**Comparative Analysis of Data Mining (DM) Platforms:** **SAS Enterprise Miner vs. Python in reference to Decision Tree (DT) induction technique.**

[Enter Author details].

**Abstract:** This paper serves to present a comparative analysis of decision tree data mining techniques using the various software platform[[1]](#footnote-1). Only two platforms, the SAS Enterprise miner and Python, are used for study from various existing data mining platforms. Data mining is offers exemplary analytics avenue from which various companies can drive future insight. As such, understanding of appropriate software would provide accurate data sourcing to reduce forecasting errors.

A side to side evaluation of each software was made using four leading proponents; Functionality, Ancillary Tasks, Support Performance, and Usability are performed on both SAS Enterprise and Python before summing up to comparative analysis to provide needed classification and performance based on above parameters. Owing to various discrepancies realized by the two software in relation to tools of comparison, there is no generalization win of any software since both portrays difference performance which varies from one aspect to another. However, the python was found to exhibit better usability than SAS, then only criteria where two differed sharply. SAS exhibit better computation performance which is attributed by controlled development and vast algorithm use to enhance the accuracy

*Index Terms*—Data mining, Decision tree, Python, SAS Enterprise Miner, Induction techniques

# **INTRODUCTION**[[2]](#footnote-2)

Data Mining finds wide ranges application in data extraction, analysis to provide needed insights [1] . Employment of this term finds it way in large data volumes, a factor that has made it popular in blockchain cryptocurrencies, industrial data volume handling, and supporting upcoming analytics platforms such as Industrial Internet of things and Industry 4.0 in what is referred to as “Big data.” This field relies on complex computational analytics employing both machine learning algorithms and statistical methodologies to arrive at needed targets [1]. It is imperative to data-mine without the reliance on intelligent computer algorithms, especially in a complex system that needs higher accurate final models and output. However, more straightforward computational approaches may be applied to small-scale methods. Data mining find wide ranges of application in the modern world with the most common application being database marketing and Customer relationship management (CRM), risk management, fraud and intrusion mitigation, data filtering, research analysis, a criminal investigation, Financial and corporate surveillance, and bioinformatics [2].

To provide an effective outcome, data mining has to break down patterns and dataset connection to help the user obtain logical relationship and sequential patterns that help in building insights into data as well as drawing a conclusion from the data. As such, this process will be done in five sequential procedures, which includes; Data selection, preprocessing, transformation, analysis (data mining), and interpretations [3]. However, according to Cross-industry standard process for data mining (CRISP-DM), these phases can be regrouped into three stages; preprocessing, data mining and result validation from which each of three steps has two sub-steps giving a total of six phases of data mining[3] that can be summarized by Figure 1.



Figure 1:Architecture of data mining

Data mining itself takes entails six task which depends on nature of data to be mined ;anomaly detection, associative rule learning, clustering, classification ,regression and summarization [2] [3] .Owing to fact , that the methods is highly computation, relying on computer science, statistics ,technology education for its success as depicted by Figure 2.This leads to need of understanding different data mining platforms and how effective they are in various aspects. Thus, different platforms and methods are used depending on user need.



Figure 2:Position of data mining as function of statistics, computer science and Technology.

This paper is limited to comparative study and analysis of two data platforms(software); the SAS Enterprise miner and Python platform. Key principle objective is to evaluate the two platforms based on four criterions with consideration of only decision tree method of data mining. Python is an open source, high level, interpreted and general purposed programing language which is dynamically typed and garbage collected [4] .Additionally, pythons is highly versatile language that can support various programing paradigms employed in structural, reflective ,imperative, object oriented and functional programing owing to multiple standard libraries available at its disposal [4].

On other SAS Enterprise Miner is a proprietary own complete software developed and serviced by SAS institutes [5]. Unlike python SAS Enterprise miner is specially design for data mining and analytics providing comprehensive solution in reference to predictive analysis, business intelligent, criminal investigations, multivariate analysis, advanced analytics and database management [5] [6].

# **OVERVIEW OF DT Induction techniques.**

DT mining is a non-parametric data technique that for building classification model using tree -like structure and grouping as the name suggests. It one of commonly used data mining methods owing to its versatility despite its simplicity of its algorithms [7]. In these algorithms, target results are usually known and as such it classified as supervised data mining technique can be deployed in mining of both categorical and numerical data [7] [8].

A typical DT taxonomy consist of roots nodes, branches and leaf nodes each representing an attribute test, outcome of test and classification or decision respectively [9]; as shown in Figure 3. DT finds a vast application in classification analysis and regression models. In classification analysis, models that describes the crucial class variables are built and finds practical application mainly in machine learning processes and pattern recognitions [7] [9].



Figure 3:Illustration of DT nomenclature

. Classification analysis follows two basic steps, learning and model checking. Learning process entails training models on data built while the latter is essential checking model accuracy used to classify the new data in discrete form, that is yes or no, safe or risk [3].A general approach for building classification model is shown by Figure4.



Figure 4:A general approach for building classification DT induction model

In regression models, models are used to predict continuous value as opposed to class labels and the output is usually mean values of observable node values.

