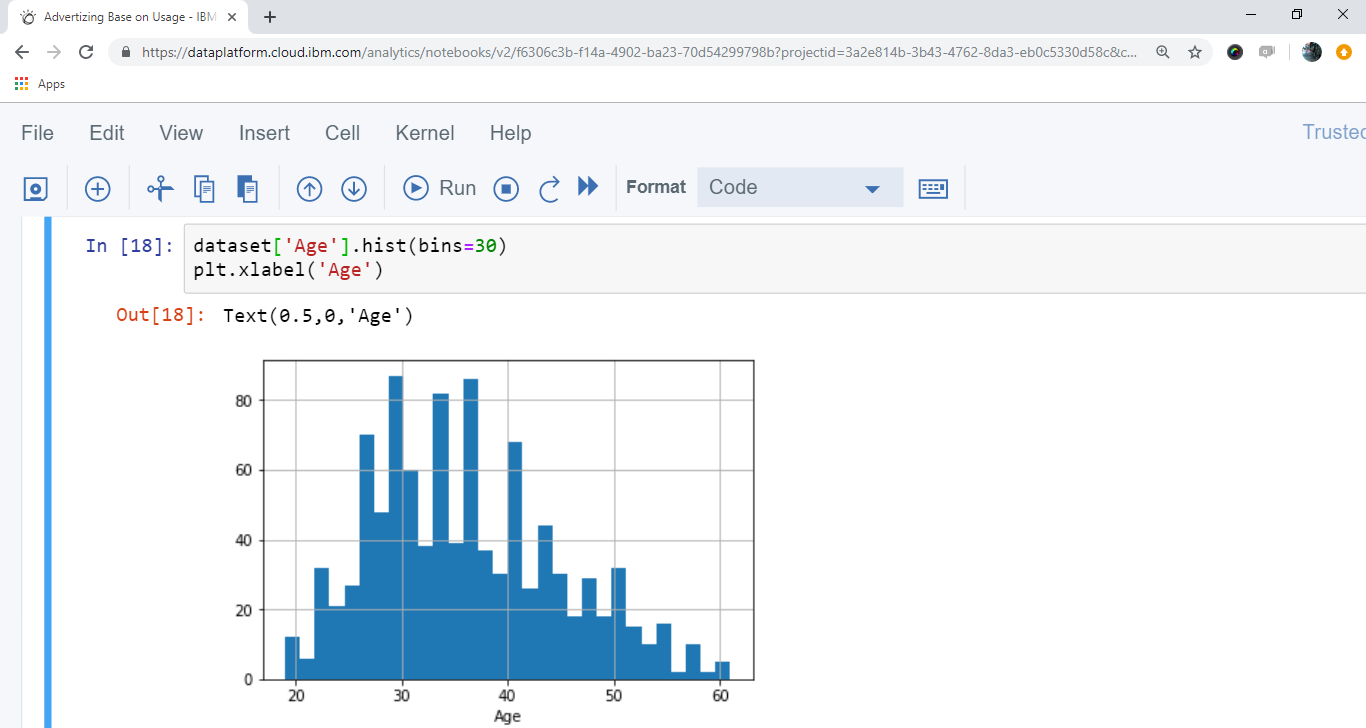


Fig 4.1.1.4 Graph for Age Variations in the Dataset

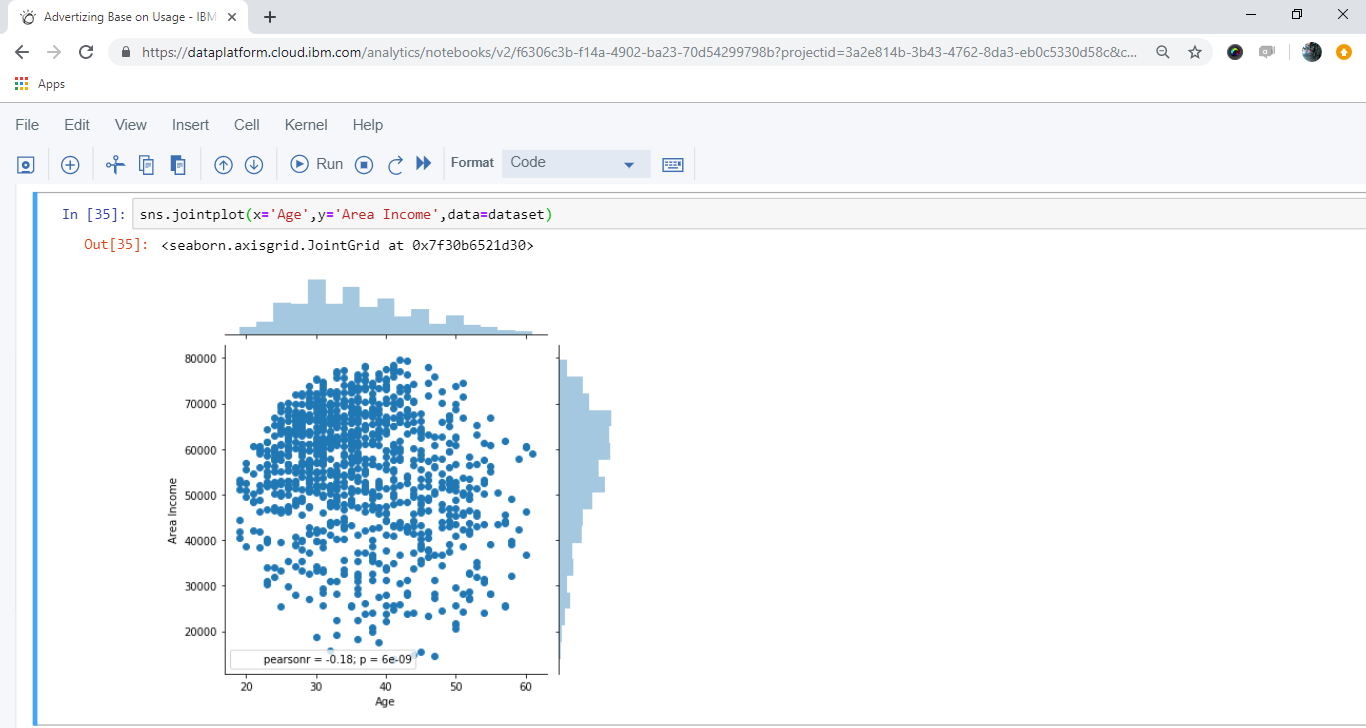


Fig 4.1.1.5 Join-Plot for Area Income and Age



Fig 4.1.1.6 Join-Plot for Daily Time Spent on Site and Age

