USE FOR GRAPHICAL USER TOOLS IN DATA ANALYTICS AND MACHINE LEARNING APPLICATION

Dr. Kamal Gulati¹, V.P. Sriram², Dr. Mukta Sharma³, Parul⁴, Sherin Eliyas⁵, Sunil L. Bangare⁶

¹Associate Professor, Amity University, Noida, Uttar Pradesh, India | Independent Consultant, Stratford University, Virginia, USA, drkamalgulati@gmail.com
²Associate Professor, Department of MBA, Acharya Bangalore B School (ABBS), Bengaluru, Karnataka, India, dr.vpsriram@gmail.com
³Associate Professor, CS & IT Department, Trinity Institute of Professional Studies, Dwarka, Delhi (Affiliated to GGSIP University), India
⁴Research Scholar, Department of Computer Science, Baba Mastnath University, Rohtak, Haryana, India, dparul79@gmail.com
⁵Assistant Professor, Department of Computer Applications, Hindustan Institute of Technology and Science, Chennai, Tamil Nadu, India, sherine@hindustanuniv.ac.in
⁶Assistant Professor, Department of I.T., Sinhgad Academy of Engineering, Pune, India, sunil.bangare@gmail.com

ABSTRACT

The study was aimed to analyze, clean, visualize, standardize and train traditional machine learning models all in a single GUI tool which is software-based. The data preprocessing includes the deletion of unwanted columns, handling missing values and dummy variables. The data analysis is supported further by data visualization. Different types of graphs can be plotted to analyze the variations in data that one cannot figure out by looking at the data set. On the other hand, this tool can also build and train machine learning models of the user's choice using the dataset. The user has the option to standardize the data before training the model. Post training, the user can get the complete Summary of the model including the evaluation metrics like accuracy, r-squared value, f-statistic, z-statistic etc. This report can also be saved in the computer. The user can also save the trained model to use it later or somewhere else. In this tool, the user can also compare the accuracies of different models trained on the same data set. This tool provides one complete solution for Data Analysts and Machine Learning Engineers as the tools used by them are all available in the tool. This tool also helps the people with less exposure to coding as it does not require a single code to be written. The interface is completely user friendly and simple.

Keywords—Data Analysis, Machine learning, Data preprocessing, Standardization, Data visualization.

I. INTRODUCTION

Machine learning is all around us, from next word suggestions while typing to movie recommendations in OTT platforms. The world generates enormous amount of data each day and is also completely dependent on this data, here are some daily statistics- 500 million tweets are sent, 294 billion messages are sent, 4 petabytes of information are made on Facebook, 4 terabytes of information are made from each associated vehicle, 65 billion messages are sent on WhatsApp, 5 billion inquiries are made, By 2025, it's assessed that 463 Exabyte of information will be made every day worldwide – that is what might be compared to 212,765,957 DVDs each day! And most of the data is used by different machine learning models implemented in several applications and websites. The data extracted is also known as raw data, which may contain some unwanted information as well. This tabular data is then processed where it is cleaned. The data is visualized to delineate any pattern that cannot be interpreted from the data set.
This data set is then used to train the model according to the problem and this model is used to predict target outputs of unknown data. The data preprocessing, profiling, visualization and standardization all fall under the data analysis part. The model training, model evaluation, and prediction all fall under the Machine learning part. This tool allows us to work on both fields simultaneously. Thus, it minimizes the effort and provides a single solution for different fields. Also, coding is not required to use this software which makes it more efficient.

The data visualization part allows the user to plot different types of graphs, correlation maps, and distribution graphs. This helps the user to understand the data more clearly, even if the user does not require to use the model training feature, they can use the tool to represent the data in different forms and decision-makers can take decisions more carefully. They are also allowed to save the graphs in picture format and use it somewhere else. Considering the fact that several similar tools are available over the internet, what makes this tool different is that it is software based while those are cloud based. For using this tool, the user will not require any kind of internet connection or any account. Making it easy to use, the user is only required to browse the data set from the window which pops out after pressing the import button.

II. BACKGROUND

Data is the basic element upon which any organization runs and flourishes. It is impossible to visualize the world without data [1]. It begins with a query, and data is only a response to that query. Superfluous to the kind of information, analysts investigate, their central goal, and crowds' vision direct them to discover the pattern to shape the story they need to tell. In this emerging critique of ‘algorithms’ carried out by scholars in law and in the social sciences, few have considered in much depth their mathematical design [2].

Although the analysis methods, modeling assumptions and restrictions are totally transparent to the user, this transparency is not achieved at the expense of analysis power [3]. For data exploration, the user needs to arrange these findings, prepared and introduced in an offered setting, to make them helpful. The fundamentals of data analytics include recovering and assembling huge volumes of information, coordinating it, and transforming it into bits of knowledge organizations can use to settle on better choices [4]. This includes figuring out enormous measures of unstructured data and getting key experiences from it.

