



## Contents

Overview .....	3
Objective .....	4
Background .....	4
About eProcurement .....	4
About Data Analytics and Artificial Intelligence .....	5
How Machine Learning can help to analyse large databases such as the e-procurement database .....	5
About the database .....	5
Participating Members.....	6
Project Overview.....	6
Algorithms used in the Project .....	6
Technology Stack used for the implementation.....	6
Key functions performed by the project.....	7
Functioned based on Data Analytics.....	7
1. Checking the validity of Pin codes and Permanent Account Number (PAN) in the dataset... ..	7
2. Categorizing the tenders based on the tender validity period .....	8
3. Finding tender-status wise number of tenders .....	8
4. Finding the month-wise and year-wise unique work-items created in the database .....	9
Functioned based on Machine Learning.....	10
1. Identifying the splitting of tenders .....	10
2. Clustering the tenders based on the differences between L1 & L2 and L1 & L3.....	12
3. Calculating the Risk Score based on the data available in 'gep_tender_work_items' and 'gep_bids' tables .....	12
Conclusion.....	13
About the Accuracy of the modules .....	13
Areas for further improvement .....	13
Recommendations for use in the Indian Audit & Accounts Department .....	14
Appendix 1 – Analytic techniques/algorithms used in the project.....	15
K Medoids Clustering to cluster work items based on difference between L1 & L2 and L1 & L3 ....	15
Gaussian Mixture Model Clustering for checking correctness of the data.....	15
K-Means Clustering for checking the ranks of accepted AOC Bids and Evaluators Value .....	16
Cosine Similarity Clustering for identifying the Splitting of tenders.....	16
Appendix 2 – Dimensions identified to calculate the degree of similarity between the work items ..	17
Appendix 3 – Using the project.....	18
Setting up of running environment .....	18
Running the project .....	18
Using different dataset in the project.....	19
Modifying the source code of the project .....	19

Banner image on the cover page taken from the website of eProcurement System, Government of Punjab ([eproc.punjab.gov.in](http://eproc.punjab.gov.in))

## Overview

An important tool for the development of the fullest potential of the people working in any organisation is the adoption of adaptive, topical and responsive training and capacity development mechanisms.

With a view to strengthen the practice –oriented research and project work undertaken in its Knowledge Centre, the Regional Training Institute (RTI), Jammu has introduced several innovative measures. Among these, is the conduct of a 9-week long internship programme for two 2<sup>nd</sup> year B. Tech students of the Indian Institute of Technology (IIT), Jammu as part of a collaborative alliance with the premier Institute.

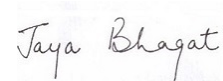
The adoption of e-procurement mechanisms by various State Governments pose new challenges with regard to the conduct of a comprehensive and focused audit. The RTI team and the interns worked together on a project that sought to use Data Analytics and Artificial Intelligence to analyse the e-procurement system of the Government of Punjab and build systemic methods to query the transactions and bids in the database.

The resultant research, outlined in this paper being shared with the reader, showcases both the strengthening of audit methodologies and the enhanced ability to analyse large volumes of data to arrive at an audit conclusion, through the use of new technology.

While sharing the project, with all offices of the Indian Audit and Accounts Department (IA&AD), our team would like to thank the office of the Principal Accountant General (Audit), Punjab, for their cooperation in providing data and feedback whenever requested during the course of the research project.

The source code is available with the RTI Team and we would be happy to share the same with any sister office in the IAAD on request.

Any suggestions for improvement are welcome.



**Jaya Bhagat**  
**Director General**

**Jammu, 12<sup>th</sup> November 2020**

## Objective

The objective of this project is to enhance the efficiency of audit teams by equipping them with an analytical tool that would help them to ascertain whether an adequate system is in place in the auditee units with regard to efficiency and transparency in public procurement. The project is intended for use by the external or internal auditors in an organization and is expected to make their job easier by way of scientific analyses of the process of handling public procurement by the Government as enhanced through automation and process re-engineering. The project should also be of help the auditors in determining as to whether the eProcurement system has facilitated greater clarity for the Government with regard to the overall picture of its procurement activities.

## Background

### About eProcurement

Public buying is an essential component of activities of major Government departments/ PSUs/ other bodies and authorities and therefore the procedure to be followed in public procurement must therefore conform to the yardsticks prescribed in the relevant clauses of the General Financial Rules, as enunciated in the Fundamental principles of public buying (for all procurements including procurement of works).

Every authority delegated with financial powers of procuring goods in the public interest shall have the responsibility and accountability to conduct such procurement keeping in view the tenets of efficiency, economy, and transparency, thereby ensuring the fair and equitable treatment of suppliers and promotion of competition in public procurement. In the present digital era, the mode of procurement to be followed is desirable through electronic means.

e-Procurement (electronic procurement, sometimes also known as supplier exchange) is the business-to-business or business-to-consumer or business-to-government purchase and sale of supplies, work, and services through the Internet as well as other information and networking systems, such as electronic data interchange and enterprise resource planning.

The e-procurement value chain consists of indent management, e-Informing, e-Tendering, e-Auctioning, vendor management, catalogue management, Purchase Order Integration, Order Status, Ship Notice, e-invoicing, e-payment, and contract management. Indent management is the workflow involved in the preparation of tenders. This part of the value chain is optional, with individual procuring departments defining their indenting process. In works procurement, administrative approval and technical sanction are obtained in electronic format. In goods procurement, indent generation activity is done online. The end result of this stage is taken as inputs for issuing the NIT.

Elements of e-procurement include request for information, request for proposal, request for quotation, RFx (the previous three together), and eRFx (software for managing RFx projects). With increased use of e-procurement, needs for standardization arise. Since the audit office is directly connected with scrutinising all the procedures for procurement, a need was felt by the RTI Team to prepare a module that would enable analysis of data relating to procurement procedures using Artificial Intelligence & data analytics. This module has been prepared in collaboration with IIT, Jammu, on the basis of data obtained from office of the PAG (Audit), Punjab.

