Roommate Compatibility Detection Through Machine Learning Techniques

Manisha Lamba  
Department of Computer science and Engineering  
Guru Gobind Singh Indraprastha University

Raunak Goswami  
Department of Computer science and Engineering  
Guru Gobind Singh Indraprastha University

Mr Vinay  
Department of Computer science and Engineering  
Guru Gobind Singh Indraprastha University

Mohit Lamba  
Department of Electronics and Communication Engineering  
Guru Gobind Singh Indraprastha University

Abstract—Our objective is to develop an artificially intelligent system which aims at checking the compatibility between the roommates of same or different sex sharing a common area of residence. There are a few key factors determining one’s compatibility with the other person. Interpersonal behaviour , situational awareness, communication skills. Here we are trying to build a system that evaluates user on these key factors not via pen paper test but through a highly engaging set of question and answers. Hence using these scores as an input to our machine learning algorithm which is based on previous trends to come up with percentage probability of user being compatible with other user. With the growing population there is always a challenge for organisation and educational institutions to make the students and their employees more and more productive and in such cases a persons social environment comes into play. A person may be a genius but as long as he is not able to work well with his peers there will always be a chance of more productive performance. It is a well- established fact that human are and have always been a social animal and this has helped in creating communities of like-minded people. Many times, even when there are a large no of people employed to do a particular task the result may not be as expected as people may not compatible in working with one another. This at the end creates performance gaps, hinders organisation success and in many cases loss of precious resources. Our intent is not to remove the non- compatible people from the picture but to find out the perfect compatible match for the person elsewhere that will not only save the resources will also enable effective use of resources. Through the use of various machine learning classification techniques, we intent to do this.

Index Terms—compatibility, Interpersonal behavior, resources, situational awareness, classification techniques, industrial readiness, artificial intelligence

I. INTRODUCTION

In computer science, artificial intelligence (AI), sometimes referred to as machine intelligence, is the intelligence presented by the machines, in relation to the natural intelligence displayed by humans [10]. Leading machine intelligence textbooks define this field as study of “intelligent bots”: devices that perceive their environment and takes steps that will maximize its chance of determining its path in achieving its goals. Machine learning (ML) is an application of artificial intelligence that is the scientific study of algorithms and various mathematical models that enable computing systems to perform a specific task [23]. Coming to the target problem it can be clearly observed that there has been a rise in cases of social fights, distrust and decrease in overall efficiency of large institution despite of having a huge and efficient human resource. Here the people working together cannot be blamed since if working individually they can work peacefully and be productive [15]. Choosing the right set of people has always been a challenge for an organization therefore we propose to suggest a solution to this ever increasing problem. As the populations increases so will the chances of selecting the sets of incompatible groups.

Fig. 1 shows the dataflow of our product that will enable people to check whether they are compatible with a particular group or particular organization. Here our software should not be confused with matrimonial compatibility that aims to find a perfect life partner but a software that is made to make a prefect group of people who can work well together and can produce better results than combining their
individual performances [1]. Initially we propose a two user model. Here the user one will be asked a set of forty seven questions, these questions will related to his general behavior and his ability of solving day to day problems [30]. Each question will have options ranging from one to five. The scale is as 1=Disagree, 3=Neutral, 5=Agree. Similarly the user two will also answer these question and their responses will be recorded [22]. On recording the responses a machine learning classification model will classify the responses entered by the users on a five point scale one being least social and five being the most social. Here the thing to note is that two people having the same score determines highest compatibility and people having huge score differences will show a least compatible relation. We have pre-trained our model in order to reduce time of execution on the web application and every time user runs the program he wont have to wait too long for the program to generate the output. After generating the output the outputs will be stored in the SQL database for future reference. The output will be displayed to the users who took the assessment through the web page in form of graphical representation.

II. METHODOLOGY
A. Data Set Selection and pre-processing

Data pre-processing is one of the important step in machine learning Tab. 1 Data gathering methods may often be loosely controlled, which can result in out-of-range values (e.g., Income: -100), impossible data combinations such as (e.g., Sex: Male, Pregnant: Yes), missing values, etc. Analyzing data that have not been carefully screened can produce misleading results [20]. Thus, the representation and quality of data is the first and the foremost before running and observing the analysis. Often, data preprocessing and cleaning is the important phase of a machine learning project, especially in computational biology and artificial intelligence [17] if there is any irrelevant, redundant, noisy and unreliable data present then knowledge discovery during the training phase is more difficult [3]. Data preparation and filtering steps can take considerable amount of processing time. Data preprocessing includes cleaning, normalization, transformation, feature extraction etc [31]. Initially the data that we have used contained many null values and it was our task to clean them up. For faster cleaning [10] we have used Microsoft azure ML studio. After uploading the data to the Microsoft Azure ML studio, we used multiple imputations using chained equations in order to clean up the data and to free up the redundant values. The data set used has been taken from Kaggle . The data set is referred to as the big five personality test. The Big Five personality traits, also called as the five-factor model (FFM) and the OCEAN model, is a taxonomy, or clustering, for various personality features. When factor analysis (a statistical technique) is applied to personality survey data, some words used to describe aspects of personality are often applied to the same person. For example, someone described as conscientious is more likely to be described as "always prepared" rather than "messy". This theory is based therefore on the association between words but not on neuropsychological experiments. This theory uses descriptors of common language and therefore suggests five broad dimensions commonly used to describe the human personality and psyche [2]. Data preprocessing may affect the way in which outcomes of the final data processing can be interpreted. This aspect should be carefully considered when interpretation of the results is a key point, such as in the multivariate processing of chemical data (chemometrics).