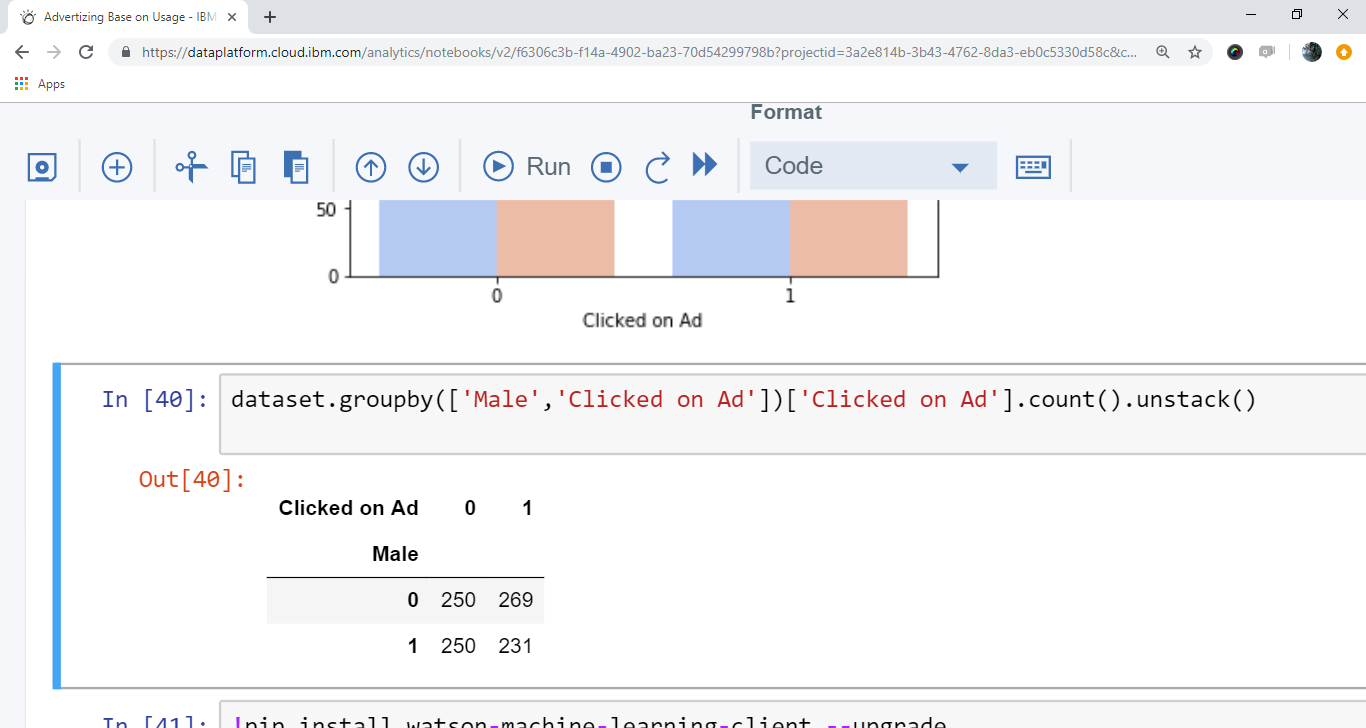


Fig 4.1.1.7 Group-By of Male and Clicked on Ad

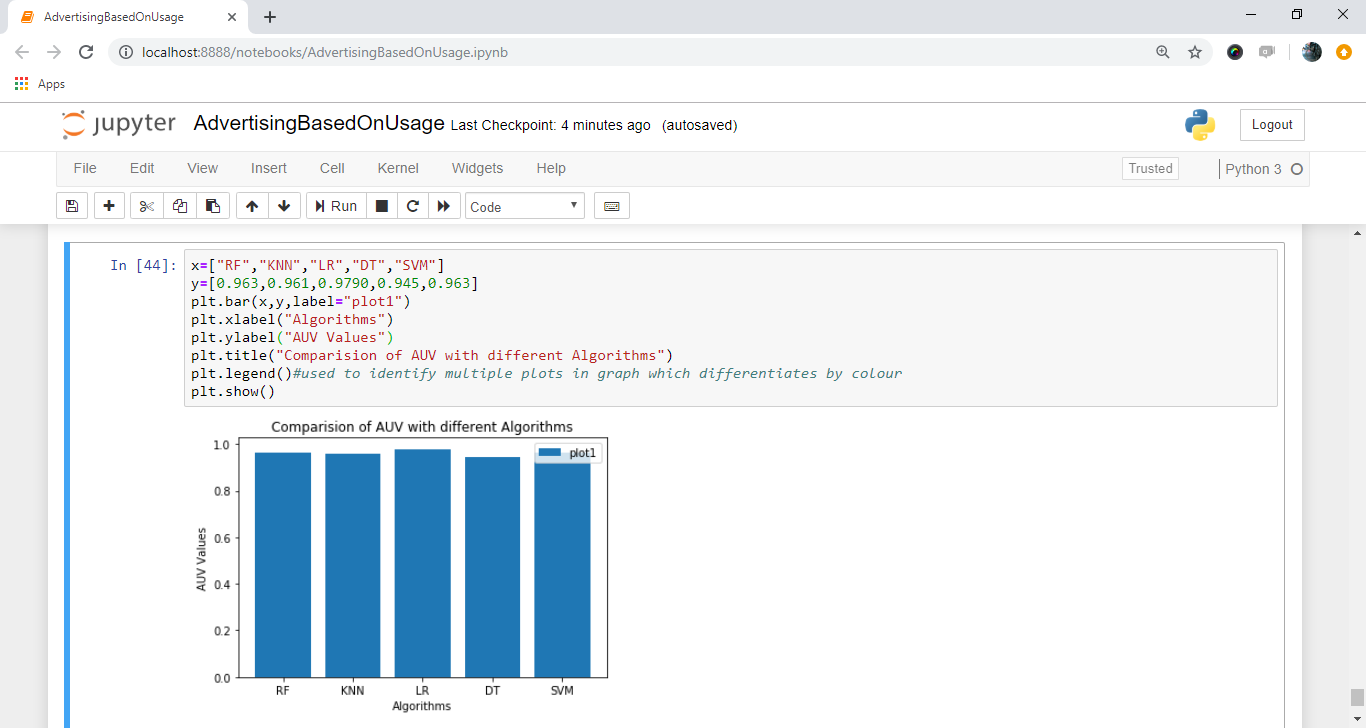


Fig 4.1.1.8 Comparison of AUC using different Algorithms

**4.2 Statistical techniques and data visualization**

By importing matplotlib.pyplot library we have drawn graphs to demonstrate the AUC-ROC curves and by using bar graphs we have visualized the percentage levels of different techniques. And we have used the co-relation function to demonstrate the impact of every factor on each other.

