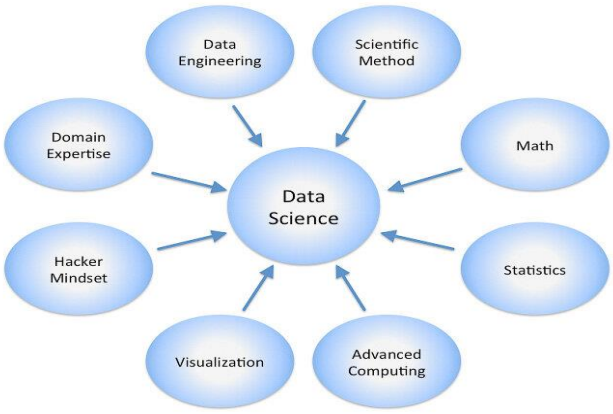
ADS Software Documentation

Contents

INTRODUCTION

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyse actual phenomena" with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge and information science. A Data Scientist is responsible for extracting, manipulating, pre-processing and generating predictions out of data using various tools and programming languages such as MS Excel, MS Access, Python, Power BI, R, Hadoop and many others.

This work seeks to delve into the learning curve involved in transforming data into models that are useful in decision making and propose an easier alternative for anyone at any level of the data science profession.



PROBLEM STATEMENT

Despite the increasing availability of data and a growing appreciation for the importance of Data Science and data-driven product design, there is a practical disconnect between the amount of work available and the number of people with the requisite skills. With data science still emerging as a career path for many people, a shortage of data science human resources is a key contributor to the disconnect. This is partly due to the tough learning curves that are required to get acquainted with the available tools and other factors. This gap

hinders the extent to which the potentials in Data Science is harnessed especially on the African continent.

Also, the few people with the requisite skills have to go through rigorous means to analyze their dataset in order to achieve their specified goals. This mostly results in spending more time on a single project which could have been used for other productive projects.

MOTIVATION

There are a lot of available software and programs to help an individual in the data science process. However, those software and programs normally requires that one has a technical skill like coding and understanding complex mathematical algorithms. In some cases, the user-friendly ones are more expensive. All these factors go a long way to limit the number of people who patronize the field of data science.

Also, the few available people who has the requisite skills have to go through complex processes in analyzing their dataset. This results in spending a lot of time on one project which could have been used for other equally important projects.

It is in this light, that our team decided to design a program that will solve all these problems. A program that will bring data science to everyone irrespective of their background as well as help automate the data science processes.

AIM OF THE APPLICATION

The primary aim of ADS studio is to enable individuals from all spheres of life to draw meaning out of their dataset without going through any complex processes. ADS Studio is built in a way that it uses mathematical algorithms to quickly perform exploratory Data analysis, data visualizations, data cleaning, preprocessing and creating of machine learning models. All the user has to do is to load the dataset and ADS studio handles the rest.

ADS Studio is also designed in a way that allows individuals already in the data science world to automate the data science process and get meaning out of their dataset in a very short period of time. Indeed, ADS studio is bringing data science to everyone.

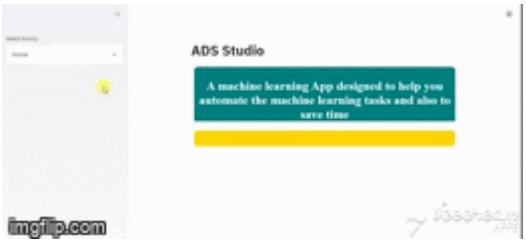
REQUIREMENTS ANALYSIS

PRODUCT PESPECTIVE

ADS Studio is a web-based system. It can be accessed using Internet Explorer 8.0 and above, Mozilla Firefox 2.0, and Google Chrome and the like.

