Artificial Intelligence Project

1st Abduallah El Maraghy
Faculty of computer science
Misr International University
Cairo, Egypt
Abduallah1808756@miuegypt.edu.eg

2nd Ahmed Osama Hussien
Faculty of computer science
Misr International University
Cairo, Egypt
Ahmed1815007@miuegypt.edu.eg

3rd Seif Nagi Seif
Faculty of computer science
Misr International University
Cairo, Egypt
Seif1805522@miuegypt.edu.eg

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Our problem statement is focusing on how to take any decision based on a scientific rule. That people always have some attributes (Data)and can't get any result based on them. So they start in thinking about having them in tables and classify them and based on that classification and mathematical operations they can reach the goal of knowing if this thing is applicable or not based on the result got by yes or no. Also, they designed many approaches to be able to use them with different types of data.

So based on the rules and operations that are done over the data we became able to reach answers from any data by applying the K-mean, Decision tree, and KNN Algorithms. So that helped as to reach to right efficient answers after applying those algorithms. another factor that can affect the answer is the dataset which plays an important role as according to the dataset we always need a big dataset to reach to high accuracy answers.

In this paper we used the weka to run differ algorithms on the differ approach of the supervised and unsupervised.Supervised using two algorithms KNN and Decision tree that focus on having a classified data , while in the unsupervised using the k-mean algorithm focus on have no classification but on clustering them in which classify the data according to random choosing of the centroids using differ ways as ecludien and manhatein .from the Decision Algorithm we can see the main decisions and the leaf nodes that the action that will be taken according to them by yes or no.while the KNN algorithm focus in applying the decisions on comparing the given case on the previous dataset to get out the result which belong to the truth.While on the other hand the the k-mean is focusing on having many differ data and start in classifying them into clusters to be able to have a correct decision.

II. Methodology

In the real life, we are surrounded by humans who can learn from their experiences using their learning ability, along with computers or robots that operate on our orders. But can a machine also learn from experiences or past data like a human does? Can a machine act like a human and learn from past experiences or data? So here comes Machine Learning's role. To perform such a process we have to find a good quality dataset to be the data that the machine will use as a memory reference to predict the incoming event when the same actions occurs again. The process is performed by training the machine on many inputs that have the same output to reach a high percentage of accuracy when testing on a real event.

Some algorithms of supervised and unsupervised learning techniques is used to reach this approach like : KNN, Decision Tree, K-mean. For every algorithm there is two processes; first process is to training the machine to find a set of weights and biases that have low loss, on average, across all examples extracted from the dataset, and the second is to test the machine to evaluate the performance reached by training it.

To start our process of implementing this project we have to pick up a good quality dataset. we have explored "Kaggle" to find such a good quality dataset with a usability rate of 9.4 that we are going to use for testing and training.

We have searched for a machine learning data-sets for classification which have a numeric values and have a high usability. To insure that this dataset is valid we have used weka tool to test our algorithms on it and save the results to test it later after implementing the project using python scripts.

III. Related Work

Many researchers have earlier tried to create strong immunity against spam email. Mostly the common approaches are Naive Bayes, Support Vector Machine(SVM), and Artificial Neural Network(ANN) to detect spam email. we've often seen researchers don't apply proper data preprocessing; as a result model performance doesn't reach up to the mark.

Mohamad. used 'content-based spam filtering' and 'non-content-based spam filtering' both to classify the spam and regular emails. they need to apply the Porter-Stemmer algorithm for data cleaning purposes and as extracted features, they need to use a hybrid feature, which is that the combination of Term Frequency Inverse Document Frequency(TF-IDF) and therefore the rough pure mathematics. Throughout their experiment, their spam email detector has detected the spam emails accurately but it predicted around 50% of normal emails as spam email, which is extremely much inconvenient.

Renuka. have mainly focused on data pre-processing. they need to use a stemmer to convert a word to a related form and to correct the missed word also. For classification purposes, they need to use the Naive Bayes algorithm and have presented a statistical comparison between the classification model performance using the Content-based spam filtering technique and Word Stemming-based spam filtering. Their proper data pre-processing technique has helped the model to succeed in 96% of the general classification accuracy. Singh. has suggested applying the Intelligent Water Drops Algorithm for spam email classification. He has applied a singular feature selection process and got around 93% accuracy. He has experimented on these pambase datasets, which contain only the TF-IDF of the words of the messages. It doesn't contain any single raw message but within the world, we'd like a classification model, which may classify the raw emails properly.

Harisinghaney. have used various machine learning algorithms like K- Nearest Neighbor (KNN), Naive Bayes, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithms to differentiate the spam emails from regular emails. they need to use a little part of 'Enron corpus datasets' and took 2500 emails for training and 2500 emails for testing. they need also to suggest applying the OCR algorithm to detect the texts from the attachments. they need to show around a forty-five increased rate of Classification model performance thanks to proper data pre-processing. They got around 87Awad. have suggested employing a neural network. they need also to use the spam base dataset and applied different types of neural networks like Multi LayerPerceptron(MLP) and Artificial Neural Network(ANN). they need applied data preprocessing and have selection alright Journal of Huazhong University of Science and Technology ISSN-1671-4512 and got the very best overall accuracy around 93% using MLP. Despite employing a proper classification algorithm, feature selection plays a crucial role in text classification. feature selection aims to calculate the relevant features to enhance the classification rate and reduce the time interval. Irrelevant features often cause a coffee classification accuracy rate and high computational cost. Thota et al. and Zhu et al. have applied feature selection but haven't mentioned the methodology to extract the important features.

Liang. have performed a feature ranking algorithm for feature selection but found a really low accuracy rate. Sasaki et al. have used a really unique method to detect spam emails. He has applied the text-based clustering method, which provides a satisfactory around 90% overall accuracy score.

A. *Some of a related projects:*

- Sentiment Analyis
- Spam Email Classification Using Supervised Machine Learning
- INTERNET ADVERTISING: RELIABILITY, DILEMMAS, AND POSSIBLE DIRECTIONS