## About Data Analytics and Artificial Intelligence

**Data analytics** is the science of analysing raw data in order to arrive at conclusions about that information. Data analytics techniques can reveal trends and metrics that would otherwise be lost in the mass of information. The scope of data analytics can be broad, from basic statistical models, business intelligence (BI), online analytical processing (OLAP) to various advanced analytics.

**Artificial intelligence** refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. **Machine learning** is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. It focuses on the development of computer programs that can access data and use it learn for themselves.

## How Machine Learning can help to analyse large databases such as the e-procurement database

With the digitization of most of the activities of the government departments, it becomes a necessity to upgrade the audit techniques as well. Use of data analytics to derive insights on a smaller dataset is generally sufficient because smaller datasets generally exhibit trends or the patterns that are simpler in nature and are easier to understand.

When it comes to working with large databases (such as the one used in this project), which hold data of multiple departments and further multiple offices and work items, data analytics may not succeed in detecting the correct patterns from the voluminous data. This is where the need for the use of Machine Learning (ML) is felt.

ML can help the auditor to obtain a general view the complete database and to detect patterns and anomalies in the database. It can also help the auditor to audit the complete population instead of a small sample thereby making the audit much more broad-based.

Since ML algorithms learn from the data that is analysed and can be programmed to learn from the user feedback, it can learn from the auditors and apply the learning to analyse data and point out anomalies that satisfy the conditions defined by the auditor.

A higher volume of data acts as an advantage when working with ML because it results in more training data being available to train the ML model. This helps the ML models to get trained for a varied range of scenarios which in turn, is reflected in more accurate results/predictions.

## About the database

The database of the Punjab Government eProcurement System has been used in the project. The database was maintained as a PostgreSQL DB which is an open-source database. The database contained the tenders for all the major departments of the Punjab Government including the PWD, Agriculture, Food Civil Supplies and Consumer Affairs, Housing and Urban Development, Power, Local Government, Transport and Water Resources. The total size of the database was 4.3 GB and the database comprised 1034 tables (including the master tables). Based on preliminary analysis, the following tables were identified for being used for the project:

1. gep\_tender\_work\_items – The table contained tender ID, work item reference number, tender title, work description, tender value, tender validity, tender inviting officer and other information pertaining to a tender.
2. gep\_bids – The table contained tenderer ID, work item ID, bid reference ID, bid status, bid rank and other details of the bids placed through the eProcurement System.
3. gep\_bid\_aoc\_details – The table contained bid ID and bid award value of the successful bids.

4. `gep_tenderer` – The table contained the details about the firms and contractors registered in the eProcurement System.
5. `gep_tender_basic_details` – The table contained basic details about the tenders including the tender reference number and tender ID.
6. `gep_orgchain_master` – This was a master table that contained information about the departments that invite tenders through the eProcurement System.
7. `gep_product_category` – This was a master table that contained the types of the works and services and their associated IDs used in the eProcurement System.

## Participating Members

### 1) The Indian Institute of Technology, Jammu

- i) Ms. Arya, Pursuing B. Tech in Computer Science and Engineering (2018-2022)
- ii) Ms. Unnam Pearly Susan, Pursuing B. Tech in Computer Science and Engineering (2018-2022)

### 2) The Regional Training Institute, Jammu

- i) Ms. Jaya Bhagat, Director General (Overall guidance)
- ii) Sh. J K Pandita, Sr AO (Training)
- iii) Sh. Aseem Beetan, AAO (OIOS)

### 3) Office of the Principal Accountant General (Audit), Punjab

- i) Ms. Monika, Sr Audit Officer (IT)

## Project Overview

### Algorithms used in the Project

The data available in the dataset was unlabelled. Therefore, the algorithms used in the project belong to the Unsupervised Learning domain which are used to draw inferences from the datasets consisting of data without labelled responses. The algorithms used in the project include K-Means clustering, K-Medoids clustering, Gaussian Mixture Model clustering and Cosine Similarity clustering. Details of the algorithms can be found in Appendix 1.

### Technology Stack used for the implementation

**Database** – The original database of which the dump was provided by the office of the Principal Accountant General (Audit), Punjab was maintained as a PostgreSQL DB. Therefore, PostgreSQL Database was used to import the original data dump and for preliminary analysis of the data.

**Programming Language** – The Python programming language has been used for all the programming done in the project. Python is one of the most commonly used programming languages in the field of Data Analytics and Artificial Intelligence as it has a wide array of libraries to support AI related programming and it offers a high level of flexibility to build new models from scratch.

R programming language which offers a similar set of libraries and functionalities would also be suitable for use in the project as an alternate programming language.

**Code Integration** – Jupyter Notebook has been used for integrating all the code for the project. It is a web-based interactive computational environment for writing and executing code. In addition to working as an Integrated Development Environment, it can also be used to run and demonstrate the project.

The following libraries of the Python Programming Language have been used for developing various functionalities of the project:

- Pandas – For manipulation of data
- Numpy – For working on large multi-dimensional arrays/matrices
- Datetime – To make computations on date and time
- Gradio – To create User Interface for the project
- Sklearn – To implement machine learning algorithms
- Matplotlib – For generating the visualizations