All DT methodologies works by splitting datasets into non overlapping and homogenous regions with application of top -down criterion [8]. Top most region is used to represent present observations which may be split into further branches. A greedy approach is employed since only current nodes between are worked on without any focused done on future nodes [9]. This subdivision is done repetitively until a stop criterion is reached. Tree pruning is later done to improve model accuracy.

DT induction mining technique is an employs algorithm that ensure there is no backtracking, that is a tree structure is developed in a top down recursive “divide and conquer manner” [3] [10]. It employs segmentation methodology crucial in analyzing large datasets with two phases; tree building phase and tree pruning phase.

Tree building phase is tasked with training data partitioning recursively until each data partition end up into small classes [11]. The splitting is done on attribute basis where most numerical attributes take the form of $a<=S$ ;where S is a real number split category for attributes form$A€C $ where C are elements or subset of A.

In tree pruning all data anomalies are training models are removing to make a tree less complex and small with key aim of improving accuracy [10] [11]. Both pre pruning and post pruning approaches may be employed depending on accuracy needed or data set under investigation. Prep running approach involves stopping construction modelling at early stages where last node becomes a leaf while post pruning involves removal of outlier branches from fully grown tree [11]. In a nutshell, prep running is a control design modeling that build a DT minimizing branches growth while latter takes effects after model designs.

* 1. *DT Induction Algorithms*

Implementation of DT induction data mining techniques deploys various algorithms depending on user application. However, all algorithm has great similarities and needs 3 basic parameters as input;

1. The data partition(D) which defines the training tuples set and their related class levels, that is, input training data [11].
2. A$ttribute \\_list$; represents attributes sets that defines a tuple
3. $Attribute \\_selection \\_method$; This represents heuristic approach for selecting attributes describes tuples according to the class in the best way possible and can be used to represent attribute selection measure as well [11].

To generate decision tree from training tuples of D, the following methodological step algorithm is commonly used [12].

* *Step 1*: Involve selecting test root node N. Branches for all possible outcomes are created.
* *Step 2:* Split in of instances into various subsets. Each branch extends from a node.
* *Step 3:* Recursively repeat each branch by using branches that reaches each instance.
* *Step 4*: Stop recursion for all branches whose instances shows same classes.

Most predominant algorithms employed by DT induction techniques includes;

*Iterative Dichotomiser 3(ID3):* This algorithm exploit on information gain to determine which subset [9] [13].It’s to be used in classifying current data subset. It one of earliest data mining algorithm that was developed by J. Ross Quinlan in 1980 before being extended other concepts of machine learning algorithm by E.B Hunt, J, and Marin [12]. The basic concept relied upon by this method is greedy top down search mechanism that involves selecting data attributes at various levels of DT, calculating information gain of each and using them as base for attribute selection, calculation of DT based on gain value of node and then building DT recursively until dataset obtain same classes [11] [12].

*C4.5*: Its successor algorithm to ID3 which has more improved aspects such as methods of handling numeric attributes, missing values, running to eliminate noisy data and generation of rules from DT [12]. It utilizes windows of cases fetched from the complete training set to develop its rules, before evaluating their goodness using criteria that measure the precision in case classifications [11]

*Classification and Regression Tree (CART):* The algorithm provides dynamic learning algorithm which can be used in both regression and classification modelling [10].

*Multivariate adaptive regression splines (MERS):* Finds wide application in multivariate regression-based mining. It was developed in 1991 by Jerome Friedman [14]. MARS can be viewed as an extension of linear modelling that can automatically model non linearities to find interaction between variables by both forward and backward passes. In this model, search knots are used as basic functions for finding data intersections between two splines [15].

*Chi-square automatic interaction detection (CHAID):* The algorithm was developed in South Africa and published by Gordon Kass in 1980 [3].It employs adjusted significance testing to perform multilevel split computation on classification DT [9].

*Condition Interference DT*: The algorithm is employs non parametric test statistical approach as the main splitting criteria, correction in multiple testing to reduce dataset over lifting [15]. As such, its results have unbiased predictor selection reducing need of pruning as in other algorithms.

## Attribute Measure Selections. `

Attribute measure refers to heuristic approach used for selecting the most appropriate splitting criteria for the DT induction algorithms [16]. These variable serves to measure degree of homogeneity of target variable in a subset providing a best-case scenario for new dataset classification [9]. Most common measure of optimal direction of splitting includes;

*Entropy(E)*

Its measure of the extent of impurities or randomness in datasets [15]. It serves to establish uncertainty in random variables commonly experienced in mined data. Its value ranges from 0 to 1 and the higher the value the more complete the data is assumed to be [15]. Entropy for multiple attribute can be expressed as in Eq.1

$$E\left(S\right)=\sum\_{i=1}^{c}-p\_{i}log\_{2}p\_{i} Eq.1$$

. Where p is likelihood of a tuple being associated with C and E(S) represent estimated information volume (on average) needed to find out class label(C) of dataset D

