**Privacy Preserving Location Data Publishing: A Machine Learning Approach**

**1. INTRODUCTION:**

Publication of data by different organizations and institutes is crucial for open research and transparency of government agencies. Just in Australia, since 2013, over 7000 additional datasets have been published on ’data.gov.au,’ a dedicated website for the publication of datasets by the Australian government. Moreover, the new Australian government data sharing legislation encourage government agencies to publish their data, and as early as 2019, many of them will have to do so. Unfortunately, the process of data publication can be highly risky as it may disclose individuals’ sensitive information. Hence, an essential step before publishing datasets is to remove any uniquely identifiable information from them. However, such an operation is not sufficient for preserving the privacy of users. Adversaries can re-identify individuals in datasets based on common attributes called quasi-identifiers or may have prior knowledge about the trajectories traveled by the users. Such side information enables them to reveal sensitive information that can cause physical, financial, and reputational harms to people. One of the most sensitive sources of data is location trajectories or spatiotemporal trajectories. Despite numerous use cases that the publication of spatiotemporal data can provide to users and researchers, it poses a significant threat to users’ privacy. As an example, consider a person who has been using GPS navigation to travel from home to work every morning of weekdays. If an adversary has some prior knowledge about a user, such as the home address, it is possible to identify the user. Such an inference attack can compromise user privacy, such as revealing the user’s health condition and how often the user visits his/her medical specialist. Therefore, it is crucial to anonymize spatiotemporal datasets before publishing them to the public. The privacy issue gets even more severe if the adversary links identified users to other databases, such as the database of medical records. That is the very reason why nowadays most companies are reluctant to publish any spatiotemporal trajectory datasets without applying an effective privacy preserving technique. A widely accepted privacy metric for the publication of spatiotemporal datasets is k-anonymity. This metric can be summarized as ensuring that every trajectory in the published dataset is indistinguishable from at least k − 1 other trajectories. The authors in, adopted the notion of k-anonymity for spatiotemporal datasets and proposed an anonymization algorithm based on generalization. Xu et al investigated the effects of factors such as spatiotemporal resolution and the number of users released on the anonymization process focused on improving the existing clustering approaches. They proposed an anonymization scheme based on achieving kanonymity by grouping similar trajectories and removing the highly dissimilar ones. More recently, the authors developed an algorithm called k-merge to anonymize the trajectory datasets while preserving the privacy of users from probabilistic attacks. Local suppression and splitting techniques were also considered to protect privacy. However, there are three major problems with the aforementioned approaches.

• Lack of a well-defined method to cluster trajectories as there is not an easy way to measure the cost of clustering when considering the distances among trajectories rather than simply the locations.

• The existing literature focuses on pairwise sequence alignment, which results in a high amount of information loss.

• There is no unified metric to evaluate and compare the existing anonymization methods. In this paper, we address the mentioned problems by proposing an enhanced anonymization framework termed machine learning based anonymization (MLA) to preserve the privacy of users in the publication of spatiotemporal trajectory datasets. MLA consists of two interworking algorithms: clustering and alignment. We have summarized our main contributions in the following bullet points.

• By formulating the anonymization process as an optimization problem and finding an alternative representation of the system, we are able to apply machine clustering algorithms for clustering trajectories. We propose to use k 0 -means 1 algorithm for this purpose, as part of the MLA framework.

• We propose a variation of k 0 -means algorithm to preserve the privacy of users in the publication of overly sensitive spatiotemporal trajectory datasets. • We enhance the performance of sequence alignment in clusters by considering multiple sequence alignment instead of pairwise sequence alignment.

• We propose a utility metric to evaluate and compare the anonymization frameworks. MLA and all algorithms associated with it are applied on two real-life GPS datasets following different distributions in time and spatial domains. The experimental results indicate a significantly higher utility levels while maintaining k-anonymity of trajectories.

**1.1 Objective of the project:**

Publishing datasets plays an essential role in open data research and promoting transparency of government agencies. However, such data publication might reveal users’ private information. One of the most sensitive sources of data is spatiotemporal trajectory datasets. Unfortunately, merely removing unique identifiers cannot preserve the privacy of users. Adversaries may know parts of the trajectories or be able to link the published dataset to other sources for the purpose of user identification. Therefore, it is crucial to apply privacy preserving techniques before the publication of spatiotemporal trajectory datasets. In this paper, we propose a robust framework for the anonymization of spatiotemporal trajectory datasets termed as machine learning based anonymization (MLA). By introducing a new formulation of the problem, we are able to apply machine learning algorithms for clustering the trajectories and propose to use k-means algorithm for this purpose. A variation of k-means algorithm is also proposed to preserve the privacy in overly sensitive datasets. Moreover, we improve the alignment process by considering multiple sequence alignment as part of the MLA. The framework and all the proposed algorithms are applied to TDrive and Geolife location datasets. The experimental results indicate a significantly higher utility of datasets by anonymization based on MLA framework.

**2. LITERATURE SURVEY:**

**“Machine learning aided anonymization of spatiotemporal trajectory datasets,”**

The big data era requires a growing number of companies to publish their data publicly. Preserving the privacy of users while publishing these data has become a critical problem. One of the most sensitive sources of data is spatiotemporal trajectory datasets. Such datasets are extremely sensitive as users' personal information such as home address, workplace and shopping habits can be inferred from them. In this paper, we propose an approach for anonymization of spatiotemporal trajectory datasets. The proposed approach is based on generalization entailing alignment and clustering of trajectories. We propose to apply k'-means algorithm for clustering trajectories by developing a technique that makes it possible. We also significantly reduce the information loss during the alignment by incorporating multiple sequence alignment instead of pairwise sequence alignment used in the literature. We analyze the performance of our proposed approach by applying it to Geolife dataset, which includes GPS logs of over 180 users in Beijing, China. Our experiments indicate the robustness of our framework compared to prior works.

**“Anonymization of longitudinal electronic medical records,”**

Electronic medical record (EMR) systems have enabled healthcare providers to collect detailed patient information from the primary care domain. At the same time, longitudinal data from EMRs are increasingly combined with biorepositories to generate personalized clinical decision support protocols. Emerging policies encourage investigators to disseminate such data in a deidentified form for reuse and collaboration, but organizations are hesitant to do so because they fear such actions will jeopardize patient privacy. In particular, there are concerns that residual demographic and clinical features could be exploited for reidentification purposes. Various approaches have been developed to anonymize clinical data, but they neglect temporal information and are, thus, insufficient for emerging biomedical research paradigms. This paper proposes a novel approach to share patient-specific longitudinal data that offers robust privacy guarantees, while preserving data utility for many biomedical investigations. Our approach aggregates temporal and diagnostic information using heuristics inspired from sequence alignment and clustering methods. We demonstrate that the proposed approach can generate anonymized data that permit effective biomedical analysis using several patient cohorts derived from the EMR system of the Vanderbilt University Medical Center.

**“Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data,”**

Human mobility data has been ubiquitously collected through cellular networks and mobile applications, and publicly released for academic research and commercial purposes for the last decade. Since releasing individual's mobility records usually gives rise to privacy issues, datasets owners tend to only publish aggregated mobility data, such as the number of users covered by a cellular tower at a specific timestamp, which is believed to be sufficient for preserving users' privacy. However, in this paper, we argue and prove that even publishing aggregated mobility data could lead to privacy breach in individuals' trajectories. We develop an attack system that is able to exploit the uniqueness and regularity of human mobility to recover individual's trajectories from the aggregated mobility data without any prior knowledge. By conducting experiments on two real-world datasets collected from both mobile application and cellular network, we reveal that the attack system is able to recover users' trajectories with accuracy about 73%~91% at the scale of tens of thousands to hundreds of thousands users, which indicates severe privacy leakage in such datasets. Through the investigation on aggregated mobility data, our work recognizes a novel privacy problem in publishing statistic data, which appeals for immediate attentions from both academy and industry.

**“Novel privacy-preserving algorithm based on frequent path for trajectory data publishing,”**

Existing location-based services have collected a large amount of location data, which contain users’ personal information and has serious personal [privacy leakage](https://www.sciencedirect.com/topics/computer-science/privacy-leakage) threats. Therefore, the preservation of individual privacy when publishing data is receiving increasing attention. Most existing methods of preserving user privacy suffer a serious loss in data usability, resulting in low usability of data. In this paper, we address this problem and present TOPF, a novel approach for preserving privacy in [trajectory data](https://www.sciencedirect.com/topics/computer-science/trajectory-data) publishing based on frequent path. TOPF aims to achieve better quality of [trajectory data](https://www.sciencedirect.com/topics/computer-science/trajectory-data) for publishing and strike a balance between the conflicting goals of data usability and data privacy. To the best of our knowledge, this is the first paper that uses frequent path to preserve data privacy. First, infrequent roads in each trajectory are removed, and a new way is adopted to divide trajectories into candidate groups. A new method for finding the most frequent path is then proposed, and then, the representative trajectory is selected to represent all trajectories within a group. Experimental results show that our algorithm not only effectively guarantees the privacy of the user but also ensures the high usability of the data.

**“Towards privacy-preserving publishing of spatiotemporal trajectory data,”**

Mobile network operators can track subscribers via passive or active monitoring of device locations. The recorded trajectories offer an unprecedented outlook on the activities of large user populations, which enables developing new networking solutions and services, and scaling up studies across research disciplines. Yet, the disclosure of individual trajectories raises significant privacy concerns: thus, these data are often protected by restrictive non-disclosure agreements that limit their availability and impede potential usages. In this paper, we contribute to the development of technical solutions to the problem of privacy-preserving publishing of spatiotemporal trajectories of mobile subscribers. We propose an algorithm that generalizes the data so that they satisfy kτ,ϵ-anonymity, an original privacy criterion that thwarts attacks on trajectories. Evaluations with real-world datasets demonstrate that our algorithm attains its objective while retaining a substantial level of accuracy in the data. Our work is a step forward in the direction of open, privacy-preserving datasets of spatiotemporal trajectories.