The brief review by Berral-García, J. L. explains the basics of machine learning algorithms, and enumerates some (but not all) of the most used algorithms for analyzing and modeling data [5]. It's hard for users to visualize, explore, and use this enormous data. The ability to visualize data is very important for scientific research. At present, computers can be used to process a large amount of data [6].

The essential thought behind ML techniques is that an algorithm for a computer is prepared to "learn" the conduct introduced as a component of past experience. Machine learning methods, on the other side, are closer to the real problems but they usually rely on problem- or user-specific parameters or thresholds making it rather art than science [7]. There are a few open-source executions of AI algorithms that can be utilized with either application programming interface (API) calls or no programmatic applications. Instances of such executions incorporate Azure ML, Power BI, and Amazon Sage maker etc. [8]

III. RELATED WORK

With the rapid boom in data science and AI advancements, tech associations have redirected their thoughts on technologies that ordinarily forestall the programming angle and offer easy to understand GUI (Graphical User Interface) so anybody with insignificant information on algorithms can utilize them to analyze data and fabricate top-notch AI models. Various organizations have actually dispatched GUI-driven tools for data analytics and machine learning which perform several tasks like data analysis by which given raw data is changed into helpful measurements, bits of knowledge, and explanations to settle on significant business choices.

1 Azure ML: Azure Machine Learning Studio is a GUI-based tool which provides comprehensive facilities to its users to build and operationalize Machine Learning tasks. It provides a visual intuitive climate with no prerequisites for coding. It follows a basic five-venture measure:

a) Import dataset of any format

b) Perform data preprocessing steps like data cleaning and wrangling.

www.turkjphysiotherrehabil.org
c) Splitting the dataset into a training set and testing set.

d) Train model using different ML algorithms

e) Evaluate the model and get predictions

2 **Amazon Sage maker:** Amazon Sage Maker is a completely oversee administration that incorporates modules that can be utilized together or individually to effectively develop, train, and deploy ML models. It is generally adopted by particular organizations that empowers data scientists and developers to build machine learning models. The product puts forth a brave attempt at making data analytics and Machine learning easier to understand. Sage Maker delivers customers who wish to utilize AI for predictive analysis, data mining and meeting market expectations.

3 **IBM Watson:** IBM Watson Studio is an AI-based application utilized for Data Processing, Machine Learning, Data analysis, etc. IBM Watson Machine Learning supports data engineers and developers to deploy AI models across any cloud platform. It offers help to perform several tasks related to data preprocessing, cleaning, visualization and model building alongside various data science dialects and analytics tools. IBM helps users to train models dynamically, to deploy them on cloud, and create APIs to construct AI-controlled applications through DevOps.

4 **Tableau:** Tableau is the most well-known data analytics tool for visualization purposes. It permits you to separate insignificant, unformatted information into useful and understandable form. Visual i.e. graphs and charts made by utilizing Tableau can undoubtedly assist you with understanding the conditions between the indicator factors. In spite of the fact that Tableau is essentially utilized for data visualization, it can further perform data analysis as per user's requirements.

5 **Power BI:** Power BI is Microsoft's commercial software for data visualization and analysis for business intelligence (BI). It tends to be used to extract data from a wide range of sources in the cloud and create interactive dashboards. Power BI permits clients to make intuitive visual reports, with an insignificant expectation to learn and adapt. Its fundamental selling point is its incredible information availability; it works flawlessly with Excel yet additionally SQL server, MySQL databases, text files, and cloud sources.

IV. METHODOLOGY

To take care of an issue, a designer needs an approach to sort out the work, execute it without losing time which assists with explaining the objective of the client. When the business issue has been plainly expressed, the analysts can characterize the scientific way to deal with the issue. This progression involves communicating the issue with regards to statistical and AI strategies which is fundamental since it recognizes what sort of patterns will be expected to address the inquiry most viably. Some of the methods involved in designing this tool are:

A. Data Analysis

At the point when given a machine learning issue, a data researcher is regularly entrusted with making illustrative factors, also called highlights that are prescient of the result of interest. Fruitful Data analysis requires the formation of highlights that not just give helpful bits of knowledge into the actual information, yet additionally considers any restrictions of the learning algorithms that is being utilized. Note that this is definitely not a minor task, as the exhibition of a given machine learning algorithmic hyper parameter is vigorously reliant upon the nature of the data input highlights.
B. Data preprocessing

The collected data naturally have noise, but the noise can affect the training of the model negatively. Therefore, the process of data cleaning must be carried out if necessary. Dataset will be provided by the user that will consist of files where columns will be Attributes and targets (whichever user wants to select). Users can navigate through the entire dataset and can delete or update columns and tuples. With the help of a dropdown menu, users can select what to do with selected items.

C. Data Visualization

Data visualization provides an effective data representation for the data collected from various sources. It is associated with the design, development, and implementation of the graphical presentation of the data. It also permits the decision makers to look at the analytics in pictorial format and makes it easier to understand the data. This helps to locate different patterns, understand statistics, and set up an opinion.