*Information gain(A)*

It measures of how best an attribute splits the training model according to the target classification [17]. It’s the most employed splitting criteria for DT induction mining technique that work by reducing number to test needed to classify any given tuple [9] [17]. The higher the gain the better data classification. It obtains by differences between two successive entropies and for any tuple in a dataset, D Information gain can be expressed by Eq.2

$$A\left(T,X\right)=\sum\_{C\in X}^{}P\left(c\right)E(c) Eq.2$$

Where T, X, P represent current state, selected state probability of a tuple being associated with class C respectively

*GINI Index*

It’s a measurement of how often random elements would be identified as incorrect, that is, it measures the impurities in a data partition, D [17]. Under these criteria attributes that have low Gini index are preferred for splitting. It widely applied in CART algorithms [12]. Eq.3 represent basic expression for GINI index.

$$Gini Index=1-\sum\_{j}^{}p\_{j}^{2} Eq.3$$

p is probability jth partition of a tuple being associated with class C

*Gain ratio*

 It measures of biasness towards test with varied outcomes helping data miner to identify usefulness of partitioning to various classifications [9] [13]. Its ratio of information gain(A) to Split information (split Info) as presented in Eq.4 while Split Info is expressed by Eq.5.

$$Gain Ratio=\frac{Information Gain}{SplitInfo} Eq.4$$

$$SplitInfo=-\sum\_{i=1}^{k}\frac{n\_{i}}{n}log\frac{n\_{i}}{n} Eq.5$$

Where,$ n\_{i}$is number of records in partition $i$ and k is partition in parent node. Other splitting methods includes; reduction in variance and Chi-Square.

# **Evaluation Criteria**

Data mining unlike other standard database file that are usually organized is associated with lot of peculiarities such as missing values, nonlinearities, inconsistency among others. Thus, data mining software developer are faced with great task of developing algorithm that offer great scalability, accuracy and reliability by offering extensive ways for handling different vagaries associated with mined datasets. Evaluation of data mining criteria will call for understanding of how different mining software work to yield needed result considering data variations [18].

Technological landscape is highly competitive and dynamic; a factor that has led to flood of data mining tool market with multiple software’s [18]. Such large basket to choose from places user at loggerhead on which software to use since just like the data mining itself, software peculiarities exist at different magnitudes. Choosing on which software to mine data with is thus left on hand of data miners since they have to select tools that meet just their mining needs. However, there are basic criteria that form base for any software selection especially when it comes to multitasking, interoperability, efficiency, hardware requirements, security etc. [19]..

Key features for data mining software selected for study will include, functionality, ancillary task support, usability and performances discussed below;

## Performance

Data mining software functionality is a core to any mining tool selection. The area focuses on quality and workability of software in reference to handling wide variety of data at different circumstances, driven by hardware configuration and algorithms dynamics [19]. The effects of software on hardware as well as data mining computation capability are great analytic functions of performance [18]. Key performance parameters;

*Working environment and Integrability:* Measures DM software can perform under different operating systems; Windows, Android, iOS, etc.

*DM Architecture:* It define how best can user customize system architecture or options available for adoption egg client -server, stand alone

*Data Access*: Define software interface required by the software.

*Data size*: Define the volume of data handled by software; that is complexity of data mining it can carry out.

*Efficiency:* Time take to produce reasonable data; speed of performance and measure how reliable DM software is.

*Interoperability:* Define how software can be interlinked with other tools i.e. use of extension, data importing or exporting.

*Reliability*: It measure of how consistence, robust or sturdy the system is, and its ability to operate without crashing. Should it fail, backup and recovery system activation are considered

Data access, size, efficiency, interoperability and reliability are the key measure of computation performance of DM software and are only one dealt with in later stages. However, architectural structure has been incorporated since hardware -software interaction affect performance to some at varying degrees.

## Ancillary Task Support

Ancillary task support refers to ability of software to support auxiliary task functions such as include processes that facilitate data fine tuning and handling discrepancies such as blanks or missing values [20],[22].Key support tasks include;

*Data filtering and cleaning*: Measure of ability of DM software to select data on given criterion and remove unwanted values respectively.

*Metadata manipulation*: Its ability to handle, modify or hold various data description; codes, descriptions, formulae or strings

*Handling blanks*: Its state how software handle blanks? Does it substitute them, can it handle blanks without corrupting the data set, or does the tool allow blanks to be substituted with user-defined value [22]

*Feedback*: How does system behave when it hit a snug? or how does it possible allow feedbacks to be used for other system modelling and alignment analysis?

*Randomisation*: Define how effective a tool can handle random data during or prior to data modelling.

Other auxiliary task support criteria include value substitution, binning, deletion and derivation of attributes.

## Functionality

It defines how well is a DM software suited in handling various data domains. How best can its handle varied algorithms and criteria it can provide to user for data mining [19]. Key parameters defining functionality includes;

*Algorithms variety and modifiability*: Its measure of how a mining tool is able to provide multiple mining algorithms; does it support DT induction mining technique [22] And how well can you change algorithms to suite user needs.

*Methodology:* Its measure of how DM tool carries out mining; sequentially (step by step), iteratively, etc.