**“Local suppression and splitting techniques for privacy preserving publication of trajectories,”**

We study the problem of preserving user privacy in the publication of location sequences. Consider a database of trajectories, corresponding to movements of people, captured by their transactions when they use credit cards, RFID debit cards, or NFC (http://en.wikipedia.org/wiki/Near\_field\_communication) compliant devices. We show that, if such trajectories are published exactly (by only hiding the identities of persons that followed them), one can use partial trajectory knowledge as a quasi-identifier for the remaining locations in the sequence. We devise four intuitive techniques, based on combinations of locations suppression and trajectories splitting, and we show that they can prevent privacy breaches while keeping published data accurate for aggregate query answering and frequent subsets data mining.

**“Towards trajectory anonymization: a generalization-based approach,”**

Trajectory datasets are becoming popular due to the massive usage of GPS and locationbased services. In this paper, we address privacy issues regarding the identification of individuals in static trajectory datasets. We first adopt the notion of k-anonymity to trajectories and propose a novel generalization-based approach for anonymization of trajectories. We further show that releasing anonymized trajectories may still have some privacy leaks. Therefore we propose a randomization based reconstruction algorithm for releasing anonymized trajectory data and also present how the underlying techniques can be adapted to other anonymity standards. The experimental results on real and synthetic trajectory datasets show the effectiveness of the proposed techniques

**“Traffic information publication with privacy preservation,”**

We are experiencing the expanding use of location-based services such as AT&T’s TeleNav GPS Navigator and Intel’s Thing Finder. Existing location-based services have collected a large amount of location data, which has great potential for statistical usage in applications like traffic flow analysis, infrastructure planning, and advertisement dissemination. The key challenge is how to wisely use the data without violating each user’s location privacy concerns. In this article, we first identify a new privacy problem, namely, the *inference-route* problem, and then present our anonymization algorithms for privacy-preserving trajectory publishing. The experimental results have demonstrated that our approach outperforms the latest related work in terms of both efficiency and effectiveness.

**“Anonymizing moving objects: How to hide a mob in a crowd?”**

Moving object databases (MOD) have gained much interest in recent years due to the advances in mobile communications and positioning technologies. Study of MOD can reveal useful information (e.g., traffic patterns and congestion trends) that can be used in applications for the common benefit. In order to mine and/or analyze the data, MOD must be published, which can pose a threat to the location privacy of a user. Indeed, based on prior knowledge of a user's location at several time points, an attacker can potentially associate that user to a specific moving object (MOB) in the published database and learn her position information at other time points.

In this paper, we study the problem of privacy-preserving publishing of moving object database. Unlike in microdata, we argue that in MOD, there does not exist a fixed set of quasi-identifier (*QID*) attributes for all the MOBs. Consequently the anonymization groups of MOBs (i.e., the sets of other MOBs within which to hide) may not be disjoint. Thus, there may exist MOBs that can be identified explicitly by combining different anonymization groups. We illustrate the pitfalls of simple adaptations of classical *k*-anonymity and develop a notion which we prove is robust against privacy attacks. We propose two approaches, namely *extreme-union* and *symmetric anonymization*, to build anonymization groups that provably satisfy our proposed *k*-anonymity requirement, as well as yield low information loss. We ran an extensive set of experiments on large real-world and synthetic datasets of vehicular traffic. Our results demonstrate the effectiveness of our approach.

**“Location privacy and its applications: A systematic study,”**

This paper surveys the current research status of location privacy issues in mobile applications. The survey spans five aspects of study: the definition of location privacy, attacks and adversaries, mechanisms to preserve the privacy of locations, location privacy metrics, and the current status of location-based applications. Through this comprehensive review, all the interrelated aspects of location privacy are integrated into a unified framework. Additionally, the current research progress in each area is reviewed individually, and the links between existing academic research and its practical applications are identified. This in-depth analysis of the current state-of-play in location privacy is designed to provide a solid foundation for future studies in the field.

**“Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints,”**

Publishing data about patients that contain both demographics and diagnosis codes is essential to perform large-scale, low-cost medical studies. However, preserving the privacy and utility of such data is challenging, because it requires: (i) guarding against identity disclosure (re-identification) attacks based on both demographics and diagnosis codes, (ii) ensuring that the anonymized data remain useful in intended analysis tasks, and (iii) minimizing the information loss, incurred by anonymization, to preserve the utility of general analysis tasks that are difficult to determine before data publishing. Existing anonymization approaches are not suitable for being used in this setting, because they cannot satisfy all three requirements. Therefore, in this work, we propose a new approach to deal with this problem. We enforce the requirement (i) by applying (k,km)-anonymity, a privacy principle that prevents re-identification from attackers who know the demographics of a patient and up to m of their diagnosis codes, where k and m are tunable parameters. To capture the requirement (ii), we propose the concept of utility constraint for both demographics and diagnosis codes. Utility constraints limit the amount of generalization and are specified by data owners (e.g., the healthcare institution that performs anonymization). We also capture requirement (iii), by employing well-established information loss measures for demographics and for diagnosis codes. To realize our approach, we develop an algorithm that enforces (k,km)-anonymity on a dataset containing both demographics and diagnosis codes, in a way that satisfies the specified utility constraints and with minimal information loss, according to the measures. Our experiments with a large dataset containing more than 200,000 electronic health records show the effectiveness and efficiency of our algorithm.

**3. SYSTEM ANALYSIS**

**3.1 Existing System**

Publication of data by different organizations and institutes is crucial for open research and transparency of government agencies. Just in Australia, since 2013, over 7000 additional datasets have been published on ’data.gov.au,’ a dedicated website for the publication of datasets by the Australian government. Moreover, the new Australian government data sharing legislation encourage government agencies to publish their data, and as early as 2019, many of them will have to do so. Unfortunately, the process of data publication can be highly risky as it may disclose individuals’ sensitive information

**Disadvantages of Existing System:**

* Less Accuracy

**3.2 Proposed System**

Now-a-days due to mobile all online applications are recording user locations and then storing them in their app and this location details they can use to track users. Sometime some malicious users can track the user location by knowing their home address and then they can match the home location with other location details to know where user is travelling like bank, hospital or any other locations. To overcome from this problem author has introduce Machine Learning based data privacy preserving technique

**Advantages of Proposed System:**

* More Accuracy

**Modules Information:**

To implement this project we have used Taxi Trajectory (TDRIVE) dataset and below is the dataset details

1. Clustering model: in this model user locations will be clusters by using KMEANS algorithm and then calculate loss value. Loss value indicates difference between correct value and predicted value and the lesser the loss the better is the algorithm. The loss value will be saved to compare with Dynamic Sequence Alignment Loss and this Dynamic Sequence is called as Heuristic Clustering Algorithm.
2. Dynamic Sequence Alignment: In this module or algorithm we will take location form cluster member and then take random locations from original dataset and both this records will be aligned to get location which has minimal loss.
3. Data Generalization: in this module user location will be generalized or anonymised by summing up location with loss values.

**FUNCTIONAL REQUIREMENTS:**

**SOFTWARE REQIREMENTS:**

**System Attributes:**

1. filename
2. dataset
3. train
4. cluster\_labels
5. kmeans\_loss, heuristic\_loss, spectral\_loss

**Data base Requirements:**

No need

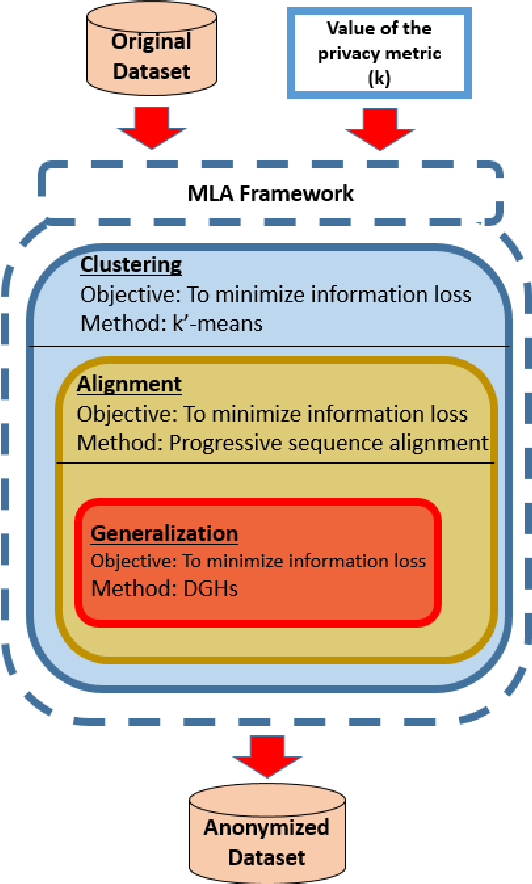
**USECASE:**

Use cases - Use cases describe the interaction between the system and external users that leads to achieving particular goals.

1. Upload Taxi Trajectory Dataset
2. Preprocess Dataset
3. Run KMeans with DynamicSA Algorithm
4. Run Data Generalization Algorithm
5. Loss Comparison Graph

**User Stories:** In this work, the authors target grouping the trajectories based on their similarity and choose a cluster head for each cluster to represent the cluster. Also, dummy trajectories were added to anonymize the datasets further

**Work down Structure:**



**Prototype:**

python 3.7.0 or 3.7.4

opencv-python==4.5.1.48

keras==2.3.1

tensorflow==1.14.0

protobuf==3.16.0

h5py==2.10.0

sklearn-extensions==0.0.2

scikit-learn==0.22.2.post1

Numpy

Pandas

**Models and Diagrams:**



**NON-FUNCTIONAL REQUIREMENT:**

**Usability:**  Usability is a quality attribute that assesses how easy user interfaces are to use. The word "usability" also refers to methods for improving ease-of-use during the design process.(how it was handle entire project easy)

**Security:** the quality or state of being secure: such as. a : freedom from danger : safety. b : freedom from fear or anxiety. c : freedom from the prospect of being laid off job security.

**Readability:** Readability is the ease with which a reader can understand a written text.

**Performance**: the execution of an action. : something accomplished : deed, feat. : the fulfillment of a claim, promise, or request : implementation. 3. : the action of representing a character in a play.

**Availability**: the quality or state of being available trying to improve the availability of affordable housing. 2 : an available person or thing.

**Scalability**: Scalability is the measure of a system's ability to increase or decrease in performance and cost in response to changes in application and system processing demands.