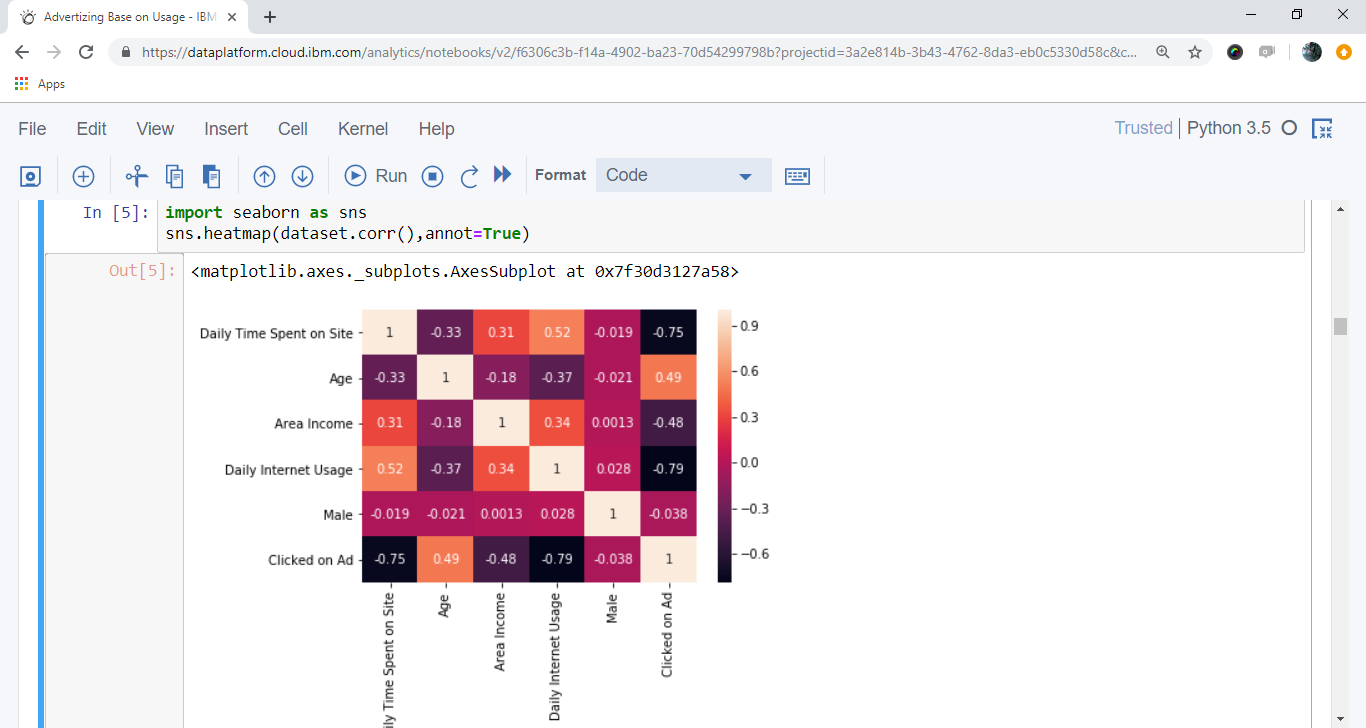


Fig 4.2.1 Heat-Map of the Dataset

**4.3 Data Modeling using supervised ML algorithms**

In general we have two types of learning algorithms, supervised and unsupervised learning algorithms. and in detail it consists of different techniques like, Support Vector Machines, logistic regression, decision trees, k-nearest neighbor algorithm.

Since our model comes under supervised learning algorithm we applied every technique of the algorithm and based on the accuracy values we obtained we chosen Logistic Regression. Logistic regression is a technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values). In this post you will discover the logistic regression algorithm for machine learning. And our model has  dependent variable which is Clicked On Ad means the output can be either a person is opening the Ad or not.