*Model Validation*: It measure of DM tool’s ability support model refinement process in addition to creations to come up with best possible mine.

*Data flexibility*: Its ability to accept or reject variety of data types; continuous, discrete, conditional etc. without binning.

*Data sampling*: Its concern with how data miner’s software can accept random samples in predictive modeling.

*Reporting*: Concern on method through which data mining software can represent data in various forms, how best it can represent detailed report, how best it can select actual data record to fit target profiles [21].

*Model Exporting*: Measure ability to integrate model format into other tools such as Excel, SQL etc.

## Usability

Define how best a DM software can withstand different level of data and number of users without compromising its functionality, quality and integrity [20]. Usability also entails how best a miner can guide effective data mining as opposed to shift into data dredging. Most common criteria defining usability includes; user interface, data visualization, error reporting, action history, domain variety, user data type and ease of learning [20]. Data visualization is the most useful aspect that define how data do DM software present the data and is used for comparative case analysis for this criterion.

# **Description of the DT Induction software**

The 4 evaluation criteria discussed on Clause III will be performed on two platforms; python and SAS Enterprise miner.

## SAS Enterprise miner

***Overview of SAS Enterprise miner****.*

This software was developed in 1960 by the SAS Institute [5]. Upon its creation, the SAS software was primarily used for the management of data, business intelligence, predictive and descriptive analysis, and many others. With time, other new statistical functions were added into this software.

It is important to note that SAS software does not depend on the operating system. This software generally provides a solution for every domain of business such as data governance, quality, management of fraud, text mining, etc. [22]. Also, this SAS software helps in decision making by leveraging data that is already in existence into business intelligence environments

***Performance Evaluation of SAS Enterprise Miner***

1. Speed/Efficiency

The average run time for the SAS Enterprise Miner is 2.15 seconds [23]. This represents a high computational speed and many tasks can be executed over a short period of time. In addition, as a result of the dragging and dropping feature, components in this software can be put in use directly without much worrying about the coding part [5].

1. Data Handling Capability

SAS has a very good data handling capability and in addition, it can handle parallel computations [24].

1. Interoperability

SAS has a high interoperability with SQL and other enterprise directory servers [24].

1. Reliability

It very robust and sturdy software since all its mining algorithm and process data are held in hard disk other than in Random Access memory (RAM) [22]. Under normal condition chances of crashing are very low

1. Hosting variety criterion:

As earlier mentioned, the SAS Enterprise Miner software is independent of the operating system. This means that it can run on any platform, be it Windows or Linux [23].

1. Architecture criterion:

The software has a stand-alone architecture. This allows the software to access huge volumes of data efficiently, while at the same time enabling timely intelligence to many users [5]. This is achieved by the use of an n-tier architecture which ensures users can distribute functionality across various resources of computer so that every bit of work can be done by those resources that are best suited for that work [22]. This architecture comprises of four tiers namely: Data sources, SAS servers, Middle tier and Clients as shown in Figure 5.



Figure 5:SAS Enterprise Miner Architecture

When the connectivity criterion is considered, SAS Enterprise Miner employs the use of Input Data Source nodes that have the capability of extracting various data sources, scheduling, sorting, and conversions of formats [23]. It also captures data from the data set of SAS or other imported data and automatically creates dataset metadata.

 ***SAS Enterprise miner Functionality***

1. Variety of algorithms for decision tree induction:

SAS Enterprise software has a provision of various algorithms for decision tree such as Classification and Regression Trees, CHAID, and C4.5 [26]. These algorithms have three identifications that are multi-split and they split for categorical variables. Out of which, two help in the splitting of categorical and input variables that are continuous, that is Gini and entropy, while the other can only split the categorical input variable, which is the Chi-square test [26]. Two other multi-split algorithms are also provided by the SAS Enterprise Miner for the regression tree. These split for variables that have a numeric target, that is the Reduction in Variance and the F-Test [21].

1. Prescribed Methodology criterion:

The SAS Enterprise Miner employs the use of the Sampling, Exploring, Modifying, Modelling, and Assessing (SEMMA) data mining technique [27].

1. Data Flexibility

The flexible nature of the SAS architecture enables it to handle any amount of data and opens and entire world of SAS to data miners who have various skills ranging from business to technical experts [5].

1. Model Validation

The SAS Enterprise Miner uses logistic regression to fit a model in existence to new data.

1. Reporting

The SAS Enterprise Miner uses the PRINT procedure in generating simple reports which show the values of all variables and observations made in the data set [5].

1. Exporting

SAS Enterprise Miner uses the POC EXPORT statement to export data to other platforms [25].

 ***SAS Enterprise miner Usability***

1. User interface:

SAS Enterprise software contains an interface that is user friendly. In this software, the building of the models is very easy because it only involves simple clicks, drag and drops objects into the workspace [28].