**3.3. PROCESS MODEL USED WITH JUSTIFICATION**

**SDLC (Umbrella Model):**

**Umbrella Activity**

**Umbrella Activity**

**Umbrella Activity**

1. Feasibility Study
2. TEAM FORMATION
3. Project Specification PREPARATION

Business Requirement Documentation

ANALYSIS & DESIGN

CODE

UNIT TEST

DOCUMENT CONTROL

ASSESSMENT

TRAINING

INTEGRATION & SYSTEM TESTING

DELIVERY/INSTALLATION

ACCEPTANCE TEST

Requirements Gathering

SDLC is nothing but Software Development Life Cycle. It is a standard which is used by software industry to develop good software.

**Stages in SDLC:**

* Requirement Gathering
* Analysis
* Designing
* Coding
* Testing
* Maintenance

**Requirements Gathering stage:**

The requirements gathering process takes as its input the goals identified in the high-level requirements section of the project plan. Each goal will be refined into a set of one or more requirements. These requirements define the major functions of the intended application, define operational data areas and reference data areas, and define the initial data entities. Major functions include critical processes to be managed, as well as mission critical inputs, outputs and reports. A user class hierarchy is developed and associated with these major functions, data areas, and data entities. Each of these definitions is termed a Requirement. Requirements are identified by unique requirement identifiers and, at minimum, contain a requirement title and textual description.



These requirements are fully described in the primary deliverables for this stage: the Requirements Document and the Requirements Traceability Matrix (RTM). The requirements document contains complete descriptions of each requirement, including diagrams and references to external documents as necessary. Note that detailed listings of database tables and fields are *not* included in the requirements document.

The title of each requirement is also placed into the first version of the RTM, along with the title of each goal from the project plan. The purpose of the RTM is to show that the product components developed during each stage of the software development lifecycle are formally connected to the components developed in prior stages.

In the requirements stage, the RTM consists of a list of high-level requirements, or goals, by title, with a listing of associated requirements for each goal, listed by requirement title. In this hierarchical listing, the RTM shows that each requirement developed during this stage is formally linked to a specific product goal. In this format, each requirement can be traced to a specific product goal, hence the term requirements traceability.

The outputs of the requirements definition stage include the requirements document, the RTM, and an updated project plan.

* Feasibility study is all about identification of problems in a project.
* No. of staff required to handle a project is represented as Team Formation, in this case only modules are individual tasks will be assigned to employees who are working for that project.
* Project Specifications are all about representing of various possible inputs submitting to the server and corresponding outputs along with reports maintained by administrator.

**Analysis Stage:**

The planning stage establishes a bird's eye view of the intended software product, and uses this to establish the basic project structure, evaluate feasibility and risks associated with the project, and describe appropriate management and technical approaches.



The most critical section of the project plan is a listing of high-level product requirements, also referred to as goals. All of the software product requirements to be developed during the requirements definition stage flow from one or more of these goals. The minimum information for each goal consists of a title and textual description, although additional information and references to external documents may be included. The outputs of the project planning stage are the configuration management plan, the quality assurance plan, and the project plan and schedule, with a detailed listing of scheduled activities for the upcoming Requirements stage, and high level estimates of effort for the out stages.

**Designing Stage:**

The design stage takes as its initial input the requirements identified in the approved requirements document. For each requirement, a set of one or more design elements will be produced as a result of interviews, workshops, and/or prototype efforts. Design elements describe the desired software features in detail, and generally include functional hierarchy diagrams, screen layout diagrams, tables of business rules, business process diagrams, pseudo code, and a complete entity-relationship diagram with a full data dictionary. These design elements are intended to describe the software in sufficient detail that skilled programmers may develop the software with minimal additional input.

  
When the design document is finalized and accepted, the RTM is updated to show that each design element is formally associated with a specific requirement. The outputs of the design stage are the design document, an updated RTM, and an updated project plan.

**Development (Coding) Stage:**

The development stage takes as its primary input the design elements described in the approved design document. For each design element, a set of one or more software artifacts will be produced. Software artifacts include but are not limited to menus, dialogs, and data management forms, data reporting formats, and specialized procedures and functions. Appropriate test cases will be developed for each set of functionally related software artifacts, and an online help system will be developed to guide users in their interactions with the software.



The RTM will be updated to show that each developed artifact is linked to a specific design element, and that each developed artifact has one or more corresponding test case items. At this point, the RTM is in its final configuration. The outputs of the development stage include a fully functional set of software that satisfies the requirements and design elements previously documented, an online help system that describes the operation of the software, an implementation map that identifies the primary code entry points for all major system functions, a test plan that describes the test cases to be used to validate the correctness and completeness of the software, an updated RTM, and an updated project plan.

**Integration & Test Stage:**

During the integration and test stage, the software artifacts, online help, and test data are migrated from the development environment to a separate test environment. At this point, all test cases are run to verify the correctness and completeness of the software. Successful execution of the test suite confirms a robust and complete migration capability. During this stage, reference data is finalized for production use and production users are identified and linked to their appropriate roles. The final reference data (or links to reference data source files) and production user list are compiled into the Production Initiation Plan.



The outputs of the integration and test stage include an integrated set of software, an online help system, an implementation map, a production initiation plan that describes reference data and production users, an acceptance plan which contains the final suite of test cases, and an updated project plan.

* **Installation & Acceptance Test:**

During the installation and acceptance stage, the software artifacts, online help, and initial production data are loa ded onto the production server. At this point, all test cases are run to verify the correctness and completeness of the software. Successful execution of the test suite is a prerequisite to acceptance of the software by the customer.

After customer personnel have verified that the initial production data load is correct and the test suite has been executed with satisfactory results, the customer formally accepts the delivery of the software.



The primary outputs of the installation and acceptance stage include a production application, a completed acceptance test suite, and a memorandum of customer acceptance of the software. Finally, the PDR enters the last of the actual labor data into the project schedule and locks the project as a permanent project record. At this point the PDR "locks" the project by archiving all software items, the implementation map, the source code, and the documentation for future reference.

**Maintenance:**

Outer rectangle represents maintenance of a project, Maintenance team will start with requirement study, understanding of documentation later employees will be assigned work and they will undergo training on that particular assigned category. For this life cycle there is no end, it will be continued so on like an umbrella (no ending point to umbrella sticks).

**3.4. Software Requirement Specification**

**3.4.1. Overall Description**

A Software Requirements Specification (SRS) – a [requirements specification](http://en.wikipedia.org/wiki/Requirements_specification) for a [software system](http://en.wikipedia.org/wiki/Software_system) is a complete description of the behavior of a system to be developed. It includes a set of [use cases](http://en.wikipedia.org/wiki/Use_case) that describe all the interactions the users will have with the software. In addition to use cases, the SRS also contains non-functional requirements. [Nonfunctional requirements](http://en.wikipedia.org/wiki/Non-functional_requirements) are requirements which impose constraints on the design or implementation (such as [performance engineering](http://en.wikipedia.org/wiki/Performance_engineering) requirements, [quality](http://en.wikipedia.org/wiki/Quality_%28business%29) standards, or design constraints).

System requirements specification: A structured collection of information that embodies the requirements of a system. A [business analyst](http://en.wikipedia.org/wiki/Business_analyst), sometimes titled [system analyst](http://en.wikipedia.org/wiki/System_analyst), is responsible for analyzing the business needs of their clients and stakeholders to help identify business problems and propose solutions. Within the [systems development lifecycle](http://en.wikipedia.org/wiki/Systems_development_life_cycle) domain, the BA typically performs a liaison function between the business side of an enterprise and the information technology department or external service providers. Projects are subject to three sorts of requirements:

* [Business requirements](http://en.wikipedia.org/wiki/Business_requirements) describe in business terms what must be delivered or accomplished to provide value.
* Product requirements describe properties of a system or product (which could be one of several ways to accomplish a set of business requirements.)
* Process requirements describe activities performed by the developing organization. For instance, process requirements could specify .Preliminary investigation examine project feasibility, the likelihood the system will be useful to the organization. The main objective of the feasibility study is to test the Technical, Operational and Economical feasibility for adding new modules and debugging old running system. All system is feasible if they are unlimited resources and infinite time. There are aspects in the feasibility study portion of the preliminary investigation:
* **ECONOMIC FEASIBILITY**

A system can be developed technically and that will be used if installed must still be a good investment for the organization. In the economical feasibility, the development cost in creating the system is evaluated against the ultimate benefit derived from the new systems. Financial benefits must equal or exceed the costs. The system is economically feasible. It does not require any addition hardware or software. Since the interface for this system is developed using the existing resources and technologies available at NIC, There is nominal expenditure and economical feasibility for certain.

* **Operational Feasibility**

Proposed projects are beneficial only if they can be turned out into information system. That will meet the organization’s operating requirements. Operational feasibility aspects of the project are to be taken as an important part of the project implementation. This system is targeted to be in accordance with the above-mentioned issues. Beforehand, the management issues and user requirements have been taken into consideration. So there is no question of resistance from the users that can undermine the possible application benefits. The well-planned design would ensure the optimal utilization of the computer resources and would help in the improvement of performance status.

* **TECHNICAL FEASIBILITY**

Earlier no system existed to cater to the needs of ‘Secure Infrastructure Implementation System’. The current system developed is technically feasible. It is a web based user interface for audit workflow at NIC-CSD. Thus it provides an easy access to .the users. The database’s purpose is to create, establish and maintain a workflow among various entities in order to facilitate all concerned users in their various capacities or roles. Permission to the users would be granted based on the roles specified. Therefore, it provides the technical guarantee of accuracy, reliability and security.

**3.4.2. External Interface Requirements**

**User Interface**

The user interface of this system is a user friendly python Graphical User Interface.

**Hardware Interfaces**

The interaction between the user and the console is achieved through python capabilities.

**Software Interfaces**

The required software is python.