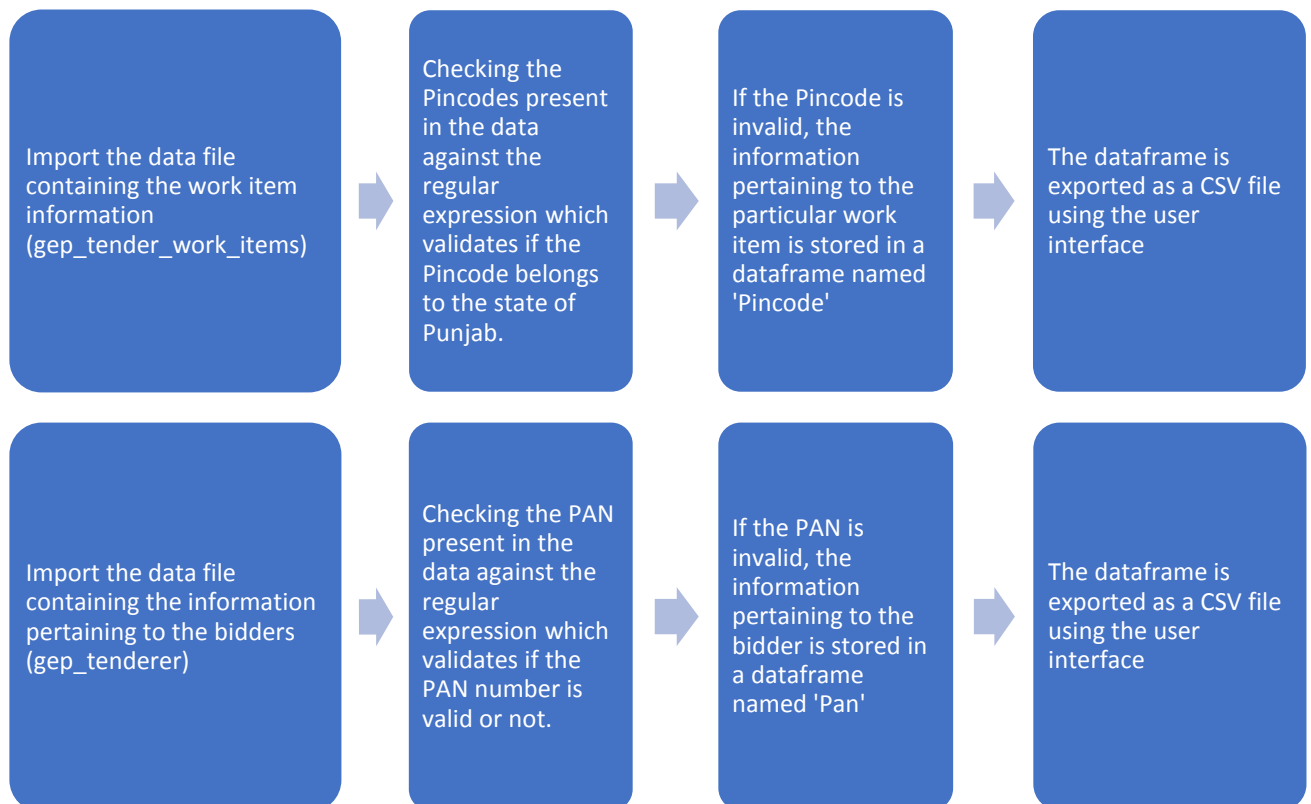
## Key functions performed by the project

### Functioned based on Data Analytics

This project demonstrates application of basic data analytics which can be used not only to point out discrepancies in the data but also the lack of controls in an IT System being used for the purpose of maintaining the data. The modules developed can perform the following functions:

#### 1. Checking the validity of Pin codes and Permanent Account Number (PAN) in the dataset

Lack of data validation can lead to invalid entries in the databases. This module checks the data for the presence of the invalid Pin codes (in this instance, Pin codes that do not belong to the state of Punjab <sup>1</sup>) and PAN. It also allows the user to export datapoints that contain invalid entries as a CSV file.

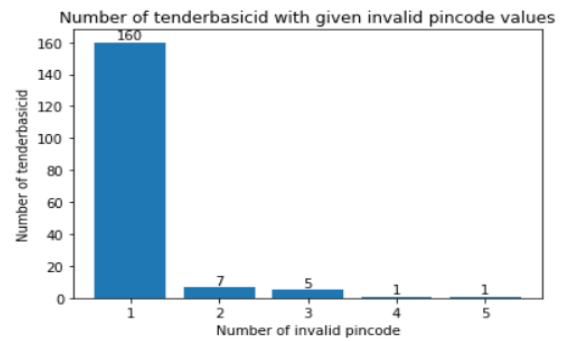


<sup>1</sup> Based on the presumption that most procured items would be through in-state or local purchase.

**SHOULD THE FILE BE EXPORTED**

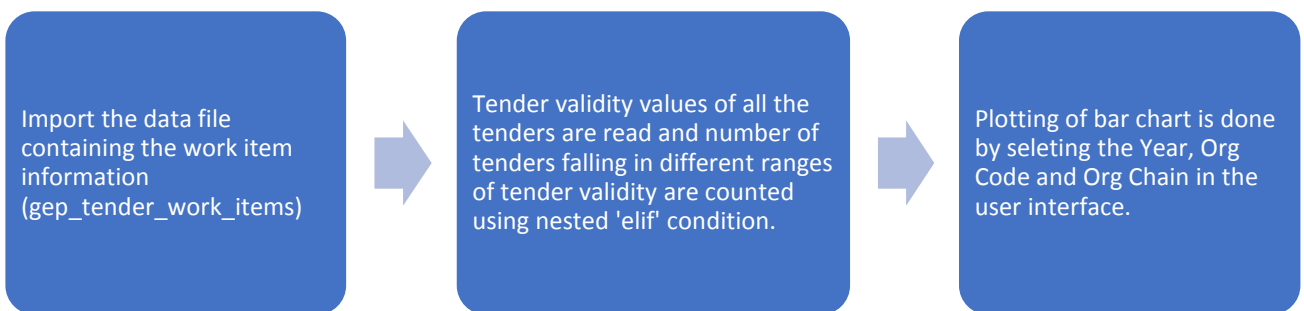
Yes

CLEAR SUBMIT



## 2. Categorizing the tenders based on the tender validity period

This module categorizes tenders into 9 categories according to the tender validity period. The user can apply the filters on the Year, Org Code<sup>2</sup> and Org Chain.<sup>3</sup>



**YEAR**

2010  2011  2012  2013

2014  2015  2016  2017

2018  2019

**ORG CODE**

ADBFP

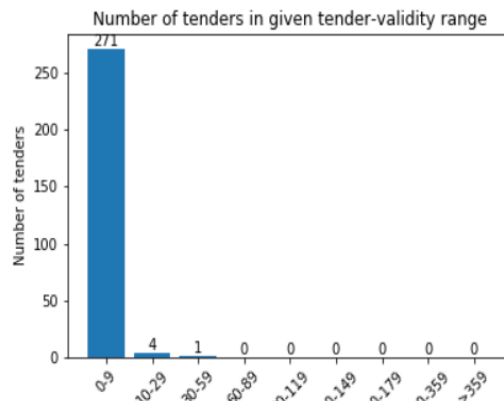
**ORG CHAIN**

CE-Bldg

**SHOULD THE TABLE BE EXPORTED**

Yes

CLEAR SUBMIT



Multiple years can be selected at a time and the Org Chain chosen should have the selected Org code. A new column named “ValidityCategory” is added to the filtered data which provides the Category as per the tender validity. This newly formatted data can be exported as a CSV file. Also, a bar graph may be generated showing the number of works in each category from the filtered data.