**5. Findings and Suggestions**

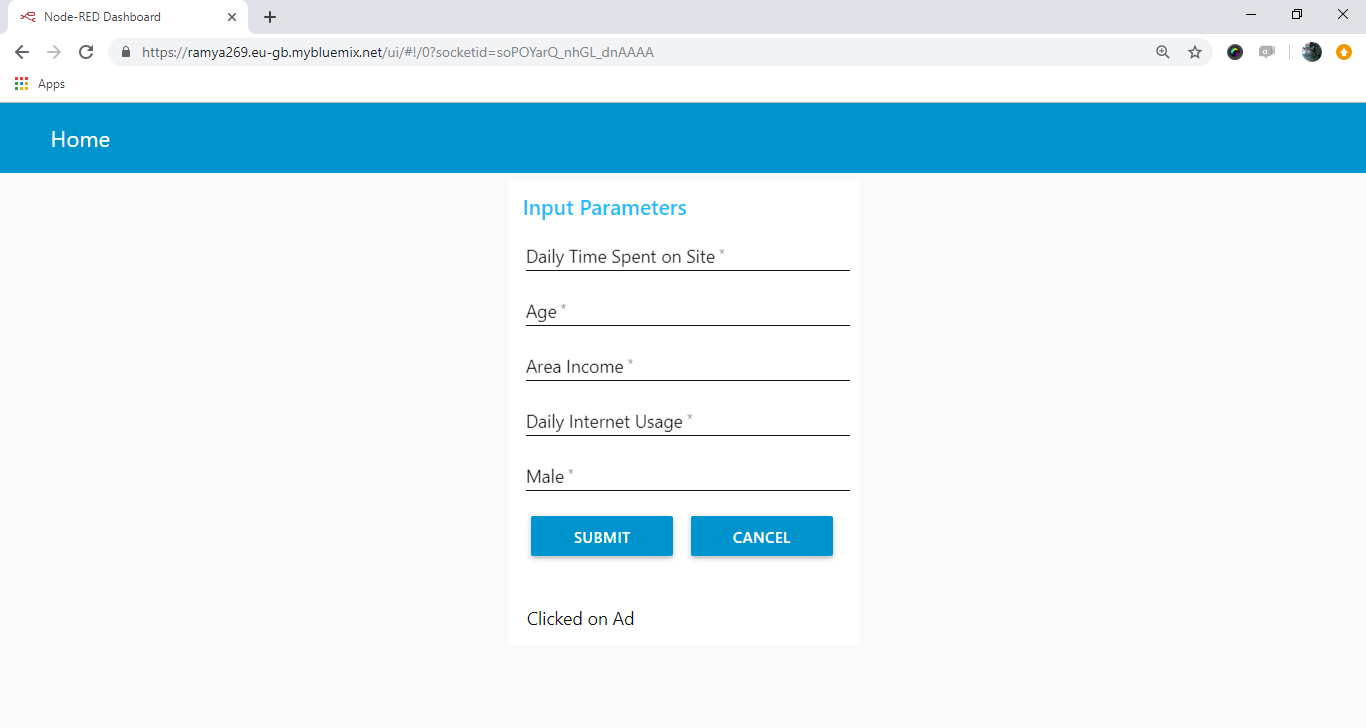