**SYSTEM REQUIREMENT:**

**HARDWARE REQUIREMENTS:**

# Processor - Intel i3(min)

* Speed - 1.1 GHz
* RAM - 4GB(min)
* Hard Disk - 500 GB
* Key Board - Standard Windows Keyboard
* Mouse - Two or Three Button Mouse
* Monitor - SVGA

**SOFTWARE REQUIREMENTS:**

* Operating System - Windows10(min)
* Programming Language - Python

**4. SYSTEM DESIGN**

**CLASS DIAGRAM:**

The class diagram is the main building block of object oriented modeling. It is used both for general conceptual modeling of the systematic of the application, and for detailed modeling translating the models into programming code. Class diagrams can also be used for data modeling. The classes in a class diagram represent both the main objects, interactions in the application and the classes to be programmed. In the diagram, classes are represented with boxes which contain three parts:

* The upper part holds the name of the class
* The middle part contains the attributes of the class
* The bottom part gives the methods or operations the class can take or undertake

****

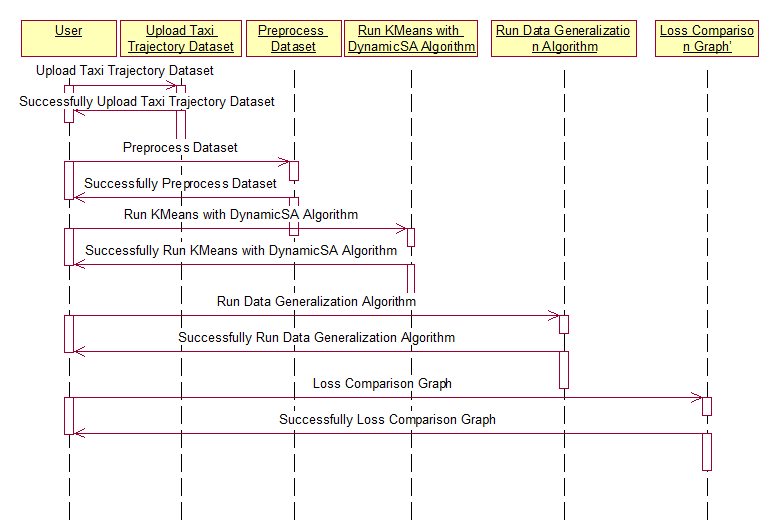
**USECASE DIAGRAM:**

A **use case diagram** at its simplest is a representation of a user's interaction with the system and depicting the specifications of a use case. A use case diagram can portray the different types of users of a system and the various ways that they interact with the system. This type of diagram is typically used in conjunction with the textual use case and will often be accompanied by other types of diagrams as we



**SEQUENCE DIAGRAM**

A **sequence diagram** is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called **event diagrams**, **event scenarios**, and timing diagrams.



**COLLABORATION DIAGRAM:**

A collaboration diagram describes interactions among objects in terms of sequenced messages. Collaboration diagrams represent a combination of information taken from class, sequence, and use case diagrams describing both the static structure and dynamic behavior of a system.



**COMPONENT DIAGRAM:**

In the Unified Modelling Language, a component diagram depicts how components are wired together to form larger components and or software systems. They are used to illustrate the structure of arbitrarily complex systems.

Components are wired together by using an assembly connector to connect the required interface of one component with the provided interface of another component. This illustrates the service consumer - service provider relationship between the two components.



**DEPLOYMENT DIAGRAM:**

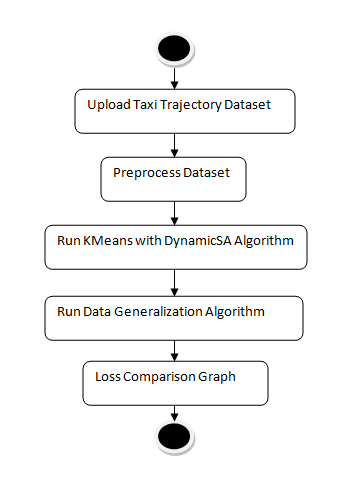
A **deployment diagram** in the Unified Modeling Language models the *physical* deployment of artifacts on nodes. To describe a web site, for example, a deployment diagram would show what hardware components ("nodes") exist (e.g., a web server, an application server, and a database server), what software components ("artifacts") run on each node (e.g., web application, database), and how the different pieces are connected (e.g. JDBC, REST, RMI).

The nodes appear as boxes, and the artifacts allocated to each node appear as rectangles within the boxes. Nodes may have sub nodes, which appear as nested boxes. A single node in a deployment diagram may conceptually represent multiple physical nodes, such as a cluster of database servers.



**ACTIVITY DIAGRAM:**

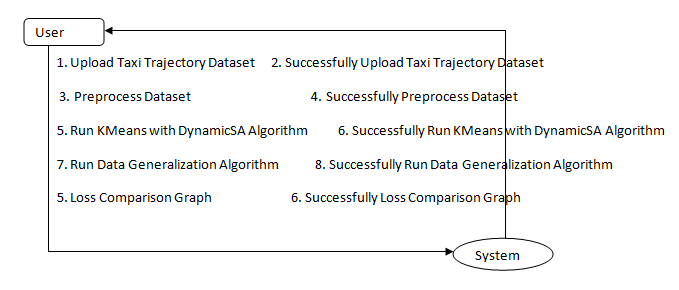
Activity diagram is another important diagram in UML to describe dynamic aspects of the system. It is basically a flow chart to represent the flow form one activity to another activity. The activity can be described as an operation of the system. So the control flow is drawn from one operation to another. This flow can be sequential, branched or concurrent



**Data flow :**

Data flow diagrams illustrate how data is processed by a system in terms of inputs and outputs. Data flow diagrams can be used to provide a clear representation of any business function. The technique starts with an overall picture of the business and continues by analyzing each of the functional areas of interest. This analysis can be carried out in precisely the level of detail required. The technique exploits a method called top-down expansion to conduct the analysis in a targeted way.

As the name suggests, Data Flow Diagram (DFD) is an illustration that explicates the passage of information in a process. A DFD can be easily drawn using simple symbols. Additionally, complicated processes can be easily automated by creating DFDs using easy-to-use, free downloadable diagramming tools. A DFD is a model for constructing and analyzing information processes. DFD illustrates the flow of information in a process depending upon the inputs and outputs. A DFD can also be referred to as a Process Model. A DFD demonstrates business or technical process with the support of the outside data saved, plus the data flowing from the process to another and the end results.



**5. IMPLEMETATION**

**5.1 Python**

Python is a general-purpose language. It has wide range of applications from Web development (like: Django and Bottle), scientific and mathematical computing (Orange, SymPy, NumPy) to desktop graphical user Interfaces (Pygame, Panda3D). The syntax of the language is clean and length of the code is relatively short. It's fun to work in Python because it allows you to think about the problem rather than focusing on the syntax.

**History of Python:**

Python is a fairly old language created by Guido Van Rossum. The design began in the late 1980s and was first released in February 1991.

**Why Python was created?**

In late 1980s, Guido Van Rossum was working on the Amoeba distributed operating system group. He wanted to use an interpreted language like ABC (ABC has simple easy-to-understand syntax) that could access the Amoeba system calls. So, he decided to create a language that was extensible. This led to design of a new language which was later named Python.

**Why the name Python?**

No. It wasn't named after a dangerous snake. Rossum was fan of a comedy series from late seventies. The name "Python" was adopted from the same series "Monty Python's Flying Circus".

**Features of Python:**

**A simple language which is easier to learn**

Python has a very simple and elegant syntax. It's much easier to read and write Python programs compared to other languages like: C++, Java, C#. Python makes programming fun and allows you to focus on the solution rather than syntax.

If you are a newbie, it's a great choice to start your journey with Python.

**Free and open-source**

You can freely use and distribute Python, even for commercial use. Not only can you use and distribute software’s written in it, you can even make changes to the Python's source code.

Python has a large community constantly improving it in each iteration.

**Portability**

You can move Python programs from one platform to another, and run it without any changes.

It runs seamlessly on almost all platforms including Windows, Mac OS X and Linux.

**Extensible and Embeddable**

Suppose an application requires high performance. You can easily combine pieces of C/C++ or other languages with Python code.

This will give your application high performance as well as scripting capabilities which other languages may not provide out of the box.

**A high-level, interpreted language**

Unlike C/C++, you don't have to worry about daunting tasks like memory management, garbage collection and so on.

Likewise, when you run Python code, it automatically converts your code to the language your computer understands. You don't need to worry about any lower-level operations.

**Large standard libraries to solve common tasks**

Python has a number of standard libraries which makes life of a programmer much easier since you don't have to write all the code yourself. For example: Need to connect MySQL database on a Web server? You can use MySQLdb library using import MySQLdb .

Standard libraries in Python are well tested and used by hundreds of people. So you can be sure that it won't break your application.

**Object-oriented**

Everything in Python is an object. Object oriented programming (OOP) helps you solve a complex problem intuitively.

With OOP, you are able to divide these complex problems into smaller sets by creating objects.

**Applications of Python:**

**1. Simple Elegant Syntax**

Programming in Python is fun. It's easier to understand and write Python code. Why? The syntax feels natural. Take this source code for an example:

a = 2

b = 3

sum = a + b

print(sum)

**2. Not overly strict**

You don't need to define the type of a variable in Python. Also, it's not necessary to add semicolon at the end of the statement.

Python enforces you to follow good practices (like proper indentation). These small things can make learning much easier for beginners.

**3. Expressiveness of the language**

Python allows you to write programs having greater functionality with fewer lines of code. Here's a link to the source code of Tic-tac-toe game with a graphical interface and a smart computer opponent in less than 500 lines of code. This is just an example. You will be amazed how much you can do with Python once you learn the basics.

**4. Great Community and Support**

Python has a large supporting community. There are numerous active forums online which can be handy if you are stuck.