D. Data standardization

In the last step of feature engineering, data standardization comes into view. When the characteristics of the input data sets have huge differences in-between their range, or we can say that whenever they are measured in dissimilar measurement units like Pounds, Meters, and Miles etc. The differences in the range of these prime features can cause problems to many machine learning models. For example, for the models that support distance computation, if one among the features has a wide range of values, the distance will be governed by this particular feature. In simple terms, Standardization is a second scaling approach where the values are concentrated on the mean with a unit standard deviation. This means that the mean of the attribute will become zero and the resultant distribution will have a unit standard deviation.

Z-score is one of the most popular methods to standardize the data, and can be done by:

$$Z = \frac{\text{value} - \text{mean of sample}}{\text{Standard Deviation}}$$
At the end, the tool provides several outputs as the result, these outputs depend on the choice of user.

- The first output is the cleaned or preprocessed data set. Users are allowed to save the dataset in .csv or .xls format. If he wants he can just press the save dataset button, from the pop-up window he can browse to the location where he wants to save the dataset. This feature is for those users who only want to do the preprocessing work and analyze the data.

- The second output is the graphs which are plotted in the data visualization part. The user is allowed to save the graph in the computer as an image. The save graph button will help the user to do so. Again, the user will encounter a pop-up window to browse to the location where he wants to save the image. This feature is for the data analysts or decision-makers who take crucial decisions by interpreting the graphs and statistics. There are more than 10 types including 'Scatter Plot', 'Box Plot', 'Bar Plot', 'Histogram Plot', 'Distribution Plot', 'Correlation Heatmap', 'Hex Plot', 'Violin Plot', 'Strip Plot', 'Swarm Plot', 'Linear Model Plot', etc.

- The third output is the Summary of the model trained using the dataset. This Summary provides a brief detail of the evaluation metrics. This includes the values of accuracy, r-squared coefficient, adjusted r-squared coefficient, p-value, etc. These values are described below:

  - **Accuracy** – The accuracy is defined as the percentage of correct predictions made by the model on the test data.

  - **R-squared coefficient** – It can be defined as the ratio of variance in the outputs that our model is capable of predicting based on its features.

  - **Adjusted r-squared** – It is the adjusted r-squared for the number of predictors in the model. It increases only if the new term improves the model more than would be expected by chance. It is always less than the r-squared coefficient.

  - **P-value** – It is the probability that any statistical measure of assumed probability distribution will be greater or less or equal to the observed result.
### VI. CONCLUSION

This research has been conducted to provide wholesome knowledge in regards to a GUI based data analytics and machine learning software. This is an interactive GUI based software for performing data analysis and exploration along with model building using traditional machine learning algorithms. It accepts dataset of various formats as input and furnishes an outline with different data profiling and statistical summary reports. Different data transformation can be performed on the dataset utilizing some pre-characterized functions. Users can investigate their information with synopsis insights, and save the Dataset into their computer. Datasets are

---

<table>
<thead>
<tr>
<th>Dep. Variable:</th>
<th>Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model:</td>
<td>OLS</td>
</tr>
<tr>
<td>Method:</td>
<td>Least Squares</td>
</tr>
<tr>
<td>Date:</td>
<td>Thu, 13 May 2021</td>
</tr>
<tr>
<td>Time:</td>
<td>21:53:02</td>
</tr>
<tr>
<td>No. Observations:</td>
<td>50</td>
</tr>
<tr>
<td>Df Residuals:</td>
<td>46</td>
</tr>
<tr>
<td>Df Model:</td>
<td>3</td>
</tr>
<tr>
<td>Covariance Type:</td>
<td>nonrobust</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>coef</th>
<th>std err</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1918</td>
<td>0.024</td>
</tr>
</tbody>
</table>

---

| R-squared: | 0.951 |
| Adj. R-squared: | 0.948 |
| F-statistic: | 296.0 |
| Prob (F-statistic): | 4.53e-30 |
| Log-Likelihood: | 76.973 |
| AIC: | -149.9 |
| BIC: | -142.3 |

---

| t     | P>|t|   | [0.025 | 0.9 |
|-------|------|--------|--------|
| 7.876 | 0.000 | 0.143 | 0.9 |

---

| Omnibus: | 14.838 |
| Prob (Omnibus): | 0.001 |
| Skew: | -0.949 |
| Kurtosis: | 5.586 |

---

| Durbin-Watson: | 1.282 |
| Jarque-Bera (JB): | 21.442 |
| Prob (JB): | 2.21e-05 |
| Cond. No. | 10.8 |

---

Figure-3.1 Summary of model (Output) -1

Figure-3.2 Summary of model (Output) -2

Figure-3.3 Summary of model (Output) -3

Figure-3.4 Summary of model (Output) -4
handily devoured by models during preparing. They are consequently retrained through algorithmic hyper parameters to track down the best model for providing predictive results.

REFERENCES

10. Kamal Gulati, S. Saravana Kumar, Raja Sarath Kumar Boddu, Ketan Sarvakar, Dilip Kumar Sharma, M.Z.M.