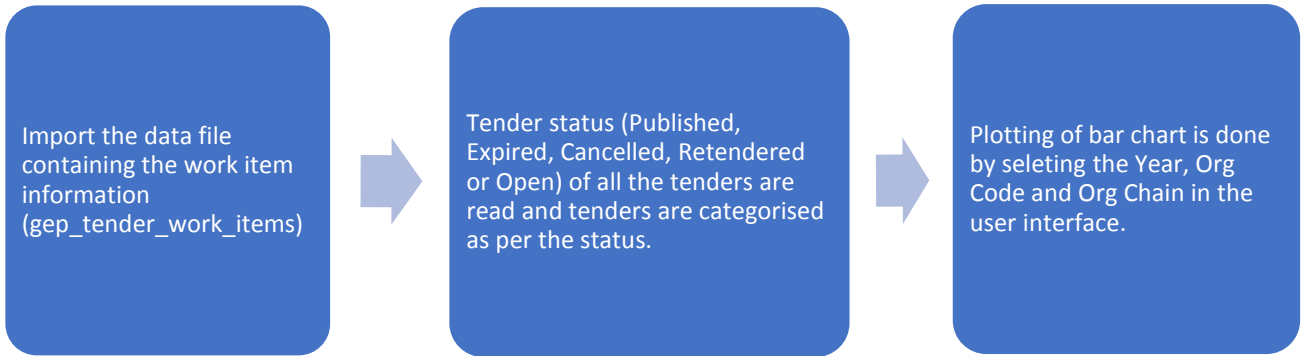
## 3. Finding tender-status wise number of tenders

Tenders in the eProcurement database used in this project have been assigned ‘tender-status’ which is based on the status of the tender at a point of time.

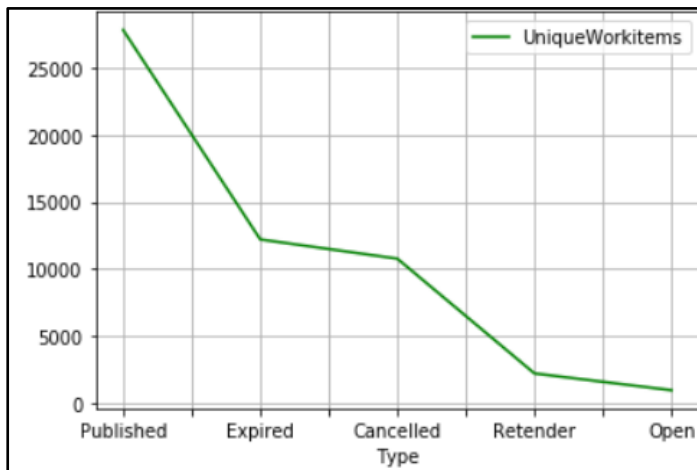
<sup>2</sup> Org Code refers to a unique alphabetical code used to identify each department in the database.

<sup>3</sup> Org Chain outlines the path from the apex level of the Organization to the specific office which has floated the tender





A tender can have any one of the following statuses: Published, Expired, Cancelled, Retender and Open. This module allows the user to see the number of tenders for each tender-status. Similar to the previous module, the user can apply filters on the Year, Org Code and Org Chain. The data satisfying the conditions entered can be exported as a CSV file.

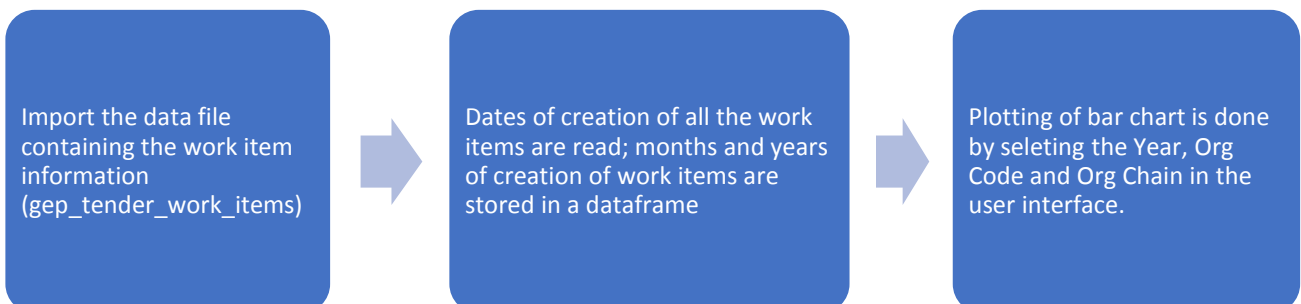


A line plot showing the number of unique work items in each tender status is generated.