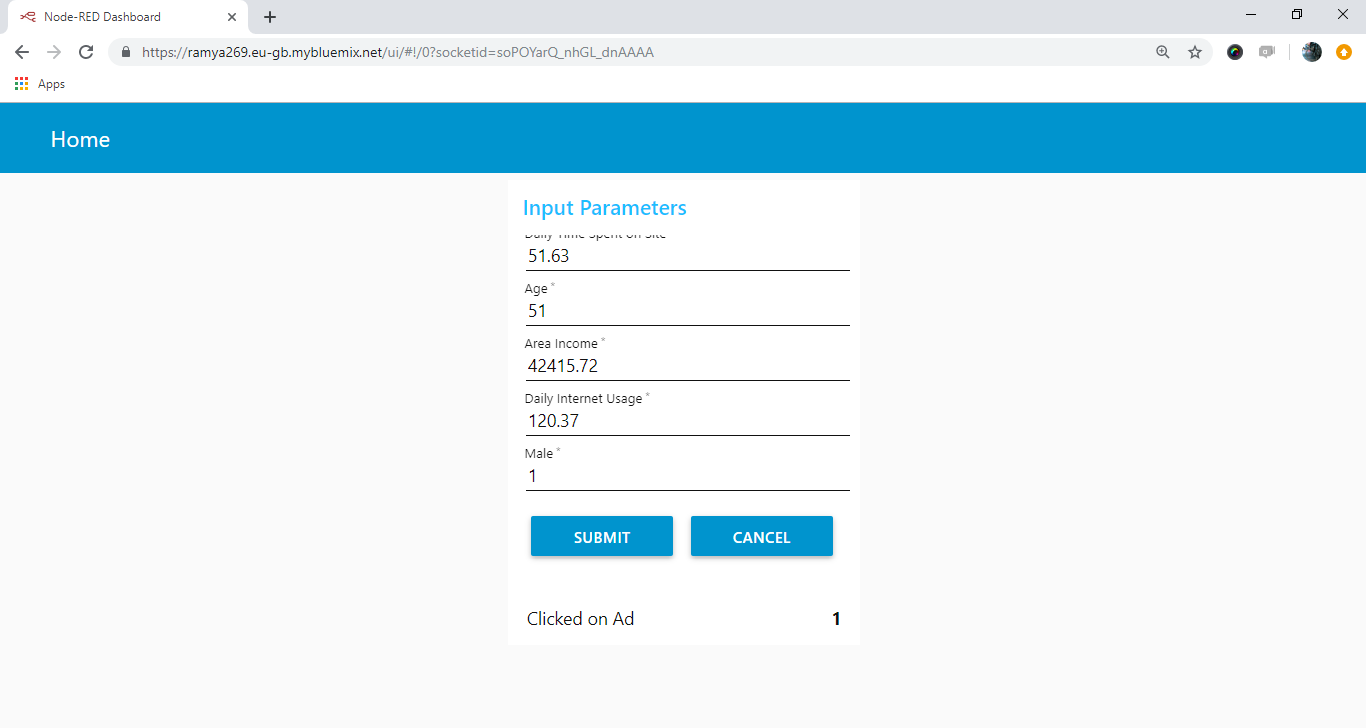
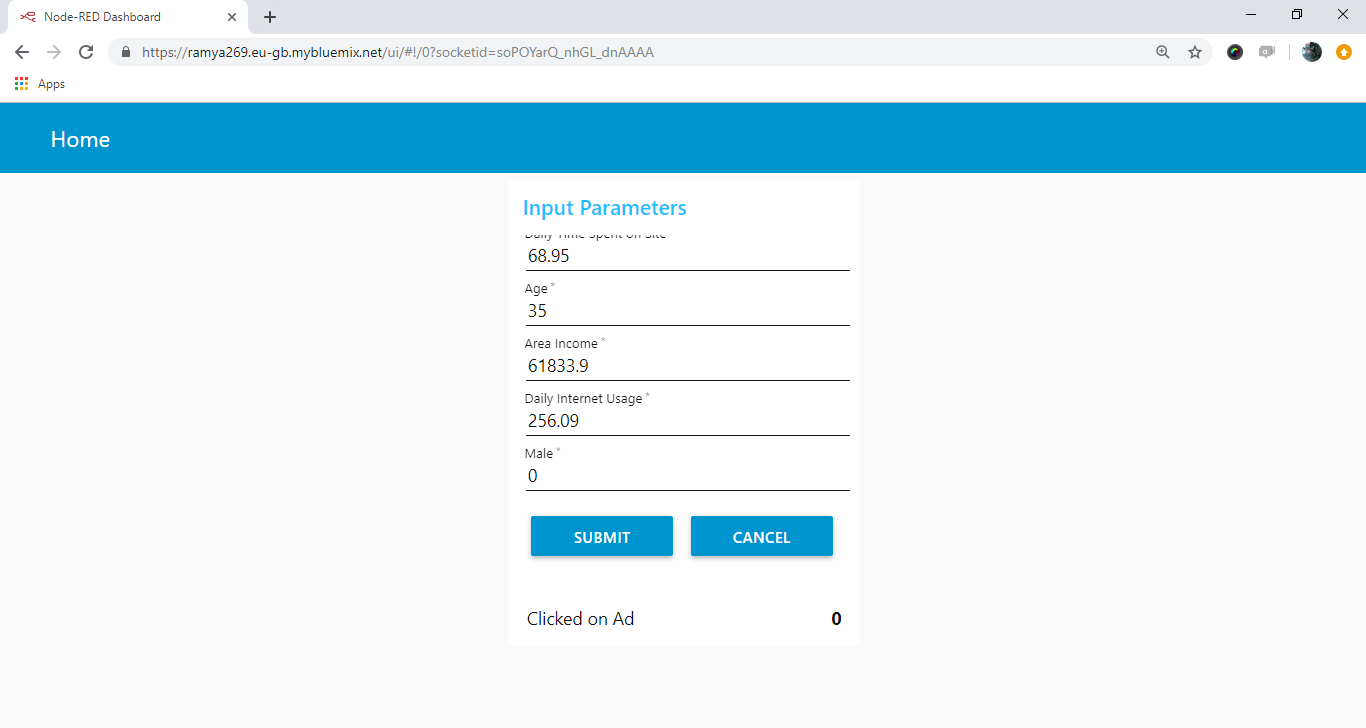