USER INTERFACE

Users are able to view the home page of the ADS Studio, load dataset, perform data exploratory analysis, data cleaning, data visualization and build models



HARDWARE INTERFACE

The ADS Studio application shall provide minimum hardware requirements. The following hardware configurations are required for a PC using the ADS Studio:

- Pentium processor
- 32 MB of free hard-drive space
- 128 MB of RAM

SOFTWARE REQUIREMENTS

This section lists the requirements that are needed to run the system efficiently. The operating system needed for the system to run effectively, the interface to run the application, the driver for running web applications, the integrated development environment to develop the application, and the third-party tool used for editing purposes are as follows:

- Operating System: Windows or MAC OS
- Web Browser: Microsoft Edge, Internet Explorer, Mozilla Firefox, or Google Chrome.
- Integrated Development Environment: Visual Studio Code, Anaconda

PRODUCT FUNCTION

The ADS Studio application would have the following basic functions:

1. Enable a user to load their own dataset (CSV, xls, json)
2. Perform Exploratory Data Analysis based on the dataset
3. Perform Data Visualization (Pie Chart, Bar chart, pairplot, scatter, correlation matrix etc)
4. Perform Data Cleaning on the dataset
5. Preprocess the dataset for Model Building
6. Build preferred model (Regression, Classification, Clustering)
7. Use the model for inferencing

USER CHARACTERISTICS

The users of the ADS Studio are broadly categorized into two:

- Experts in the data science field who wishes to automate the data science process and save time.
- Non- data science experts who seeks to draw meaning out of their dataset.

The studio has been designed in a way that it is user friendly and it is just about clicking on the functionality you desire and then the ADS Studio generates your desired result for you. No need for coding or performing any complex mathematical operations.

CONSTRAINTS

- Hardware Limitations: The minimum hardware requirement for the system is 128MB of RAM and a 32 MB hard disc- drive
- Currently, the application is a web- based product and can only be accessed with internet connectivity

ASSUMPTIONS AND DEPENDENCIES

The assumptions and dependencies of ADS Studio are as follows:

- The system is dependent on the availability of python and Streamlit to run.

- We assume that system users adhere to the system's minimum software and hardware requirements.
- This system will use third-party software, and it is assumed that system users are familiar with the software (Microsoft Azure).

IMPLEMENTATION

This chapter includes the detailed design used to build the ADS Studio. The system's design is used to create the functions and operations of the gathered requirements in detail, including process diagrams, and other documentation.

DETAILED SCOPE

The ADS Studio has 5 sections, which are explained below:

1. The first section is the Home page which gives a welcome address to the users
2. EDA (Exploratory data Analysis): It contains subsections needed in performing basic exploratory data analysis.
    - show shape,
    - show columns
    - show summary
    - value counts which are needed in performing EDA.
    - Show null values
3. Data Cleaning section that enables users to clean their dataset
4. Visualization: This section enables users to visualize various types of graphs. It has subsections like:
    - bar graph
    - pie chart
    - Histogram
    - Pairplot
    - scatter plot
    - correlation matrix.
5. Machine Learning: This section enables users to create models using different machine learning algorithm. It has subsections like:
    - Regression
    - Classification
        - SVC
        - KNN
        - Decision Tree
        - Random Forest
        - Naïve Bayes
    - clustering.

ACTIVITY DIAGRAM

This section lists the activity diagram and describes the flow of activities in the ADS Studio. A detailed description is then given after the figure for each activity. The figure below demonstrates the activity flow for the ADS Studio. The flow begins when the user comes to the Home page. The user can browse through the available list of categories and then choose the one that suits his or her interest. When the user decides to select EDA, he or she will be allowed to input his or her preferred dataset (CSV/XLS/JSON) and will be presented with the options below:

- show shape,
- show columns
- show summary
- value counts which are needed in performing EDA.
- Show null values

Likewise, when the user selects Visualization, he or she will be allowed to visualize a number of plots based on what he or she chooses. Some of the plots that the user can visualize are listed below:

- bar graph
- pie chart
- Histogram
- Pairplot
- scatter plot
- correlation matrix.

However, a user can decide to go to the plot section directly without first going through the EDA section. Also, a dataset loaded in the EDA section remains in memory and can be used in the Plot section without necessarily reloading.