**5.2 Sample Code:**

from tkinter import messagebox

from tkinter import \*

from tkinter import simpledialog

import tkinter

import numpy as np

from tkinter import simpledialog

from tkinter import filedialog

from Bio import pairwise2

from Bio.Seq import Seq

import numpy as np

import pandas as pd

from Bio.Seq import Seq

from sklearn.cluster import KMeans

from sklearn.metrics import accuracy\_score

import random

import matplotlib.pyplot as plt

from sklearn import cluster

main = tkinter.Tk()

main.title("Privacy Preserving Location Data Publishing: A Machine Learning Approach")

main.geometry("1300x1200")

global filename

global dataset

global train

global cluster\_labels

global kmeans\_loss, heuristic\_loss, spectral\_loss

trajectory\_append = []

store\_loss = []

sa\_correct = 0

def uploadDataset():

global filename

global dataset

text.delete('1.0', END)

filename = filedialog.askopenfilename(initialdir="TaxiDataset")

pathlabel.config(text=str(filename)+" loaded")

dataset = pd.read\_csv(filename,nrows=100)

dataset['querydate']= pd.to\_datetime(dataset['querydate'])

text.insert(END,str(dataset.head()))

def processDataset():

global train

global dataset

text.delete('1.0', END)

dataset.fillna(0, inplace = True)

train = dataset[['latitude','longitude']]

text.insert(END,"Total records contains in dataset : "+str(train.shape[0])+"\n")

text.insert(END,"\ndataset preprocessing completed\n")

def dynamicSA(src\_lat,src\_lon,cls\_id):

global sa\_correct

dups = []

max1 = 0

max2 = 0

choosen\_lat = 0

choosen\_lon = 0

while len(dups) < 10:

random\_record\_dataset = 0

flag = True

while flag:

random\_record\_dataset = random.randint(0,(len(dataset)-1))

if random\_record\_dataset not in dups:

dups.append(random\_record\_dataset)

flag = False

des\_lat = dataset[random\_record\_dataset,2]

des\_lon = dataset[random\_record\_dataset,3]

seq1 = Seq(str(src\_lat))

seq2 = Seq(str(des\_lat))

seq3 = Seq(str(src\_lon))

seq4 = Seq(str(des\_lon))

alignments1 = pairwise2.align.globalxx(seq1, seq2)

alignments2 = pairwise2.align.globalxx(seq3, seq4)

for match in alignments1:

score = match[2]

if score > max1:

max1 = score

choosen\_lat = des\_lat

for match in alignments2:

score = match[2]

if score > max2:

max2 = score

choosen\_lon = des\_lon

cls = 0

if max1 <= 5 and max2 <= 5:

cls = 0

else:

cls = 1

if cls == cls\_id:

sa\_correct = sa\_correct + 1

print(str(sa\_correct)+" "+str(cls\_id)+" "+str(max1)+" "+str(max2))

return str(choosen\_lat)+","+str(choosen\_lon),(max1+max2)/2

def runKmeansSA():

text.delete('1.0', END)

global trajectory\_append

global store\_loss

global train

global dataset

global cluster\_labels

global kmeans\_loss

global heuristic\_loss

global sa\_correct

sa\_correct = 0

trajectory\_append.clear()

store\_loss.clear()

kmeans = KMeans(n\_clusters=2)

kmeans.fit(train)

predict = kmeans.predict(train)

cluster\_labels = kmeans.labels\_

for i in range(0,10):

predict[i] = 3

acc = accuracy\_score(cluster\_labels,predict)

kmeans\_loss = 1.0 - acc

dataset['clusterID'] = cluster\_labels

dataset = dataset.values

for i in range(len(cluster\_labels)):

src\_lat = dataset[i,2]

src\_lon = dataset[i,3]

cls\_id = dataset[i,4]

trajectory\_value, trajectory\_loss = dynamicSA(src\_lat,src\_lon,cls\_id)

trajectory\_append.append(trajectory\_value)

store\_loss.append(trajectory\_loss)

text.insert(END,"Processed Location Data : "+trajectory\_value+" with loss : "+str(trajectory\_loss)+"\n")

text.update\_idletasks()

heuristic\_loss = sa\_correct / 100.0

text.delete('1.0', END)

text.insert(END,"KMEANS Loss on Dataset : "+str(kmeans\_loss)+"\n\n")

text.insert(END,"Heuristic Loss on Dataset : "+str(heuristic\_loss)+"\n\n")

def dataGeneralization():

text.delete('1.0', END)

global trajectory\_append

global store\_loss

for i in range(len(trajectory\_append)):

arr = trajectory\_append[i].split(",")

lat = float(arr[0])

lon = float(arr[1])

lat = lat + store\_loss[i]

lon = lon + store\_loss[i]

text.insert(END,"Latitude After Generalization : "+str(lat)+" Longitude After Generalization : "+str(lon)+"\n")

def graph():

global heuristic\_loss

global kmeans\_loss

global spectral\_loss

height = [heuristic\_loss,kmeans\_loss, spectral\_loss]

bars = ('Heuristic Loss','KMeans Loss','Spectral Loss')

y\_pos = np.arange(len(bars))

plt.bar(y\_pos, height)

plt.xticks(y\_pos, bars)

plt.title("Kmeans & Heuristic Loss Comparison Graph")

plt.show()

#defining spectral clusteering function

def runBisectKMeans():

global spectral\_loss

#creating spectral clustering object with number of clusters as 2 and gamma iteration to filter dataset is 5000

spectral = cluster.SpectralClustering(n\_clusters=2,random\_state=1,affinity='rbf',gamma=5000)

#below line starts clustering by using given 'train' dataset

spectral.fit(train)

#calculate cluster label for each cluster record

cluster\_labels = spectral.labels\_

#perform prediction on test data

predict = spectral.fit\_predict(train)

for i in range(0,5):

predict[i] = 3

#calculate correctly predicted accuracy between predicted cluster and trained cluster

acc = accuracy\_score(cluster\_labels,predict)

#calculate loss of spectral clustering

spectral\_loss = 1.0 - acc

for i in range(len(cluster\_labels)):

src\_lat = dataset[i,2]

src\_lon = dataset[i,3]

cls\_id = dataset[i,4]

trajectory\_value, trajectory\_loss = dynamicSA(src\_lat,src\_lon,cls\_id)

trajectory\_append.append(trajectory\_value)

store\_loss.append(trajectory\_loss)

text.insert(END,"Processed Location Data : "+trajectory\_value+" with loss : "+str(trajectory\_loss)+"\n")

text.update\_idletasks()

heuristic\_loss = sa\_correct / 100.0

text.delete('1.0', END)

text.insert(END,"KMEANS Loss on Dataset : "+str(kmeans\_loss)+"\n\n")

text.insert(END,"Heuristic Loss on Dataset : "+str(heuristic\_loss)+"\n\n")

text.insert(END,"Spectral Clustering Loss on Dataset : "+str(spectral\_loss)+"\n\n")

font = ('times', 16, 'bold')

title = Label(main, text='Privacy Preserving Location Data Publishing: A Machine Learning Approach',anchor=W, justify=LEFT)

title.config(bg='black', fg='white')

title.config(font=font)

title.config(height=3, width=120)

title.place(x=0,y=5)

font1 = ('times', 13, 'bold')

uploadButton = Button(main, text="Upload Taxi Trajectory Dataset", command=uploadDataset)

uploadButton.place(x=50,y=100)

uploadButton.config(font=font1)

pathlabel = Label(main)

pathlabel.config(bg='DarkOrange1', fg='white')

pathlabel.config(font=font1)

pathlabel.place(x=50,y=150)

processButton = Button(main, text="Preprocess Dataset", command=processDataset)

processButton.place(x=50,y=200)

processButton.config(font=font1)

kmeansButton = Button(main, text="Run KMeans with DynamicSA Algorithm", command=runKmeansSA)

kmeansButton.place(x=50,y=250)

kmeansButton.config(font=font1)

bisectkmeansButton = Button(main, text="Run Spectral Clustering with DynamicSA Algorithm", command=runBisectKMeans)

bisectkmeansButton.place(x=50,y=300)

bisectkmeansButton.config(font=font1)

generalButton = Button(main, text="Run Data Generalization Algorithm", command=dataGeneralization)

generalButton.place(x=50,y=350)

generalButton.config(font=font1)

graphButton = Button(main, text="Loss Comparison Graph", command=graph)

graphButton.place(x=50,y=400)

graphButton.config(font=font1)

font1 = ('times', 12, 'bold')

text=Text(main,height=28,width=100)

scroll=Scrollbar(text)

text.configure(yscrollcommand=scroll.set)

text.place(x=480,y=100)

text.config(font=font1)

main.config(bg='DarkOrange1')

main.mainloop()

**6. TESTING:**

**Implementation and Testing:**

Implementation is one of the most important tasks in project is the phase in which one has to be cautions because all the efforts undertaken during the project will be very interactive. Implementation is the most crucial stage in achieving successful system and giving the users confidence that the new system is workable and effective. Each program is tested individually at the time of development using the sample data and has verified that these programs link together in the way specified in the program specification. The computer system and its environment are tested to the satisfaction of the user.

## Implementation

## The implementation phase is less creative than system design. It is primarily concerned with user training, and file conversion. The system may be requiring extensive user training. The initial parameters of the system should be modifies as a result of a programming. A simple operating procedure is provided so that the user can understand the different functions clearly and quickly. The different reports can be obtained either on the inkjet or dot matrix printer, which is available at the disposal of the user. The proposed system is very easy to implement. In general implementation is used to mean the process of converting a new or revised system design into an operational one.

## Testing

Testing is the process where the test data is prepared and is used for testing the modules individually and later the validation given for the fields. Then the system testing takes place which makes sure that all components of the system property functions as a unit. The test data should be chosen such that it passed through all possible condition. Actually testing is the state of implementation which aimed at ensuring that the system works accurately and efficiently before the actual operation commence. The following is the description of the testing strategies, which were carried out during the testing period.

### System Testing

Testing has become an integral part of any system or project especially in the field of information technology. The importance of testing is a method of justifying, if one is ready to move further, be it to be check if one is capable to with stand the rigors of a particular situation cannot be underplayed and that is why testing before development is so critical. When the software is developed before it is given to user to use the software must be tested whether it is solving the purpose for which it is developed. This testing involves various types through which one can ensure the software is reliable. The program was tested logically and pattern of execution of the program for a set of data are repeated. Thus the code was exhaustively checked for all possible correct data and the outcomes were also checked.

**Module Testing**

To locate errors, each module is tested individually. This enables us to detect error and correct it without affecting any other modules. Whenever the program is not satisfying the required function, it must be corrected to get the required result. Thus all the modules are individually tested from bottom up starting with the smallest and lowest modules and proceeding to the next level. Each module in the system is tested separately. For example the job classification module is tested separately. This module is tested with different job and its approximate execution time and the result of the test is compared with the results that are prepared manually. The comparison shows that the results proposed system works efficiently than the existing system. Each module in the system is tested separately. In this system the resource classification and job scheduling modules are tested separately and their corresponding results are obtained which reduces the process waiting time.