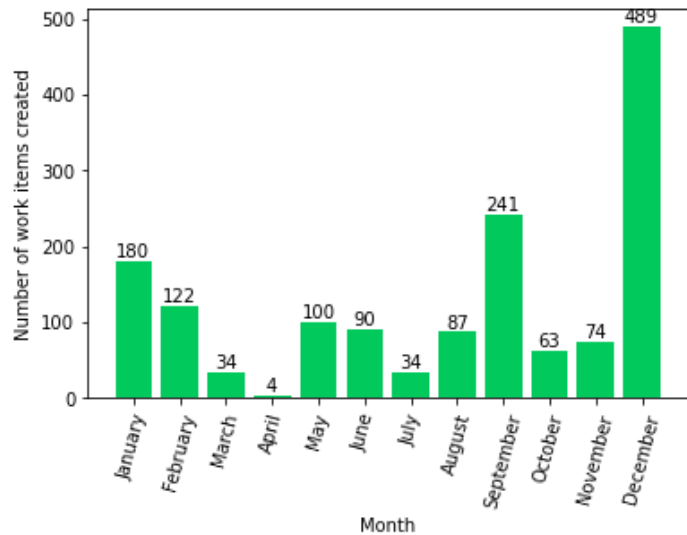
Also, irrespective of filters selected, the graph for the whole table is also displayed in the Jupyter notebook.

#### 4. Finding the month-wise and year-wise unique work-items created in the database

Work items in the eProcurement database may sometimes have similar names and descriptions. Each work item has a unique work item ID associated with it which can be used to identify each work item. With the help of work item IDs, this module finds the number of unique work items created in each month of a year. The user selects the Year, Org Code and Org Chain using the user interface and has an option to export the filtered data as a CSV file.



A bar graph showing the number of unique work items in each month of the selected year for the selected Org Chain is generated.

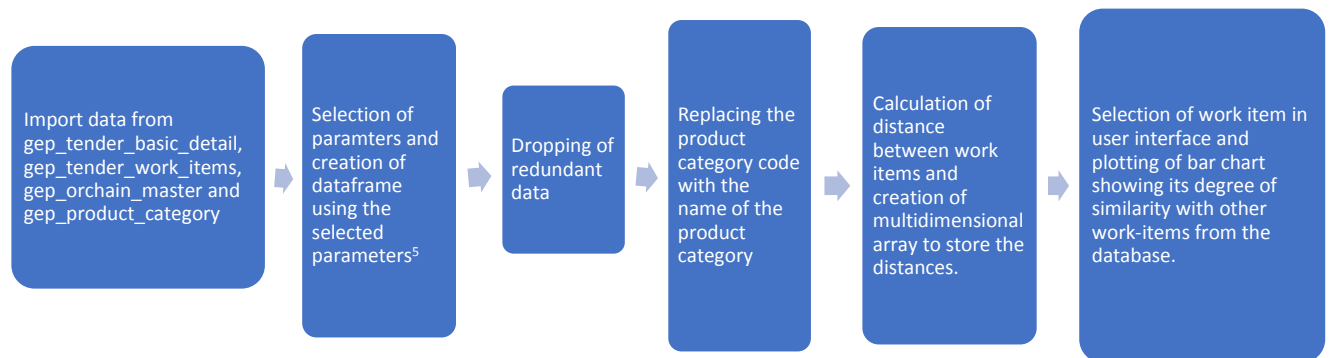


## Functioned based on Machine Learning

This project further demonstrates the application of the Machine learning in the field of audit with the help of different modules that have been developed based on the audit criteria followed when auditing eProcurement activities of a government department. The modules developed using the Machine Learning algorithms can perform the following functions:

### 1. Identifying the splitting of tenders

The module computes the similarity between the work items based on multiple factors<sup>4</sup> and therefore indicates the tenders which are likely to have been split from a single tender thereby helping the auditor to identify work items that need to be scrutinized in depth. This helps to avoid the problem of ‘salami slicing’ in procurement.



For extracting the details of tenders, the tender\_basic\_details and tender\_work\_items tables have been used. Out of the 54000 tenders, 27570 work items have corresponding entries in the tender\_basic\_details table. After merging the tables, all the null columns are dropped. The product categories in the dataset are represented by 143 numerical codes. These numbers are then replaced by their text counterparts using the master table gep\_product\_category. For the categorical data types, distance is calculated using the dice metric, where distance is considered ‘1’ when the values are unequal and ‘0’ otherwise. For the continuous text data types, the distance between the two texts is inversely proportional to similarity. The similarity is calculated using the Cosine Similarity method. For the continuous real values, the distance is the absolute difference between the two values divided

<sup>4</sup> Appendix 2 - Dimensions identified to calculate the degree of similarity between the work items

by their range. For the continuous DateTime values, the number of days between the days divided by the range of days from a beginning date (for dataset used, 17th June 2010).

The distance between two tenders is therefore calculated using the individual distance metrics for different types of columns. For each column, the distance between them is calculated and multiplied with weights assigned to each column based on their impact on the splitting results and added. Using this method, for any tender selected by the user, the twenty most similar tenders are identified using the distance calculation and exported to a CSV file.

The source code can be modified to generate a list containing more than 20 work items. However, in the instant case, the most similar work items that may exhibit a case of ‘salami slicing’ usually form part of the first 10-15 items and remaining items may exhibit a low degree of similarity with the selected work item.

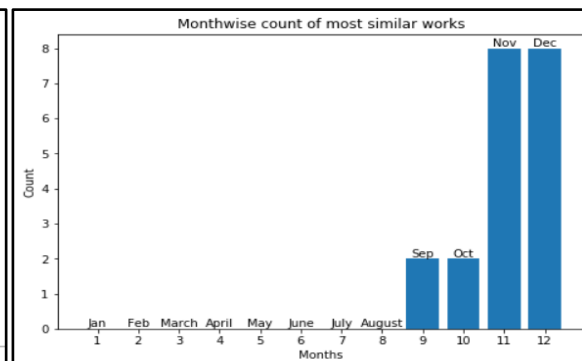
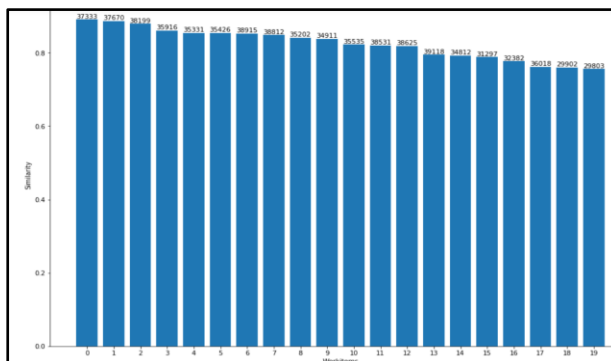
<b>ORG CODE</b> <input type="text" value="Coop"/>	<b>OUTPUT</b> <input type="text"/>
<input type="button" value="CLEAR"/> <input type="button" value="SUBMIT"/> <input type="button" value="SCREENSHOT"/>	

