



COMPARATIVE ANALYSIS OF HARRASSMENT PREDICTION BY USING MACHINE LEARNING ALGORITHMS

Lakshmi Priya¹ and Nancy Jasmine Golden²

¹PG Student, Department of Computer Applications and Research Centre, Sarah Tucker College (Autonomous), Tirunelveli, Tamil Nadu, India.

²Associate Professor, Department of Computer Applications & Research Centre, Sarah Tucker College (Autonomous), Tirunelveli, Tamil Nadu, India.

Abstract:

Sexual harassment has become a stumbling block which prevents both men and women especially women from moving up and social structure effectively. Public attention has been drawn due to the issues of workplace or surrounding's sexual harassments. This is widespread and extensive research. It may also affect morale, family relationships and health also. In other words, it can be explained as an unwanted physical contact, invading one's physical space, offensive gestures etc., Sexual behaviour and sexual learning are ideal for analysis from a Behaviour systems perspective. Such an approach is encouraged because sexual behaviour consists of a number of different types of responses, each of which is controlled by a different set or category of stimuli. Here the people who are affected by these factors need some counselling and awareness programs to overwhelm it. This prediction can be applicable in health care centre, psychological counselling etc., The model can predict the rate of human sexual behaviour by taking a survey based on many scales. The Machine Learning (ML) algorithms such as Naive Bayes (NB) and K Nearest Neighbour (KNN) are used to predict who affected more either men or women on our society. For this project, Kaggle's multidimensional self-sexual conceptual questionnaires dataset is used. The values range from low acceptability to high acceptability. In this work two algorithms are compared with the help of performance is evaluated to calculate Accuracy.

Keywords: Behaviour Prediction, Harassment, Harassment Research, KNN, Machine Learning, Naive Bayes, Prediction model, Prediction Evaluation.

Introduction:

Humans always strive to understand what other people will do. Human behaviour describes how people behave and interact. It is the result of interconnected factors. A complex interaction of actions, intellect, and emotions determines human behaviour. Behaviour can help someone better understand why and how people behave, as well as how to manage people within an organisation. Body language, speech, facial gestures, emotional reactions, and thinking are a few aspects of behaviour. For many teenagers, starting sexual activity marks a significant transitional point into adulthood. People need to be made aware of both good and unpleasant touches. Bad touches are ones that are abusive and involve touching on private body areas, which is a criminal. Good touches are those that are not abusive. People should put an end to these behaviours by avoiding those who act aggressively and by studying self-defence. The outline of the work is shown in **Figure 1**.

Outline of the Work:

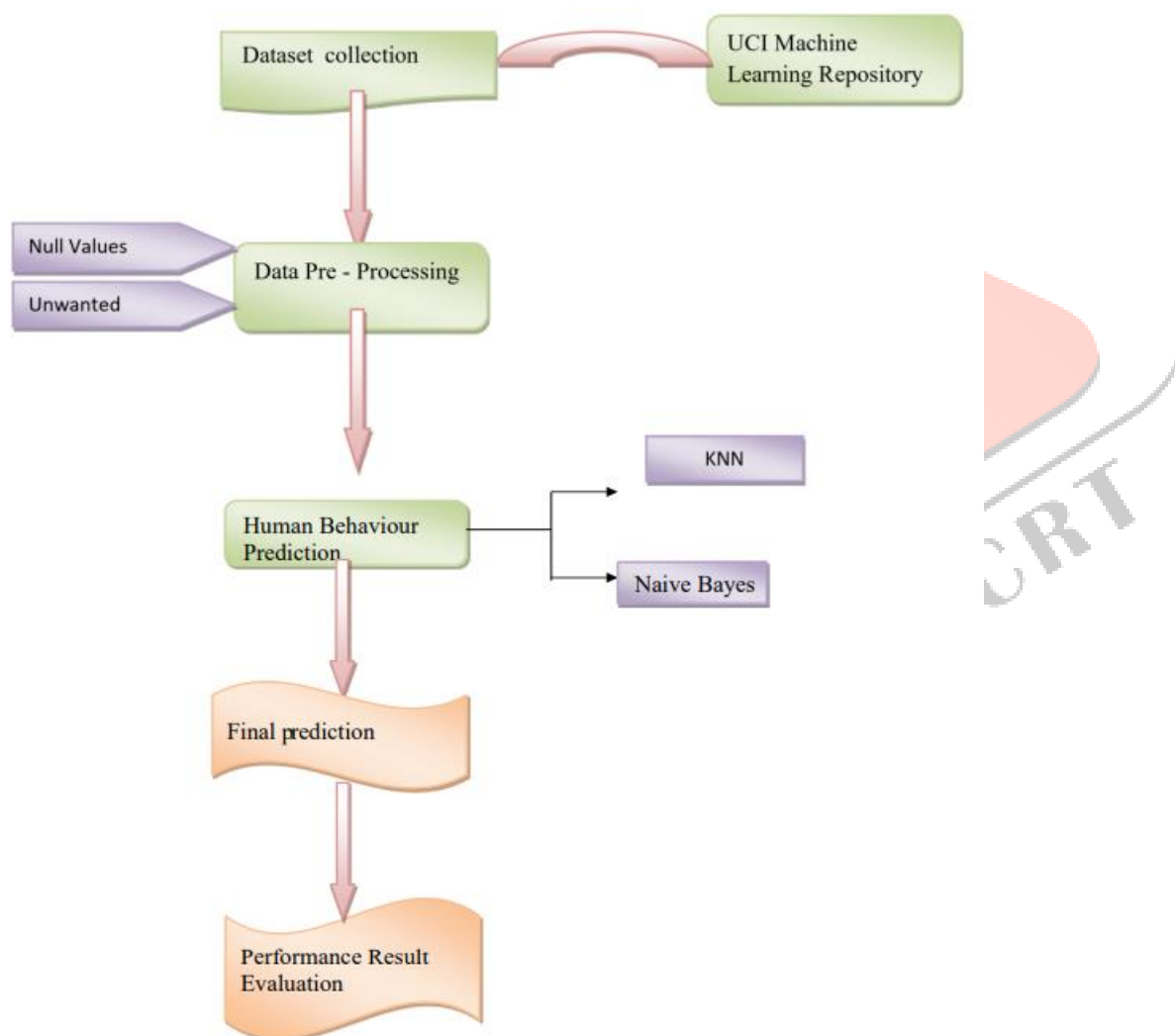


Figure 1: Outline of the Work

Dataset Description:

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. Here, in this research we are using recent dataset and is collected from the Kaggle machine

learning repository. It consists of several analysing variables and one target variable which is dependent to that.

The main objective of the research is to predict that who is affecting by the factors either male or female.

Pre-processing:

Data pre-processing is the process of transforming raw data into something that can be used by a machine learning model. In order to build a machine learning model, it is the first and most important stage. The obtained data may contain inconsistent data due to missing values. Pre-processing must be carried out to improve the model's efficiency in order to obtain the desired results on the final proposed systems. Additionally, it is the fundamental and important component of any classifications or predictions model. The pre-handling of data is also another name for this procedure. Both the method and interpretation are made simpler. To increase the precision of the prediction score, several techniques are used to remove duplicates or inconsistencies in the data.

Feature Selection using Pearson correlation:

The strength of the correlations between the various variables is evaluated by the Pearson correlation, often known as the Pearson R statistical test. As a result, it is always a good idea for the person conducting the analysis to calculate the correlation coefficient value in order to determine how strong the association between the two variables is. The link between the variables is positively correlated and both values decline or increase together if the value is in the positive range. The relationship between the variables is said to be negatively correlated if the value is in the negative range, in which case both values will move in the opposite direction.

Methodology:

Even if there are various approaches, models, and studies available, it can be challenging to determine which data mining method, classifier, or system is ideal for obtaining the desired results. Every system depends on a unique attribute dataset; hence it is crucial to create a solid model that can be adjusted to each attribute dataset. They also lack effective methods for data processing and variable selection. As a result, it is strongly recommended to examine categorization approaches before using them to analyse applicant data. Additionally, classifiers aid in predicting exact outcomes so that outcomes can be obtained.

Naïve Bayes Classification:

The Bayes theorem is the foundation of the probabilistic machine learning method known as Naive Bayes, which is utilised for a variety of classification problems. Naive Bayes differs from Bayes theorem in that it assumes conditional independence, whereas Bayes theorem does not. This indicates that there is no connection between any of the input features.

KNN:

One of the simplest machine learning algorithms, based on the supervised learning method, is K-Nearest Neighbour. The KNN method considers that the new case/data and the existing cases are comparable, and it places the new case in the category that is most similar to the existing categories. The KNN algorithm maintains all the information that is accessible and categorises new data points based on similarity. This means that as fresh data is generated, it may be quickly and accurately categorised using the KNN method. Although the KNN technique is most frequently employed for classification problems, it can also be utilised for regression.

Performance Evaluation:

The performance evaluation metrics are used to calculate the performance of your trained machine learning models. This helps in finding how better your machine learning model can perform on a dataset that it has never seen before. Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

S.No	Classifier	Accuracy
1	Naïve Bayes	69%
2	KNN	76%

Conclusion:

This comparative study easily predicts the human behaviours, which makes the people an easy and an understanding way on behavioural prediction because people should get some fundamental sexual education and information on a regular basis. We should examine people feelings attitudes and behaviours with any relationship. To build uniformity and scale some organization also provides some prevention and response to such behaviours. The aim of this study is to analyse, two machine learning classification methods were implemented, and their results were compared with different statistical measures. The experimental result shows that the accuracy of KNN of our dataset is 76% which is the highest among the rest. KNN is also giving highest accuracy for this dataset. Among two different machines learning algorithms applied, all the models produced good results for some parameter like precision, recall sensitivity etc.

Future Enhancement:

In future, this research can be implemented on other domain such as deep learning, artificial intelligence and so on. So many #metoo cares are going on in cyber-crime, so that dataset can be collected and used for further analysis. Sexual self-concept based multi-dimensional queries have been taken in this research work, in future real time images, videos, sexual behaviours can be predicted by using another different dataset. As per now, two algorithms are implemented in future deep learning algorithms can also be implemented.

References:

1. https://www.who.int/violence_injury_prevention/violence/global_campaign/en/chap6.pdf
2. <https://www.who.int/en/news-room/fact-sheets/detail/violenceagainstwomen>
3. Moreno MA (2015) Sexual assault and rape among college students. *JAMA Pediatr* 169(12):1184. <https://doi.org/10.1001/jamapediatrics.2014.2152>
4. Rosellini AJ, Monahan J, Street AE, Petukhova MV, Sampson NA, Benedek DM ... Kessler RC (2017) Predicting sexual assault perpetration in the US Army using administrative data. *Am J Prevent Med* 53(5):661–669
5. https://www.who.int/violence_injury_prevention/violence/world_report/factsheets/en/sexualviolencefacts.pdf?ua=1
6. Scott KM, Koenen KC, King A, Petukhova MV, Alonso J, Bromet EJ ... Kessler RC (2018) Post-traumatic stress disorder associated with sexual assault among women in the WHO World Mental Health Surveys. *Psychol Med* 48(1):155
7. Loughnan S, Pina A, Vasquez EA, Puvia E (2013) Sexual objectification increases rape victim blame and decreases perceived suffering. *Psychol Women Q* 37(4):455–461