**Integration Testing**

After the module testing, the integration testing is applied. When linking the modules there may be chance for errors to occur, these errors are corrected by using this testing. In this system all modules are connected and tested. The testing results are very correct. Thus the mapping of jobs with resources is done correctly by the system.

**Acceptance Testing**

When that user fined no major problems with its accuracy, the system passers through a final acceptance test. This test confirms that the system needs the original goals, objectives and requirements established during analysis without actual execution which elimination wastage of time and money acceptance tests on the shoulders of users and management, it is finally acceptable and ready for the operation.

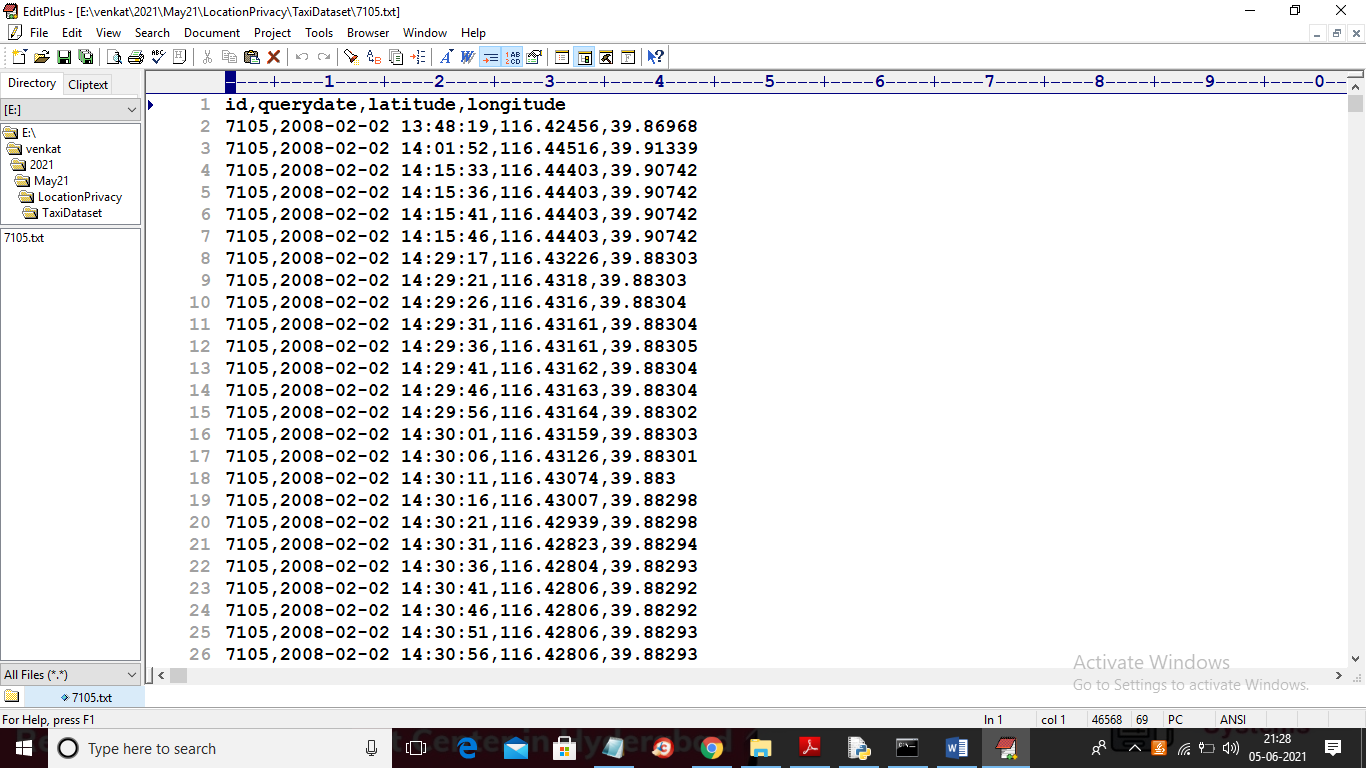
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Test Case Id** | **Test Case Name** | **Test Case Desc.** | **Test Steps** | | | | **Test Case Status** | **Test Priority** |
| **Step** | **Expected** | | **Actual** |
| 01 | Upload Taxi Trajectory Dataset | Verify Upload Taxi Trajectory Dataset  or not | If Taxi Trajectory Dataset  may not upload | we cannot do any further operations | we can do further operations | | High | High |
| 02 | Dataset Preprocessing | Verify Dataset Preprocessing or not | If Dataset Preprocessing may not upload | we cannot do any further operations | we can do further operations | | High | High |
| 03 | Run KMeans with DynamicSA Algorithm | Verify Run KMeans with DynamicSA Algorithm  or not | If KMeans with DynamicSA Algorithm  May not be Run | we cannot do any further operations | we can do further operations | | High | High |
| 04 | Run Data Generalization Algorithm | Verify Run Data Generalization Algorithm  or not | If Data Generalization Algorithm  May not Run | We cannot run  operation | We can Run the Operation | | High | High |
| 05 | Loss Comparison Graph | Verify Loss Comparison Graph  or not | If Loss Comparison Graph  May not be plot | we cannot do any further operations | we can do further operations | | High | High |

**7. SCREENSHOTS:**

Privacy Preserving Location Data Publishing: A Machine Learning Approach

Now-a-days due to mobile all online applications are recording user locations and then storing them in their app and this location details they can use to track users. Sometime some malicious users can track the user location by knowing their home address and then they can match the home location with other location details to know where user is travelling like bank, hospital or any other locations. To overcome from this problem and to provide security to user location many data anonymization techniques such as K-Anonymity and data perturbation was introduce where data perturbation will add noise to user data and K-Anonymity will adjust data into groups so user location cannot be identified.

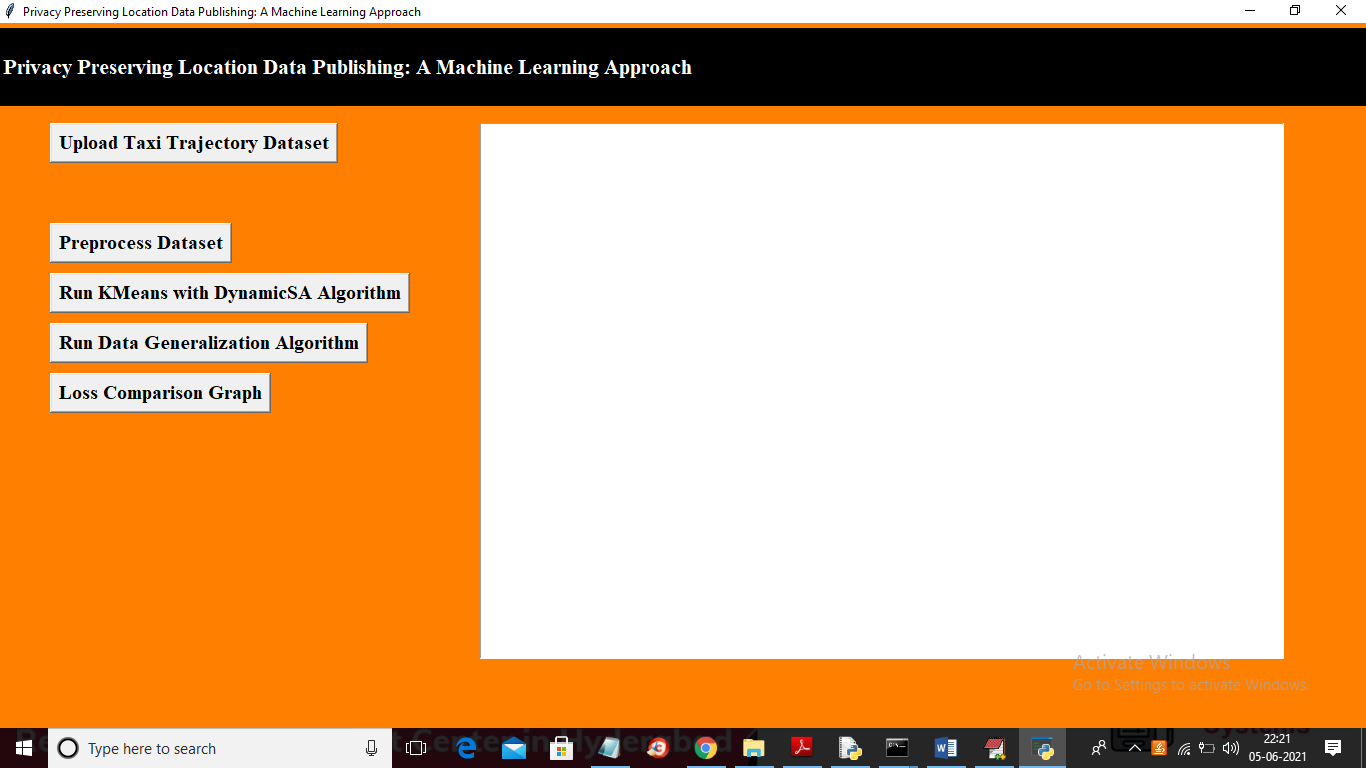
But above techniques are not reliable as malicious users can identify how to crack groups and noise data to know user location. To overcome from this problem author has introduce Machine Learning based data privacy preserving technique which consists of 3 models and this 3 models will provide more security and anonymize or generalized which cannot be easily understand or crack.