1. Visualization:

This software provides various clustering results for graphs and charts. With regards to the decision tree, SAS a tree diagram containing roots, nodes, and leaf is provided by the SAS Enterprise Miner. This tree diagram helps in explaining the rules of the decision tree. Additionally, the SAS Enterprise Miner software provides charts that are very important for lifting. The software also provides a Report Node that helps in consolidating the results of these nodes within the flow diagram of the process in an HTML report [27]. These results are then displayed in a Web browser. It gives summarized statistics for both data that is interval-valued as well as categorical-valued data. However, the graphic capabilities of SAS are functional and will require additional concept of SAS graphical package which make customization of graphics difficult.

***SAS Enterprise miner Auxiliary Task Support***

1. Data Cleaning:

There are various ways provided by the SAS Enterprise Miner in solving issues of missing data values. Replacement nodes are provided that help in filling the missing values following some accurate data. When decision trees are considered, missing values are treated as an accepted value. Additionally, the SAS Enterprise software contains filter Outliers that get rid of missing data values from the current data values [28]. Tabs replacement and handling of blank is even.

This is accomplished in two ways: elimination of missing values from the flow chart of the process, and keeping values that are missing during the analysis. Besides, the software clears out situations of variables that are categorical-valued that does not occur. Besides, the SAS Enterprise Miner software helps in the removal of observations that are out of range in case of interval-valued variables [28].

1. Binning criterion:

This software contains a Transform Variables node that transforms variables that are interval-valued in the original set of data. The node does so by the use of three options for the transformation: buckets, quantile, and Optimal Binning. The bucket binning divides the values of data into equal intervals, whereas the quantile binning divides these values into equal classes [28]. The optimal binning on the other hand is for the Relationship to Target binning where it splits a variable into groups that have a binary target.

## The Python Language

 ***Overview of Python Language*.**

Python is a high-level structured programming language that is designed to be easy for a user to interact with several extensions. A python interpreter converts instructions written in python into machine language instructions. Python programming can be performed in either the interactive (listener or shell) or script modes. This language often uses keywords that are used in English and it contains fewer syntax constructions compared to other programming languages. The development of Python was seen in 1989 in the Netherlands at the National Research Institute for Mathematics and Computer Science [29].

***Python Language Computational Performance***

1. Speed/Efficiency

The average run time for Python is 2.44 seconds [24]. This represents a high computational speed and many tasks can be executed over a short period of time. Also, Python, being a high-level programming language, it provides an ideal choice for applications that are very critical yet fast. In the world of data analytics, Python is very easy to learn hence very efficient. However, it does not have a widespread Graphical User Interface (GUI) but it has Python notebooks that have become very popular in providing the user with documentation features and tutorials [28].

1. Data Handling Capability

Python has a very good data handling capability and in addition, it can handle parallel computations.

1. Interoperability

It’s easy to integrate it to lot of platforms swift for Tensor Flow and many other programming languages support Python [29]. One can easily import Python modules from the programming languages, call Python functions and convert values between Python and other programming languages.

1. Reliability

Its RAM dependent and thus subject to crashing when limit is exceeded [24].

1. Hosting variety criterion

Python contains an Operating System module that makes use of operating system dependent functionality. This allows the user to interact with the underlying OS that Python runs on.

1. Architecture criterion:

Python uses the setup.py file that describes the interaction of the user and the project, whether building, packaging, or installing it on the targeted operating system as shown in Figure 6.



Figure 6:Architecture of Python

1. Connectivity criterion

Python uses DB-API standard for the database interfacing, such as with MySQL, MySQL, Gadfly, etc. [29].

***Python Functionality***

1. Algorithms for decision tree induction

Python uses (CART) for predictive modeling of problems. This representation model is a binary tree [29]. Assuming the variable to be numeric, each node is a representation of a single input variable. The leaf nodes represent the output that is used to make predictions.

1. Many programming paradigms

Python can fully support object-oriented as well as structured programming. Also, Python employs a dynamic type of system and also automatically manages memory [31]. These features help the developer in creating very huge and complex software applications.

1. Robust standard library

This library enables the user to select from a various range of modules according to his or her specific needs [31]. These modules further allow the user to add more functionalities to the software application without any writing of additional code. Also, Python has a standard documentation library where users can browse and get information on a variety of modules [32].

1. Data Flexibility

Python has highly flexible features for documentation and sharing. It is very simple and has massive libraries for data manipulation. It also provides the user with a wide range of tools for analyzing and visualizing data [32]

1. Model Validation

Python uses holdout set to do model validation. This means holding back some subset of the data from the training model and then using and then using this tool to check the performance of the m Reporting Python uses *ply Reports* functions, which is a small and light module that provides a simple way of generating reports from databases using Python[31-32].

1. Methodology

Python uses Series (array-like objects that have an index data frames, rows and columns with two indices.) methodology to execute its instructions.

1. Exporting

Python uses Comma Separated Values (CSV) file that is a plain text file to export data to other platforms [27].

***Python Usability***

1. User interface criterion

Python programming language is very user friendly because it uses the most English keywords. This feature allows for easy readability of the code, which saves time while writing applications. Also, one can write customized applications without any need for additional code.