In a similar way, a user can decide to select the data cleaning section where the user tidy his or her dataset.

When a user selects Machine learning, ADS Studio automatically cleans the dataset, and perform label encoding on all the categorical columns. It then requests the user to select the target column and also select the feature columns. The user has the option to enter the test value and the random state. It then presents the user with the subsections below:

- Regression
- Classification
- Clustering

When the user selects Regression, he or she is presented automatically with the regression results. The user is then asked to select a number of visualizations based on the metrics generated. The Regression Analysis is supposed to be applied on continuous dataset.

However, when the user selects Classification the user is presented with the following models to select from:

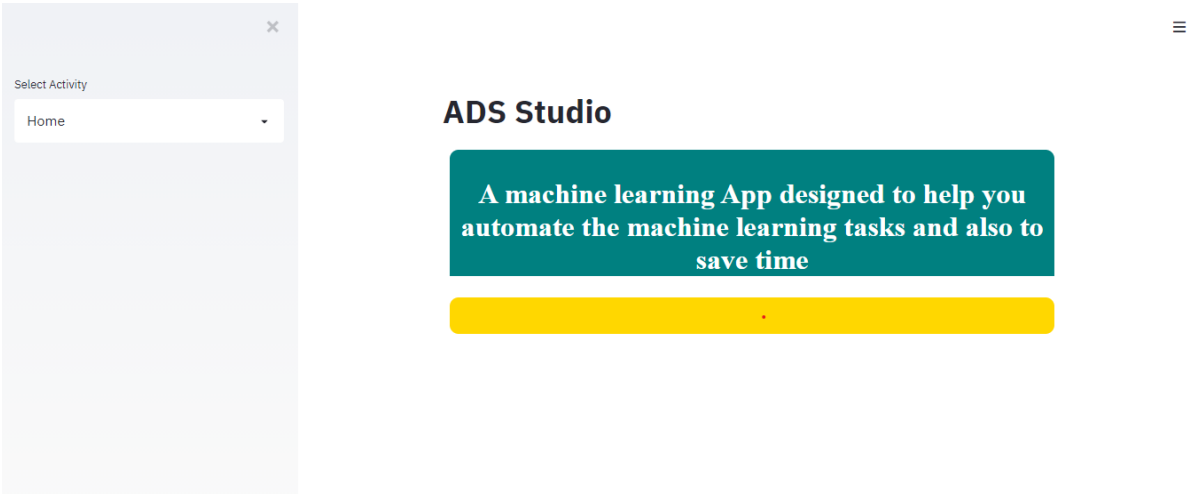
- SVC
- Decision Tree
- Random Forest
- KNN
- Naïve Bayes

When the user selects any of the models above, ADS automatically generates the classification results. With each model the user can decide to view the following metrics with their corresponding results.

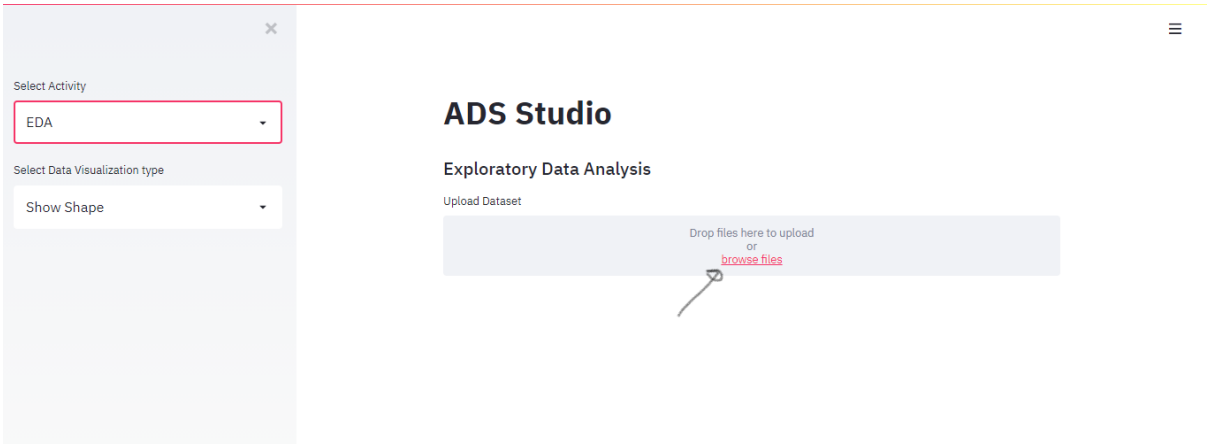
- Confusion Matrix
- ROC-Curve
- Precision- Recall.