B. Algorithmic Implementation

In mathematics and computer science, an algorithm is defined as series of instructions to solve a class of problems or perform a computation. Algorithms are unambiguous specifications for performing calculation, data processing, automated reasoning, and other tasks [26]. The model in which every decision is based on the comparison of two numbers within constant time is called a decision tree model [25]. It was introduced to establish computational complexity of sorting and searching. The simplest illustration of this lower bound technique is for the problem of finding the smallest number among n numbers using only comparisons. In this case the decision tree model is a binary tree. Algorithms for this searching problem may result in n different outcomes (since any of the n given numbers may turn out to be the smallest one). It is known that the depth of a binary tree with n leaves is at least log n, which gives a lower bound of Omega (log n) for the searching problem [27].

1) Linear regression: In statistics, linear regression is an approach to modeling the relationship between a scalar response (or dependent quantity) and one or more (or independent variables). The case of one explanatory quantities is called simple linear regression [16]. For more than one explanatory quantities, the process is called multiple or more than one linear regression. This term is different from multivariate linear regression, where multiple correlated dependent variables are estimated [14], rather Than is in the given example a single scalar quantity [24]. Given a data of n units, a regression model forms a hypothesis that the relationship between the dependent variable y and the p-vector of regressors x is linear [6]. This relationship is modeled through a disturbance term or error variable e, an unobserved unobserved random variable that adds “noise” to the linear relationship between the dependent variable and regressors. Thus the model takes the form

\[ y = mx + c \]  (1)

In Eq. (1) y is the independent variable m is the slope of the equation x is the set of dependent variables c is the constant value if present. Value of mean and standard deviation for our dataset in linear regression is 0.4793 and 0.013 respectively.

2) Logistic regression: Logistic regression is a learning model that in its base form uses a logistic function to make a binary dependant quantity [11], although many more complex extensions exist. In regression analysis, logistic regression is determining the parameters of a logistic model (a form
TABLE I
DATASET SNIPPET WHEREIN EXT 1 TO EXT 10 EXCLUDING EXT 3 ARE DEPENDENT VARIABLES AND EXT 3 IS INDEPENDENT VARIABLE.

| EXT1 | EXT2 | EXT3 | EXT4 | EXT5 | EXT6 | EXT7 | EXT8 | EXT9 | EXT10 |
|------|------|------|------|------|------|------|------|------|------|
| party hard | (don’t talk) | (people friendly) | (humble) | (talkative) | (hesitate to speak) | (extravert) | (shy) | (ego-centric) | (conservative) |
| 4 | 1 | 5 | 2 | 5 | 1 | 5 | 2 | 4 | 1 |
| 3 | 5 | 3 | 4 | 3 | 3 | 2 | 5 | 1 | 5 |
| 2 | 3 | 4 | 4 | 3 | 2 | 1 | 3 | 2 | 5 |
| 3 | 2 | 2 | 3 | 4 | 2 | 2 | 4 | 1 | 4 |
| 3 | 3 | 3 | 3 | 5 | 3 | 3 | 5 | 3 | 4 |
| 3 | 3 | 4 | 2 | 4 | 2 | 2 | 3 | 4 | 3 |
| 4 | 3 | 4 | 3 | 3 | 3 | 5 | 3 | 4 | 3 |
| 3 | 1 | 5 | 2 | 5 | 2 | 5 | 2 | 3 | 2 |
| 2 | 2 | 3 | 3 | 4 | 2 | 2 | 2 | 4 | 4 |
| 1 | 5 | 3 | 5 | 2 | 3 | 2 | 4 | 5 | 4 |
| 3 | 3 | 2 | 3 | 3 | 2 | 4 | 3 | 3 | 5 |
| 3 | 1 | 5 | 3 | 5 | 1 | 5 | 5 | 5 | 3 |
| 4 | 1 | 5 | 4 | 5 | 1 | 4 | 1 | 5 | 2 |

III. RESULTS
The results are pretty much clear that as compared to other conventionally used algorithms like support vector machines and decision trees with depth 4 our algorithm performs better in the area of variance and mean absolute error however the difference is not very much when it comes to measuring the execution time for large datasets.

TABLE II
COMPARISON OF FIVE ALGORITHMS

| ALGORITHM        | MEAN | STANDARD DEVIATION |
|------------------|------|--------------------|
| Linear Regression| 0.4793 | 0.013 |
| Logistic Regression | 0.4793 | 0.013 |
| Linear Discriminant Analysis | 0.4802 | 0.020 |
| Naive Bayes      | 0.4740 | 0.018 |
| Support Vector Machine | 0.4869 | 0.018 |

IV. CONCLUSION
Though there has been research in the field of personality correlation but this is the first end to end implementation of compatibility detection software. Support Vector machine algorithm is best fit for the dataset that has been used. For further improvisations modified version of SVM can be used.

REFERENCES
[1] Tett, Robert P., and Patrick J. Murphy. "Personality and situations in co-worker preference: Similarity and complementarity in worker compatibility." Journal of Business and Psychology 17.2 (2002): 223-243.