1. Visualization criterion

The Python language has a provision of very advanced graphical visualization tool for the tables, animations, histograms, and distribution displays aided by *Plotly,* *VisPy, and Matplotlib* function [29]. Also, Python helps in providing an evaluation of visualization such as Gains, Profit, Return on investment charts, Lift, and Response [30]

***Auxiliary Task Support provided by Python Language***

1. Data cleansing

Python can handle values that are missing in the data set by filling those values basing on already defined intervals. This is done by the use of either of these three options: by retaining the missing data, by estimating the missing data using simple methods, and by estimation of the missing data by the use of the complex method [31]. Support of blanks and tabs in non-uniform way is a key misdoing of the language

1. Binning criterion

Python has a provision of four binning options: Bins that are based on knowledge, bins that have equal range, bins that have equal size, and bins that are based on gaps [31].

# **Comparative Analysis in terms of the relevant criteria**

## Performance Comparison

SAS Enterprise Miner takes efficient and sturdy in its operation. Owing to the fact that most of its process data is stored in permanent memory, it eliminated holding of computational data reducing rate of crashing during operation. These storage increase binning and access time and thus average computational time increases significantly but increases data capacity and volume handled by software [32].

 On other hand, data handling capacity handled by python is limited by RAM and will crash whenever exceeded. Despite RAM attached issues python is quite sturdy. Data storage in RAM gives it very high execution time due to reduce access time which makes it faster than SAS miner.

Both software can be integrated into various platform and extension. Python exhibit seamless interlinkage to practically all external source when compared to SAS whose extension and linkages are limited mainly to Java based extensions. Ability of SAS miner to select and quickly recommend important variables makes it output very accurate [33]. Same effects are experienced in python (TensorFlow) which is computationally quick to select crucial variables to improve accuracy.

Table 1:

Comparative analysis summary of SAS Enterprise Miner and Python Language

|  |  |  |
| --- | --- | --- |
| Criteria | **SAS Enterprise Miner** | **Python Language** |
| 1.**Computational Performance** | *Speed /Efficiency* | Very high (average execution time of 2.45 second).Very high data accuracy | Very High (average execution time of 2.15 seconds)Relatively reduced accuracy in output |
| *Data Size/Capacity* | Very wide volume since data processed is stored in permanent memory (computer hard disks) | Medium data volume due to RAM limitations  |
| *Interoperability* | Integration with other software is not easy | Very flexible to integrate in multiple platforms |
| *Reliability* | Very efficient and study; not limited by memory. | Subject to crashing if memory is exceeded |
| 2**.Functionality** | *Data flexibility* | Both can handle numbers, characters, logical, complex, and arbitrary data types. |
| *Algorithms* | CART, CHAID, and C4.5 | Mainly uses CART (Keras and Tensor flow are key to building DT algorithms |
| *Model Validation* | Enhanced by use of logistic regression to fit a model in existence to new data | Enhanced validation by uses holdout set to do model |
| *Reporting* | uses the PRINT procedure in generating simple reports | Uses *ply Reports* function, which is a small and light module that provides a simple way of generating reports from databases |
| *Methodology* |  Use SEMMA approach which is stepwise  | Can use either parallel and serial execution that offer stepwise executions |
| *Exporting* | Allow exporting to multiple platforms |
| 3**.Usability** | *Visualization* | Provide only functional graphics, customizing plot is difficult and call for secondary SAS graphical package tools | Very versatile graphics tool aided by Plotly*,* *VisPy, Matplotlib function.* No need of secondary knowledge of extensions |
| User interface  |  Highly customizable and user friendly |
| 4.**Auxiliary Task support** | *Data cleaning, filtering, and* *Missing values*  | Missing values are treated as accepted values by *MI* and *MIANALYZE* functionsData filtering done via filter OutliersRemoval of out of range value Transformation, replacement and dropping is also catered for.Uniform tabs and blanks handling | Missing value is possible by filling those values basing on already defined intervals through *SimpleFill*, *MICE* (Multivariate Imputation by Chained Equations) and *Soft Impute* functionsRemoval of out of range value Transformation, replacement and dropping is also catered forTabs and blank handling are non uniforms |
| *Binning* | Support 3 binning options; buckets, quantile, and Optimal Binning |  Provision of four binning options: Bins that are based on knowledge, bins that have equal range, bins that have equal size, and bins that are based on gaps |

##  Functionality Criteria.

When it comes to functionalities, the two-software exhibit equal level of magnitude in all functionality criteria used. Both can handle numbers, characters, logical, complex, and arbitrary data types and thus handling data different data types is a major issue, However, python exhibit larger capacity and ability to perform more advanced data mining due to multiple libraries that enhance various DT operation; scikit-learn, TensorFlow and Keras while SAS only uses SAS/INSIGHT as primary entry to any data mining task [31].

Inters of handling DT induction technique algorithms, SAS provide 3 algorithmic implementation criteria via CART, CHIAD and C4.5. CART and C4.5 splitting continuous categorical variables either by use of entropy or Gini. CHIAD provide split by use of Chi-square. Multi-split of algorithm values is also e enhanced by other split criteria such as F-test and reduction variance [28]. In python, CART is mainly deployed in predictive DT modelling and thus model growth is limited by algorithms used.