EXAMPLE OF A USE CASE (Performing regression Analysis on the University Admissions Dataset)

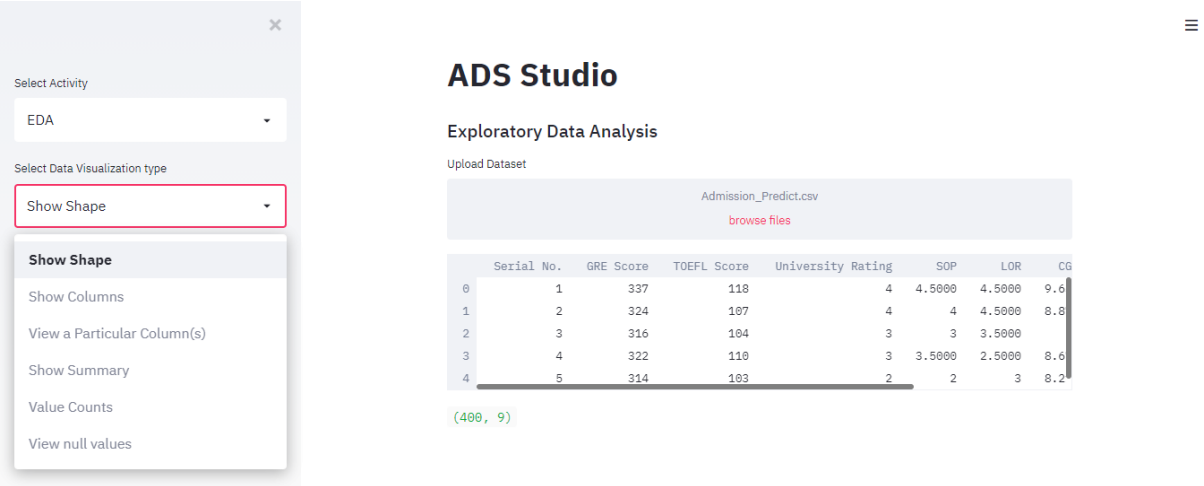
The user navigates to the web App and come to the ADS Studio Home page as displayed below.



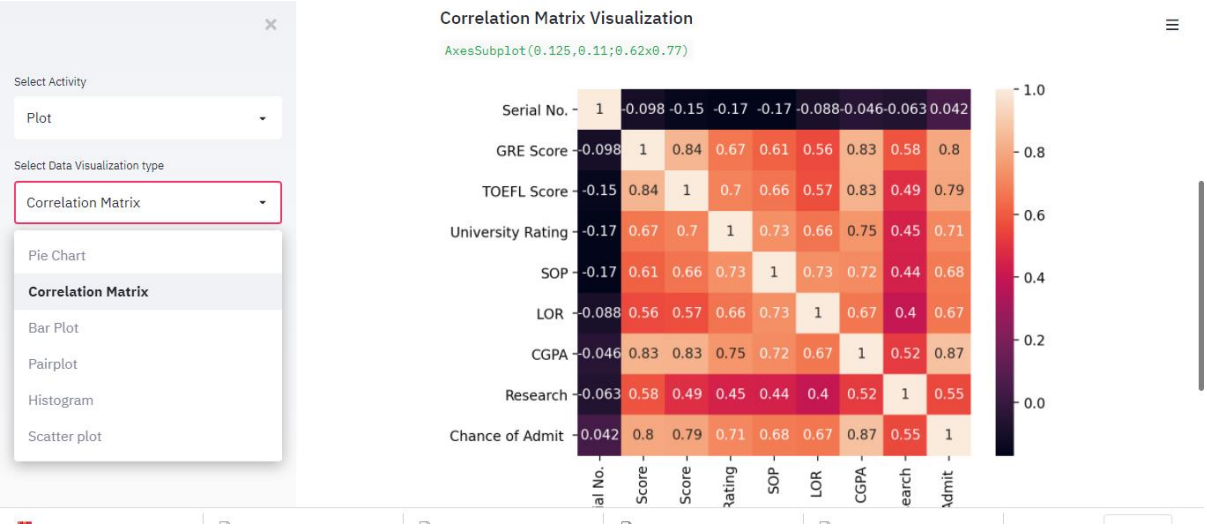
The user then uploads his or her dataset in order to begin the data science process