<b>ORG CHAIN</b> <input type="text" value="Deaprtment of Cooperati..."/>	<b>OUTPUT</b> <input type="text" value="Continue"/>
<input type="button" value="CLEAR"/> <input type="button" value="SUBMIT"/> <input type="button" value="SCREENSHOT"/>	

<b>WORK-ITEM ID</b> <input type="text" value="37479"/>	<b>OUTPUT</b> <input type="text"/>
<input type="button" value="CLEAR"/> <input type="button" value="SUBMIT"/> <input type="button" value="SCREENSHOT"/>	

<b>SHOULD THE FILE BE EXPORTED</b> <input type="text" value="Yes"/>	<b>OUTPUT</b> <input type="text"/>
<input type="button" value="CLEAR"/> <input type="button" value="SUBMIT"/> <input type="button" value="SCREENSHOT"/>	

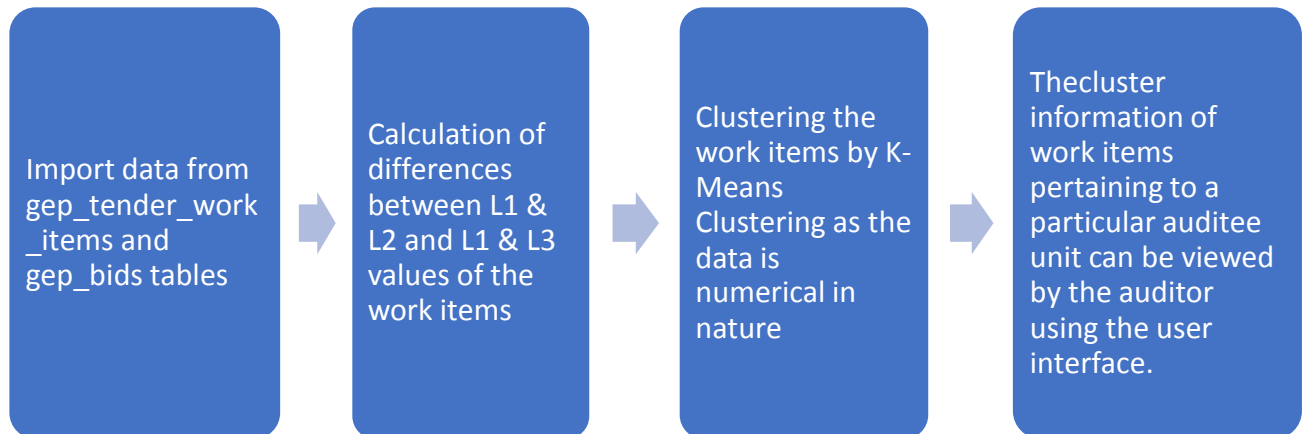
The details of the 20 work items are also generated in a tabular form which can be exported as .csv file for further analysis.



Bar charts depicting the ‘degree of similarity of the predicted work items and the selected work item’ and ‘month-wise number of similar works issued’ are produced for the user.

## 2. Clustering the tenders based on the differences between L1 & L2 and L1 & L3

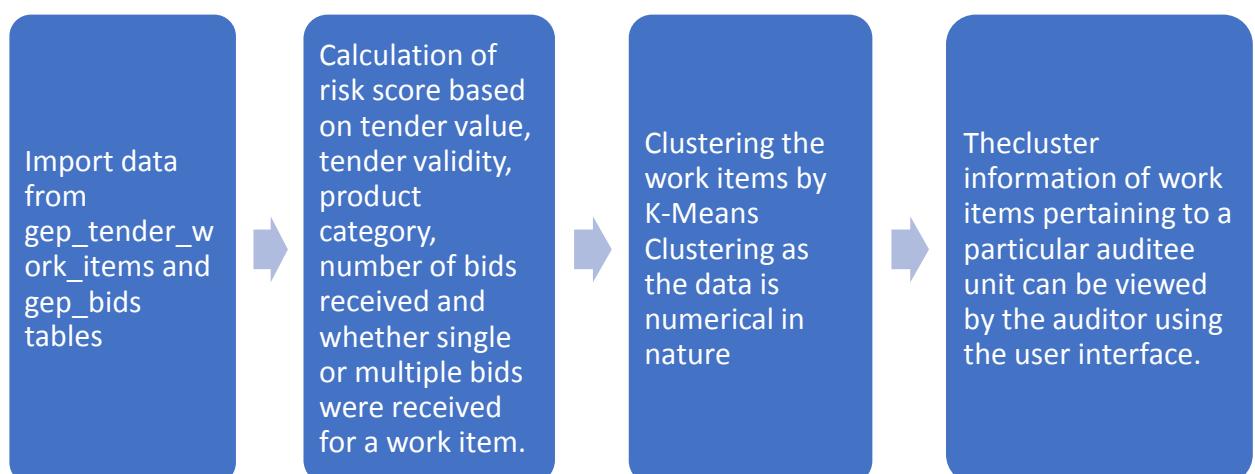
Differences between the L1, L2 and L3 generally indicate the degree of competition in the tendering process. A substantial difference between the lowest bid values can sometimes be due to collusion in tendering by the bidders. The absolute differences in the bid values may vary widely depending on the nature and value of the bid. Therefore, the module clusters the tenders based on the relative differences in the bid values with the view to find out the tenders for which the differences are unrealistically high.



It calculates and assigns a risk score to the work items based on these differences. Tenders with high risk scores can subsequently be looked into in detail by the auditor as part of a risk-based audit.