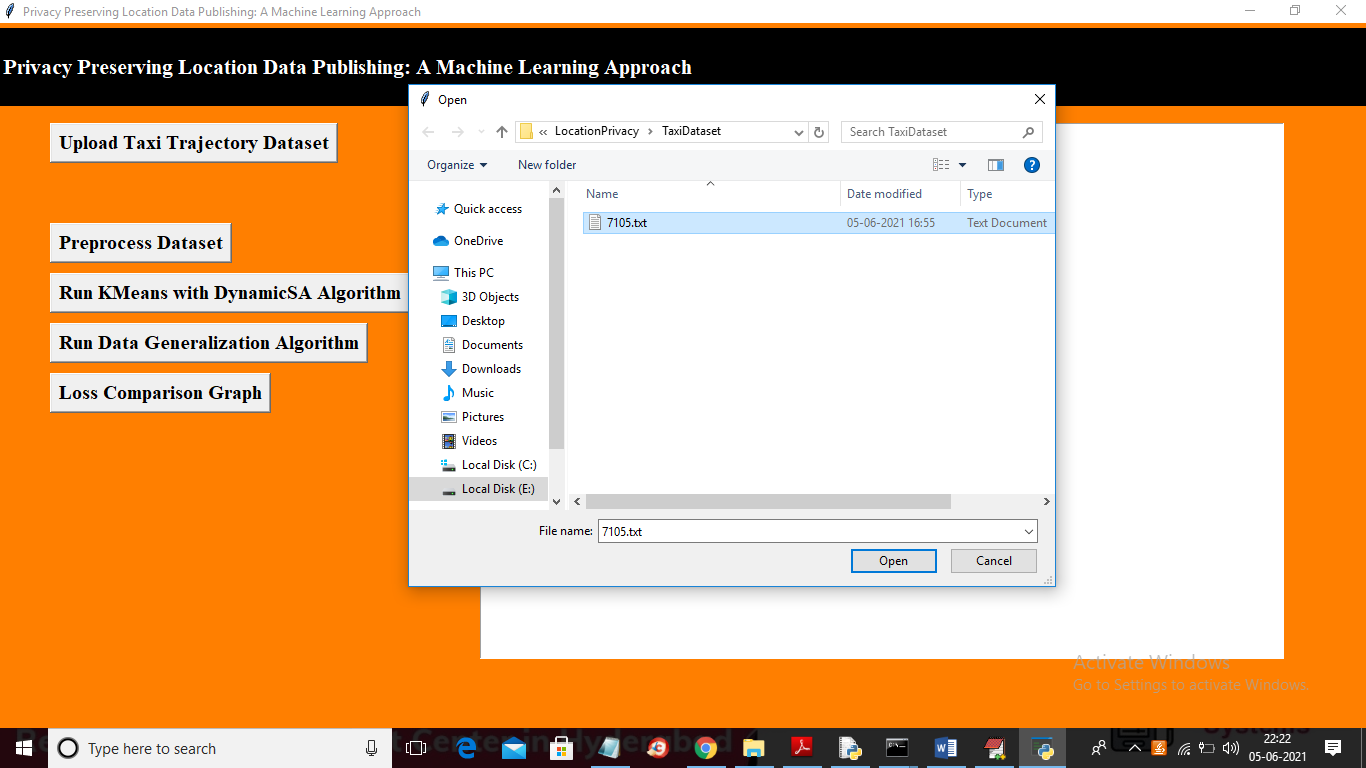
In above screen first rows contains dataset column names and remaining rows contains dataset values and in each row represents users current location latitude and longitude values.

SCREEN SHOTS

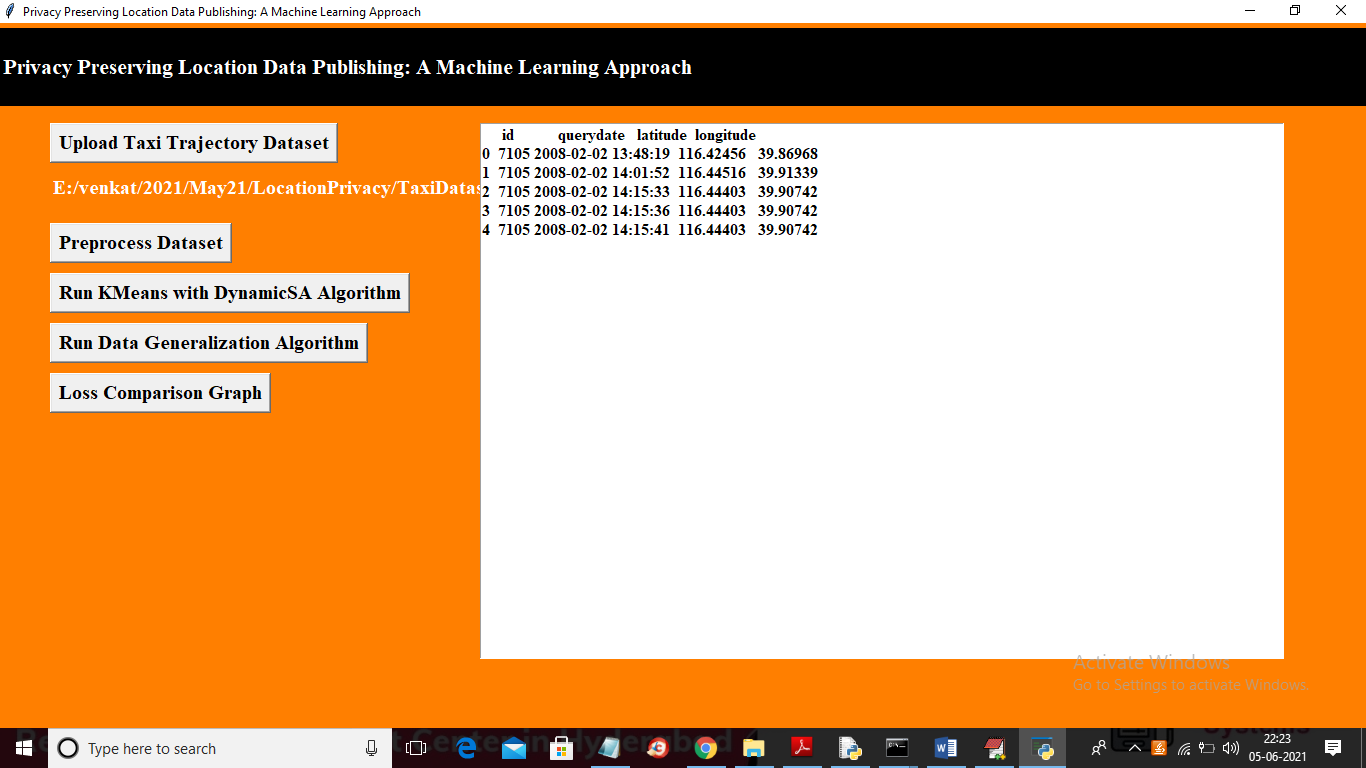
To run project double click on ‘run.bat’ file to get below screen



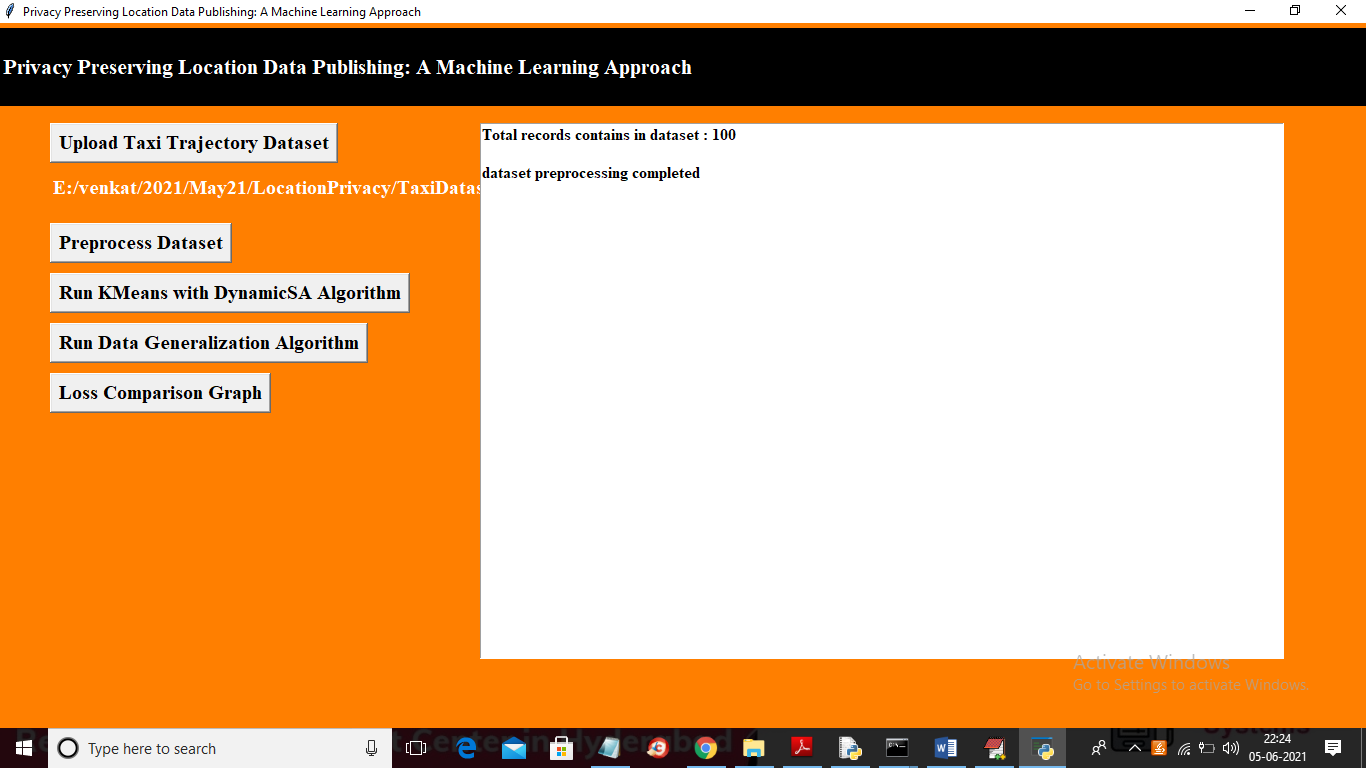
In above screen click on ‘Upload Taxi Trajectory Dataset’ button to upload dataset