Fig 5.1 User Interface



**6. References**

* <https://www.kaggle.com/fayomi/advertising>
* John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. The Journal of Machine Learning Research, 12:2121–2159, 2011.
* Zhipeng Fang, Kun Yue, Jixian Zhang, Dehai Zhang, and Weiyi Liu. *Predicting click-through rates of new advertisements based on the Bayesian network*. Mathematical Problems in Engineering, 2014, 2014.
* David D Lewis. Naive (bayes) at forty: *The independence assumption in information retrieval. In Machine learning*: ECML-98, pages 4–15. Springer, 1998. 5
* Steffen Rendle. Scaling factorization machines to relational data. In Proceedings of the VLDB Endowment, volume 6, pages 337–348. VLDB Endowment, 2013.

**7. Conclusion**

We investigated methods for data reduction using feature selection. We intelligently removed few fields from the data by viewing the co-relation between the columns . It is evident that unnecessary features act as noise and degrade the performance of the system. Therefore, it is essential to remove these fields. We tried several supervised classification algorithms to see which one will work best for our dataset. The main challenge that we faced in the project was to find ways to efficiently handle huge data with limited resources. The main advantage of using Logistic Regression with stochastic gradient was that we could do computation by reading the data one line at a time. We could not implement Support Vector Machines because the output that SVM gives is not a. So, we leave the investigation of converting SVM output into probability for future work. We could not implement Random Forest due to infrastructure limitations. We plan to try implementing it in the future.