## 3. Calculating the Risk Score based on the data available in 'gep\_tender\_work\_items' and 'gep\_bids' tables

This module uses the data available in the 'gep\_tender\_work\_items' and 'gep\_bids' to train an ML model and cluster the work items as per the risk score which is calculated based on factors such as tender value, tender validity, product category, number of bids received, L1, L2 and L3 values, differences between L1 and L3 values and whether retendering was done if a single bid was received for the work item. Thereafter the users can enter the hypothetical bid details to calculate a risk score for those bids. The risk score helps the auditors to identify tenders/work items that need to be scrutinized. This module can be used by the auditors to calculate the risk score of any bid for which basic information is available.



In a situation when required information is not available for a few options, the user can input an approximate value to calculate the risk score.

TENDER VALUE	
TENDER VALIDITY	
EXPIRE PERIOD	
PREBIDOPTION	0
CONTRACTTYPE	Tender
PRODUCT CATEGORY	
TENDERSTATUS	Expired
TENDER_STAGE	AOC
ISITEMWISEEVALUATION	f
L1	
L2	
L3	
NUMBER OF BIDS	
TIMESTAMP	
PINCODE ID	
CLEAR	SUBMIT

## Conclusion

### About the Accuracy of the modules

The implemented Machine Learning models will perform well even for the updated database as long as data of a similar nature is added to the database and the models are retrained at regular intervals. However, if an altogether a new dataset which is not at all similar to the data present in the database is used for training the models, the models may not perform well because the project has been developed based on the structure and type of data that was available in the eProcurement database of the Government. The project can be made to work for the database of a similar structure by gradually training the model using the new data.

The results of the Data Analytics and Machine Learning models also depend on the data present in the database. Presence of validation checks in the system ensures that there are no invalid entries in the dataset and hence improves the results of the analysis. Data cleaning techniques are applied to remove erroneous/invalid and duplicate entries from the dataset before training a model. If the dataset does not have significantly high number of invalid entries, data cleaning helps to refine the dataset which leads to accurate results by the model. However, if the number of erroneous/invalid entries is high, as in case of the eProcurement database, valuable data is lost during the data cleaning phase. This loss of data resulted in lesser data and insights for the model to learn on. The data validation methods can substantially help in creating very large and authentic datasets that will result in the algorithms continuously refining and learning more in league with the human knowledge in the Audit Domain.

### Areas for further improvement

Currently, the modules use Cosine Similarity and Jaccard Similarity while comparing the texts. These methods work with mathematical transformations on the text, instead of their meaning. These methods, therefore, cannot identify synonyms and understand the text with the nuances of the English language. Natural language processing algorithms can be more effective in such cases but they could not be used for this project they require a large corpus of documents related to procurement and tendering to train the models and this was not available.

The current risk analysis works on unsupervised algorithms. With labelled data in place indicating which tenders have a high risk and are more prone to be erroneous, supervised algorithms can be used, which have the potential to learn how to categorise the tenders. These models try to mimic human behaviour and try to apply this in terms of mathematical algorithms and formulae.

### Recommendations for use in the Indian Audit & Accounts Department

The modules developed as part of this project can be used in audit for the functions described above. It is to be noted that the results given by the models are generally based on a pattern or a trend and cannot be expected to be hundred percent accurate. However, the results generated will help the auditors to identify the risk areas more efficiently and enable them to spend their time and effort during an audit in a focussed manner.

## Appendix 1 – Analytic techniques/algorithms used in the project

### K Medoids Clustering to cluster work items based on difference between L1 & L2 and L1 & L3

Clustering is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points in the same group should have similar properties and features, while data points in different groups have profoundly different properties and features.

The default metric for calculating the distance between each of the points while training the model is the Euclidean distance. In comparison, the results are useful for datasets with continuous variables, but the metric fails to consider the concept behind the categorical values. For example, while computing the distance between two data points having Contract Type value as either 1, 2 or, 3, where one is Tender, two is Rate Contract, and three is Empanelment. Euclidean distance considers the difference between 1 and 3 to be more than 1 and 2, where due to the data's categorical nature, one is as distant from two as three. Therefore, the Dice distance metric is used for computing the distances between categorical points. Since the dataset is mixed, we use a distance metric called Gower Distance. Gower metric incorporates the Dice distance for the categorical columns and Manhattan distance (absolute difference between the values) for the continuous columns.

The clustering algorithm used for this module is **K-Medoids, (also called as Partitioning Around Medoid)**. A medoid is a point in the cluster whose dissimilarities with all the other points in the group are minimum. In K-Means clustering which is commonly used method of clustering numerical data, the representative point of a cluster is the mean of all points within a cluster. The representative point of a cluster in K-Medoid clustering is a datapoint inside that cluster which is comparable to the median which is robust and preferable to use when the data is categorical in nature. Let the number of clusters be 'k.' Then the algorithm initializes 'k' random medoids out of the data-set and assigns clusters to all the points in the dataset. The cost is calculated using the dissimilarity scores between medoids and data points. After this, a random point in the cluster is chosen to be the medoid. If the cost increases this way, the swap is undone.

The optimal number of clusters is decided by calculating the algorithm's performance on different values, ideally ranging from two to eight. For this project, the inertia values took a steep decrease (Elbow Method) when the number of clusters was 6. Thus, the dataset was fitted on the K-Medoids algorithm with six number of clusters.

### Gaussian Mixture Model Clustering for checking correctness of the data

(By checking the ranges of awarded values of accepted AOC Bids)

The awarded values in the database are in a similar range. So to generate meaningful clusters, Gaussian Mixture Model Clustering is used because the clusters formed from this clustering are stretched and can be helpful in the cases where all the data points are very similar.

To find the number of clusters to be used, the Bayesian Information Criterion (BIC) Score was used; the lower the value, the simpler the model, the better the performance. Usually the BIC Score decreases with an increasing number of clusters. Therefore, for the computational purposes, it was decided to have five as the number of clusters.

Bids were clustered based on the reasonability of the bid values and probability of a datapoint (bid value) lying in each cluster is computed using sklearn libraries in-built functionality predict probability.