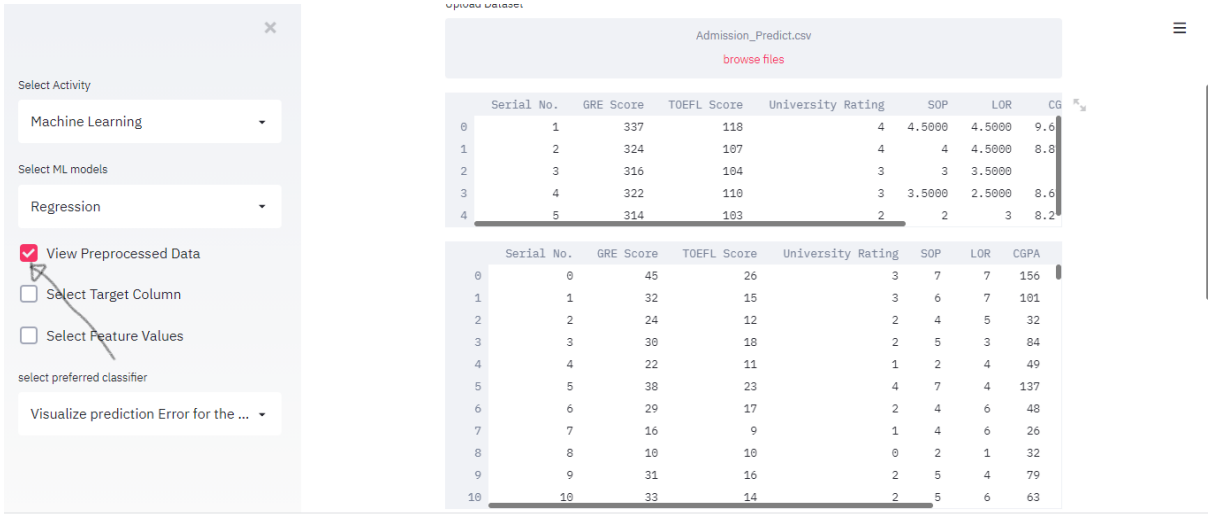
After importing the dataset to the system, the user can then perform exploratory data analysis on the dataset.