In both cases, the two tools have parallel algorithms implementation with enhanced reporting in both cases. Implementation of algorithm is systematic in both case despite different criteria, where SAS deploys SEMMA, while python relies on serial arrays procedures for stepwise algorithms implementation [24]. Both provide rooms for data export through python has more advanced data export via plyReport functions. However, CSV files are common in python. Equally, model validation is enhanced in both cases where SAS relies on logistic regression to fit a model in existence to new data while python uses holdout set to do model validation.

## Usability Criteria.

 Both have user friendly interfaces since building codes is easy through simple clicks, drags and dropping of objects into the workspace.

Comparing visualization and graphical performance, Python by far provide excellent visualization platform than SAS. Graphics via python can be modified after creation mainly through Plotly*,* *VisPy, Matplotlib function* with no need of secondary knowledge of extensions. Chart lifting, adjustment and changing graph parameters are well enhanced. In SAS, visualization is also enhanced since DT can easily be drawn. However, its graphic capacity is functional; modification is an up task and call for understanding of SAS graphical package tools

## Ancillary Task Support.

Both software provide elaborate methods of cleaning, filtering, substitution ,handling missing value and dealing with blanks .In SAS, Missing values are treated as accepted values by *MI* and *MIANALYZE* functions while data filtering done via filter Outliers while in python language mining missing value is possible by filling those values basing on already defined intervals through *SimpleFill*, *MICE* (Multivariate Imputation by Chained Equations) and *Soft Impute* functions[34].

 Both have well elaborate binning criteria and option where SAS has 3 option; buckets, quantile, and Optimal binning while Python language offer 4 binning criteria; Bins that are based on knowledge, bins that have equal range, bins that have equal size, and bins that are based on gaps. A major difference in these areas comes in uniformity of handling tabs and blanks for which in SAS its uniform unlike in python.

1. **CONCLUSION**

DT induction technique is possible in both tools discussed through there exist slight variations from one tool to another. Such differences have been brought by architecture, proprietary and somehow operation period of the software. SAS Enterprise is an upcoming data mining software and thus has long development needed to make all it component satisfactory. Its proprietary software developed by SAS institute and thus costly and subject to slow development since it not opens for liaison development by other developers as in the case of python which is free source.

The differences in development and cost makes them ideal in different application; SAS used in government and enterprises while python is general proposed language. It thus obvious that python is the most ideal for startups due to its lightweight nature and high advancement in deep learning; a key area in DM.

In comparative study, it evident that none can explicitly be branded better than other in overall comparison since both exhibit varying score cards in various criteria evaluated. For instance, SAS possess better accuracy but slow execution time which is totally opposite to python. Overall, SAS poses better computation performance due to lack of hardware limitation as in case of python. Functionalities and auxiliary task support are equally same despite use of difference execution procedure of handling algorithms, data variety, cleaning and binning. Auxiliary support role in blanks and tabs uniformity differed in two platform with python shows uneven replacement. A sharp difference arises in usability criteria where python is by far the best; thanks to its complex graphical capabilities that can allow easy creation and modification.