### K-Means Clustering for checking the ranks of accepted AOC Bids and Evaluators Value

The data values used for this model are bid ranks and evaluator's value. The evaluator's value was not recorded in a few cases, and these empty places were replaced by 0. The bid ranks contain textual data that cannot be clustered, so it was changed into numerical values. The bid ranks which were not recorded were given the rank Nil. To change the bid ranks into numerical values, the rank similarity was computed with L1, L2, L3, H1, H2, and H3. The Jaccard Similarity method was used on the ranks. The Jaccard similarity can be calculated as

$$\text{Jaccard Similarity} = \frac{A \cap B}{A \cup B}$$

Where A, B are two words, and the set operations are done on the letters of the word. The Jaccard Similarity was calculated for both L1,L2,L3, and H1, H2,H3 and the maximum of these two was considered. After finding out the percentage of similarity, if the percentage is less than 100, it is replaced by 0 to get proper clusters. The data was normalized and the model was fit based on the Jaccard Similarity and evaluators' value. K-Means clustering algorithm was selected for this model. The number of clusters for this algorithm was determined by the silhouette score, a measure of how similar an object is to its cluster compared to other groups. The cluster number was decided to be 6 based on the silhouette score and the efficiency of the model.

### Cosine Similarity Clustering for identifying the Splitting of tenders

Cosine similarity measures the similarity between two datapoints of multiple dimensions. Based on the similarity the data points are clustered such that the work items that seem very similar as per the dimensions defined while development would fall together in a cluster.

Fourteen major dimensions<sup>5</sup> were identified and the similarity between the work items was calculated based on those fourteen dimensions and the weightage assigned to each of these dimensions using the Cosine Similarity method. For the categorical data types, distance is calculated using the dice metric, where distance is considered '1' when the values are unequal and '0' otherwise. For the continuous text data types, the distance between the two texts is inversely proportional to similarity. For the continuous real values, the distance is the absolute difference between the two values divided by their range to normalize them. For the continuous DateTime values, the number of days between the days divided by the range of days from a beginning date (for this dataset, 17th June 2010).

The distance between two tenders is therefore calculated using the individual distance metrics for different types of table columns. For each column, the distance between them is calculated and multiplied with weights assigned to each column based on their impact on the splitting results and added. Using this method, for any chosen tender, the ten most similar tenders are identified using the distance calculation and exported to a CSV file. The graphs for the year-wise and month-wise distribution of these ten most similar tenders are also created. Another bar graph to plot the ten work items (work item ID) against their similarity index. The similarity index is reciprocal of 1+distance between the tenders.

---

<sup>5</sup> Appendix 2 - Dimensions identified to calculate the degree of similarity between the work items



## Appendix 2 – Dimensions identified to calculate the degree of similarity between the work items

### 1) Categorical dimensions

- i) createdby - User ID that floated the tender on the E-procurement platform
- ii) actualorgid - Organisation that floated the tender
- iii) contracttype - The type of contract, Tender, Rate Contract, or Empanelment
- iv) pincodid – Pin code of the location of the tender
- v) tendercategoryid - The superset of product categories (can be 1, 2, or 3)

### 2) Continuous (String) dimensions

- i) tendertitle - The tender title
- ii) workdesc - A brief description of the tender floated
- iii) bidopeningplace - The bid opening place
- iv) productcategory - The product category, for example, Civil Works - Roads
- v) location – Location of work
- vi) invitingofficer - The designation of the officer floating the tender
- vii) invitingofficeraddress - The office address of the officer who invited the tender

### 3) Continuous (Float/Real) dimensions

- i) tendervalue - The final value of the tender

### 4) Continuous (Datetime / Timestamp) dimensions

- i) createddate – The date when the tender was created in the eProcurement System

## Appendix 3 – Using the project

### Setting up of running environment

Following are the software requirements for running the project on a computer:

- Python programming language (including Integrated Development Environment for Python - IDLE) - The complete project has been developed in the Python programming language that is platform independent and thus can be installed on both Windows and Linux Operating Systems. Python can be downloaded from [here](#).
- Jupyter Notebook – It is an open-source web application which can be used to create and share documents that contain live code, equations, visualizations and narrative text. Jupyter Notebook can be installed in a computer having Python preinstalled. It can be installed by following the instructions given [here](#).
- The following libraries of Python which have been used in the project need to be installed in the computer:
  - Pandas
  - Numpy
  - Sklearn
  - Random
  - Datetime
  - Matplotlib
  - Gradio
  - Scipy
  - Seaborn

Libraries can be installed by running the command 'pip install <library name>' (without quotes) in the commandline interface. Detailed information on installing/upgrading libraries can be found [here](#).

### Running the project

The source code of the project is available with the Regional Training Institute, Jammu. The institute may be contacted for requesting sharing of the same. The project can then be run by following the steps given below:

- Place the source code and the tables (.csv files) in the same folder on the computer on which Python, Jupyter Notebook and the required libraries are installed.
- Open Jupyter Notebook by running the command 'jupyter notebook' (without quotes) in the commandline interface.
- The jupyter notebook being a web application, will open in the default browser.
- In the jupyter notebook, browser to the location where the source code and the csv files have been placed and open the Python code file with the extension '.ipynb'
- The code would be loaded in the jupyter notebook in some time.
- The notebook consists of code snippets, comments and visualizations of the project.
- The code snippets can be executed by selecting the code snippet and pressing 'Shift + Enter' keys on the keyboard or by clicking the 'Run' button in the interface at the top.



- The complete code in the notebook can be executed by clicking on highlighted button.



- The functions performed by the project can be accessed with the help of Gradio user interface in the notebook as shown under 'Key functions performed by the project'.

### Using different dataset in the project

The project will perform well if a different dataset which is similar in structure to the original dataset is used in the project. A different dataset can be used in the project by changing the contents of the csv files to be used in the project while keeping the structure and data types of the columns unchanged. After any change in the data is made, the complete notebook should be run again in order to retrain the models.

### Modifying the source code of the project

The source code can be easily modified as per the requirements of the user to add new functionalities or to improve the existing functionalities of project. After the modification has been made, all the succeeding code snippets must be executed again to get the desired results.