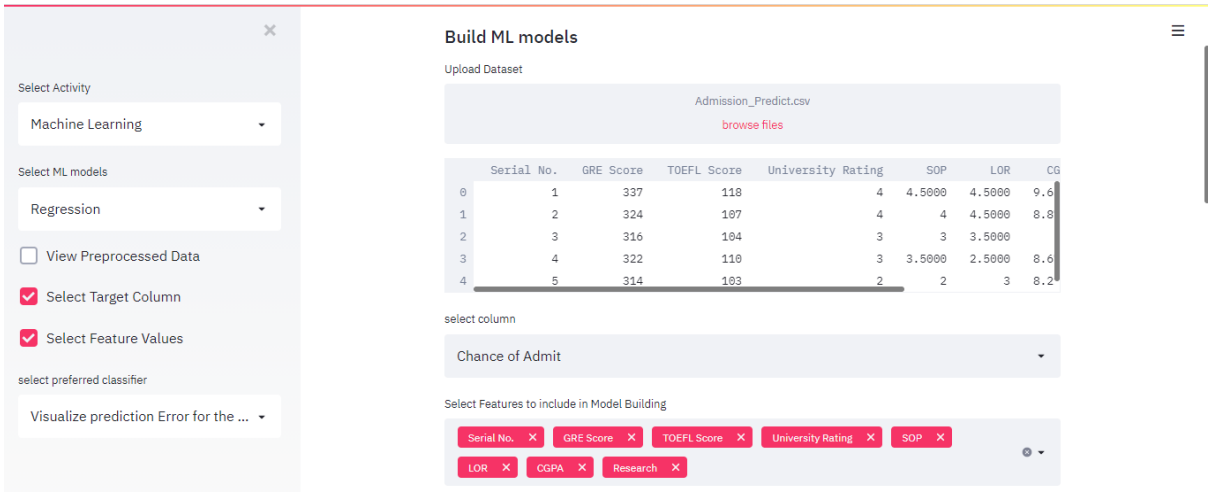
The user can then visualize the dataset using various graphs as displayed at the extreme lefthand side. In this case, the user decides to visualize the correlation matrix.



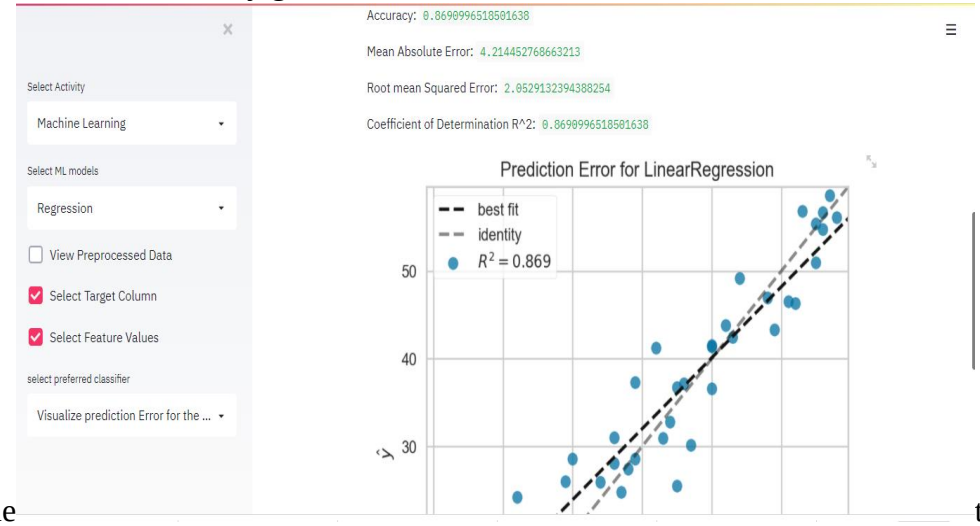
Immediately the user selects Machine learning, ADS automatically preprocessed the dataset by normalizing using MinMax, removing null rows and label encode the categorical dataset.



The user can then select the desired Machine learning building algorithm depending on the dataset. In this case, it is a regression problem so the user selects Regression. The system then requests the user to select the target column(dependent column) and also select the Feature columns(Independent columns).



ADS Studio then automatically generates the Prediction results and also shows the regression metrics results.



One can also visualize bar plot showing the errors in the actual and predicted values

CONCLUSION

This work aims to reduce the time an individual spends on a particular project and also to enable non-technical people to easily go through the data science process without any stress.

WAY FORWARD

We seek to add more features in the future and also to make the interface more user friendly.

ACKNOWLEDGEMENT

Azubi Africa

Microsoft4Afrika

Samuel Manu

Richmond Chris-Koka

Sharon Bosire

COLLABORATORS

Emmanuel Akpe

Emmanuel Osabutey

Alfred Ametepey