# **References**

|  |
| --- |
| 1. A. Twin, "Data Mining," 18 August 2019. [Online].Available: https://www.investopedia.com/terms/d/datamining.asp. [Accessed 29 April 2020].
 |
| 1. R. P, "Top 14 useful applications for data mining," 20 August 2014. [Online]. Available:https://bigdatamadesimple.com/14-useful-applications-of-data-mining/. [Accessed 29 April 2020].
 |
| 1. L. A. K. a. P. Musilek, "A survey of Knowledge Discovery and Data Mining process models," *Journals of Knowledge Engineering Review,* vol. 24, no. 1, pp. 1-24, 2004.
 |
| 1. D. Kuhlman, A Python Book: Beginning Python, Advanced Python, and Python Exercises, 2012.
 |
| 1. "About SAS," SAS Institute, [Online]. Available:https://www.sas.com/en\_us/company-information.html#history. [Accessed 29 April 2020].
 |
| 1. Denver, "SAS empowers crime fighters to crack complex cases," SAS global Forum, 09 April 2018. [Online]. Available: https://www.sas.com/en\_us/news/press-releases/2018/april/fraud-law-enforcement-visual-investigator-sgf181.html. [Accessed 29 April 2020].
 |
| 1. "Decision Tree Algorithm Examples In Data Mining," 16 April 2020. [Online]. Available:https://www.softwaretestinghelp.com/decision-tree-algorithm-examples-data-mining/. [Accessed 30 April 2020].
 |
| 1. "Data Mining Classification: Decision Trees,"[Online].Available: http://staffwww.itn.liu.se/~aidvi/courses/06/dm/lectures/lec3.pdf.
 |
| 1. A. F. H. A. a. A. S. A. A.-H. Radhwan H. A. Alsagheer, "Popular Decision Tree Algorithms of Data Mining Techniques: A Review," *International Journal of Computer Science and Mobile Computing,* vol. 6, no. 6, p. 133 – 142, 2017.
 |
| 1. "Data Mining - Decision Tree Induction," [Online].Available:https://www.tutorialspoint.com/data\_mining/dm\_dti.htm. [Accessed 30 April 2020].
 |
| 1. "Decision Tree Induction," [Online]. [Accessed 29 April 2020].
 |
| 1. S. c. a. Dr.B.Rama, "A Survey on Comparative Analysis of Decision Tree Algorithms in Data Mining," *International Journal of Advanced Scientific Technologies ,Engineering and Management Sciences (IJASTEMS),* vol. 3, no. 1, pp. 91-95, 2017.
 |
| 1. A. a. K.Arunesh, "Role of Decision Tree Classification in Data Mining," *International Journal of Pure and Applied Mathematics,* vol. 119, no. 15, pp. 2533-2543, 2018.
 |
| 1. M. R. a. M. Seema, "Decision Tree: Data Mining Techniques," *International Journal of Latest Trends in Engineering and Technology (IJLTET),* vol. 1, no. 3, pp. 150-155, 2012.
 |
| 1. N. S. Chauhan, "Decision Tree Algorithm — Explained," Towards Data Science, 24 December 2019. [Online]. Available: https://towardsdatascience.com/decision-tree-algorithm-explained-83beb6e78ef4. [Accessed 30 April 2020].
 |
| 1. D. Nettleton, Commercial Data Mining, Elsevier Inc., 2014.
 |
| 1. S. G. P. a. D. K. I. L. Bhaskar N. Patel, "Efficient Classification of Data Using Decision Tree," *Bonfring International Journal of Data Mining,* vol. 2, no. 1, pp. 6-12, 2012.
 |
| 1. "Selection Criteria for Data Mining Software: A Study," *International Journal of Computer Science,* vol. 10, no. 3, pp. 308-312, 2013.
 |
| 1. V. V. a. S. Dhawan, "Methodology for Selection of a Data Mining Tool," *International Journal of Software and hardware Research in Engineering,* vol. 2, no. 5, pp. 189-194, 2014.
 |
| 1. F. N. a. O. a. G. N. Ogwueleka, "Methodology and tool selection criteria in data mining," *Journal of Engineering, Technology and Innovation,* vol. 1, no. 6, pp. 131-137, 2012.
 |
| 1. C. G.-C. a. O. Povel, "Characterising Data Mining software," *Intelligent Data Analysis,* p. 181–192, 2003.
 |
| 1. "Python vs SAS vs R," [Online]. Available:https://mindmajix.com/python-vs-sas-vs-r. [Accessed 02 May 2020].
 |
| 1. "Python vs R vs SAS," Besant Technologies, 23 April 2019. [Online]. Available:https://www.besanttechnologies.com/python-vs-r-vs-sas. [Accessed 02 May 2020].
 |
| 1. K. Jain, "Python vs. R vs. SAS – which tool should I learn for Data Science?," 12 September 2017. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/sas-vs-vs-python-tool-learn/. [Accessed 01 May 2020].
 |
| 1. H. Eric, "SAS Data Quality – A Technology Overview," *SAS Inc..*
 |
| 1. M. C. N. a. J. P. Jim Brittain, "Data Scientist’s Analysis Toolbox: Comparison of Python, R, and SAS Performance," *SMU Data Science Review,* vol. 1, no. 2, 2018.
 |
| 1. B. Soifua, "A Comparison of R, SAS, and Python Implementations of Random Forest," All Graduate Plan B and other Reports. , 2018.
 |
| 1. U. S. a. H. Qaiser, "A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)," *International Journal of Innovation and Scientific Research,* vol. 12, no. 1, pp. 217-222, 2014.
 |
| 1. L. W. a. P. Pfeiffer, "A Qualitative Analysis of the Usability of Perl, Python, and Tcl," 2002.
 |
| 1. C. Hayward, "Usability, the Soul of Python: An Introduction to Programming Python Through the Eyes of Usability," [Online].Available: https://cjshayward.com/python/. [Accessed 02 May 2020].
 |
| 1. "2019 SAS, R, or Python Survey Update: Which Tool do Data Scientists & Analytics Pros Prefer?," Burtch Works, Evanston, 2019.
 |
| 1. J. M. R. a. F. Ortin, "A Comprehensive Evaluation of Common Python Implementations," University of Oviedo, 2015.
 |
| 1. "SAS vs R vs Python – The Battle for Data Science!," TechVidvan, 24 February 2020.[Online].Available: https://techvidvan.com/tutorials/sas-vs-r-vs-python/. [Accessed 02 May 2020].
 |
| 1. B. C. S. a. C. M. Ken Collier, "A Methodology for Evaluating and Selecting Data Mining Software," in *Proceedings of the 32nd Hawaii International Conference on System Sciences*, 1999.
 |

1. This word is constantly used as opposed to software since python is a programing language while SAS Enterprise miner is a complete proprietary software application written in C programming languages. As such, use of platform will equal magnitude level for comparative study and analysis. [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)