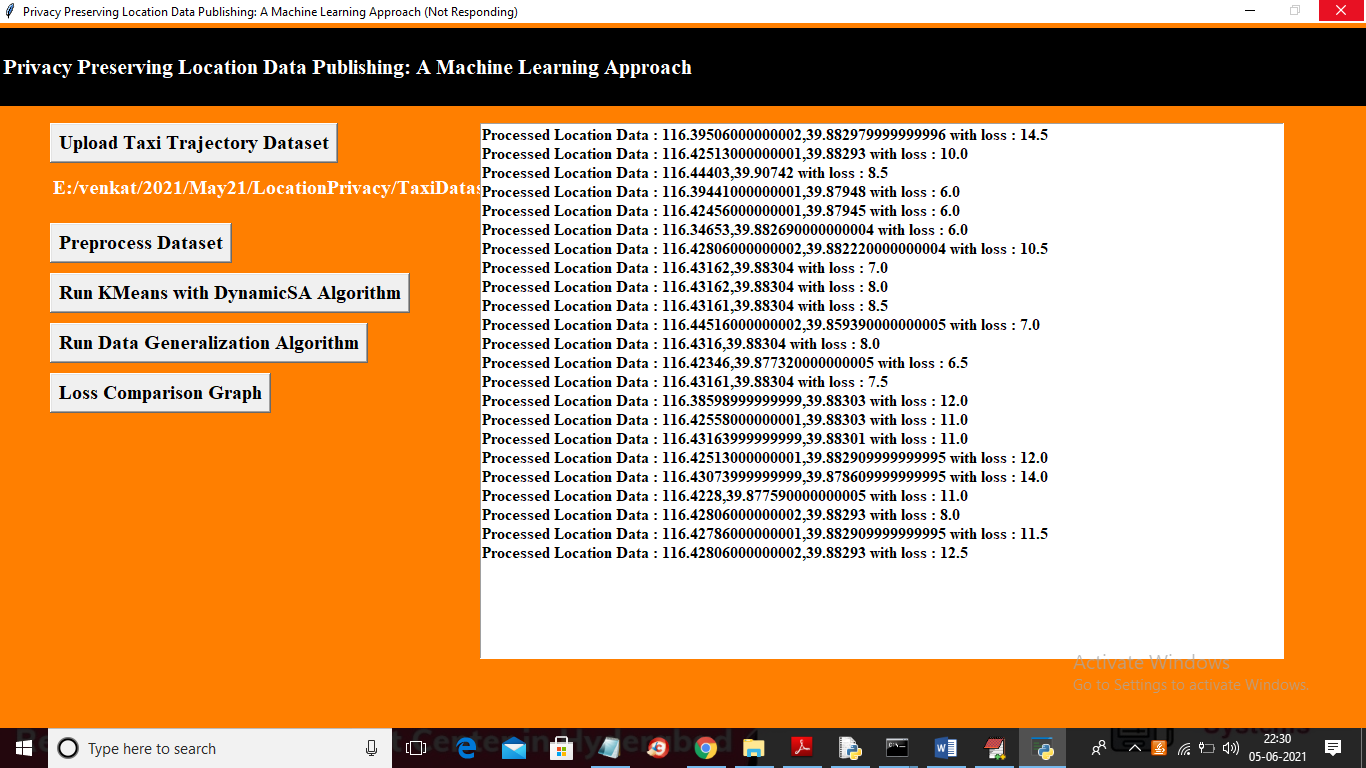
In above screen selecting and uploading taxi trajectory file and then click on ‘Open’ button to load dataset and to get below screen



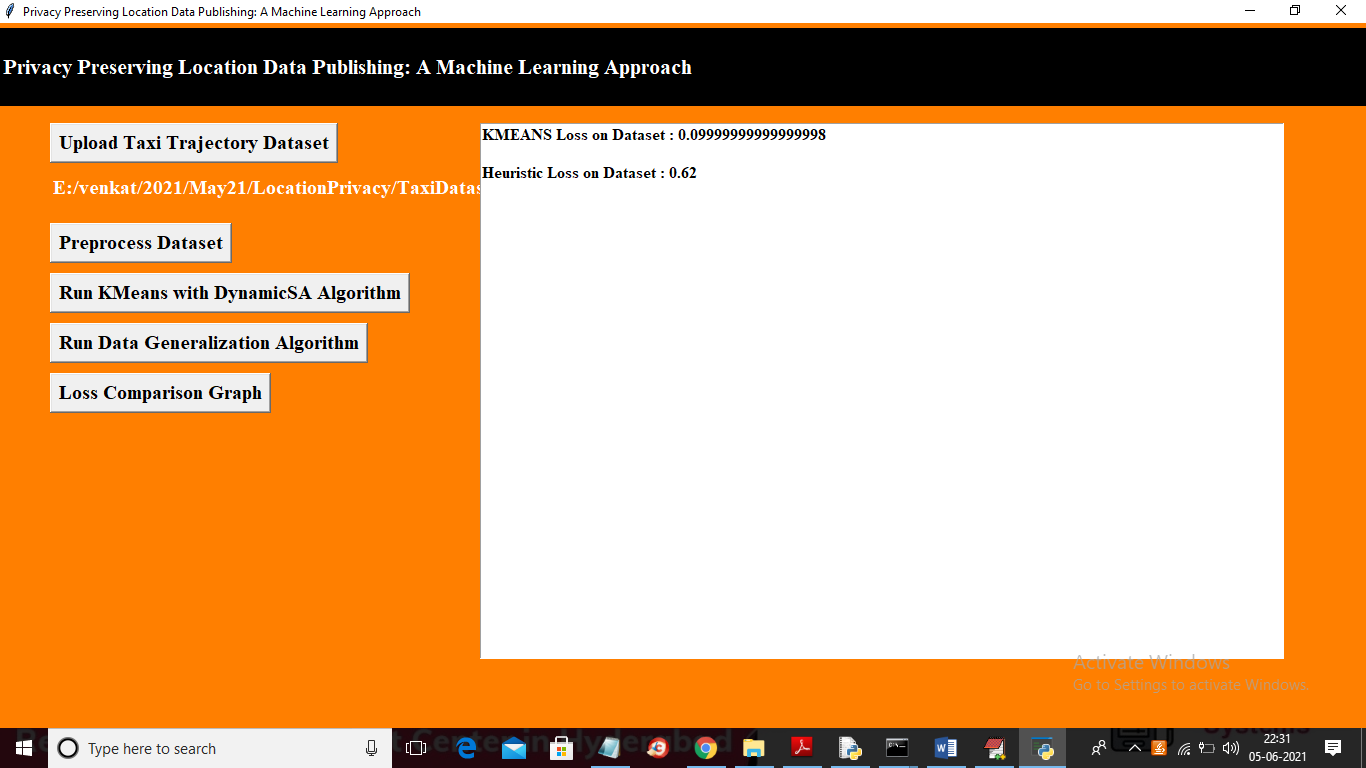
In above screen dataset loaded and now click on ‘Preprocess Dataset’ button to remove empty values and then extract latitude and longitude location from above dataset



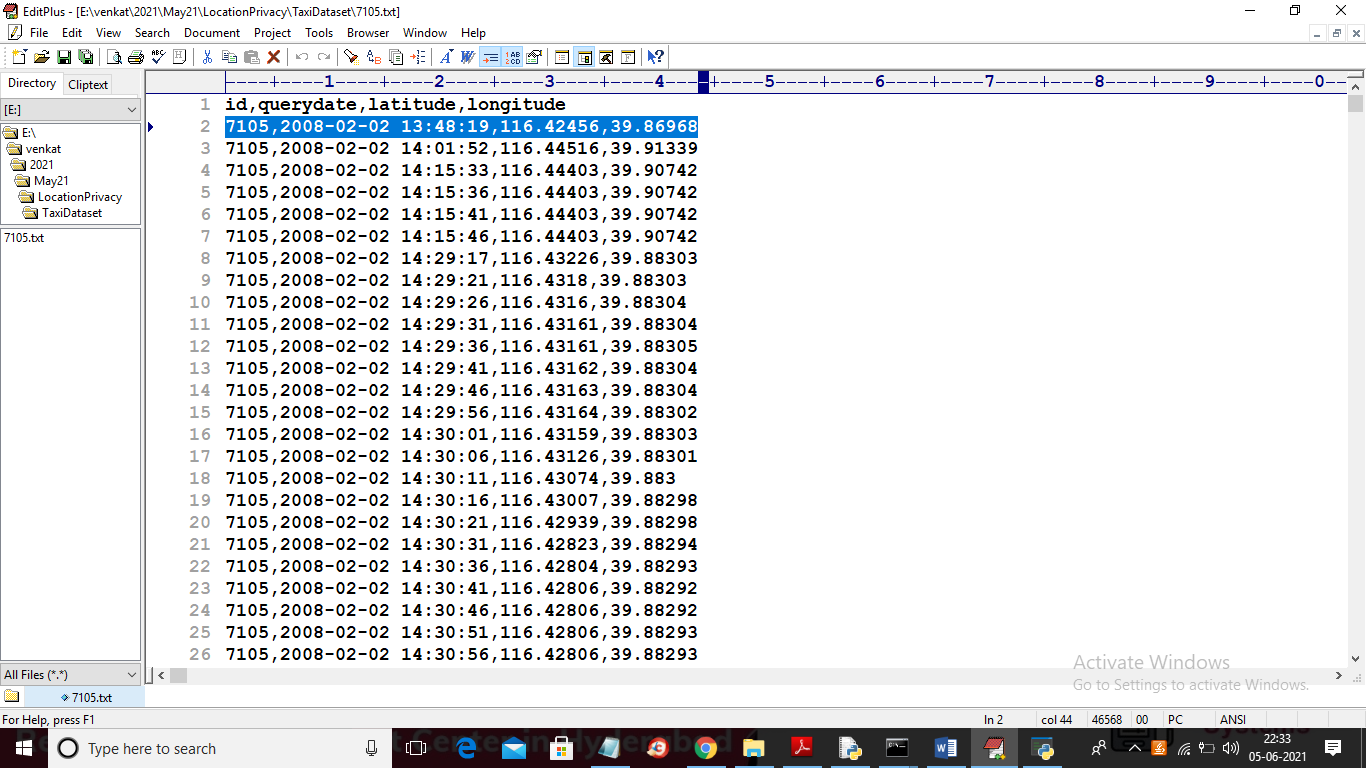
In above screen dataset preprocessing completed and now click on ‘Run KMeans with DynamicSA Algorithm’ button to run KMEANS on dataset with Dynamic SA. This algorithm will group all similar location into same cluster and then perform DYNAMIC SA.



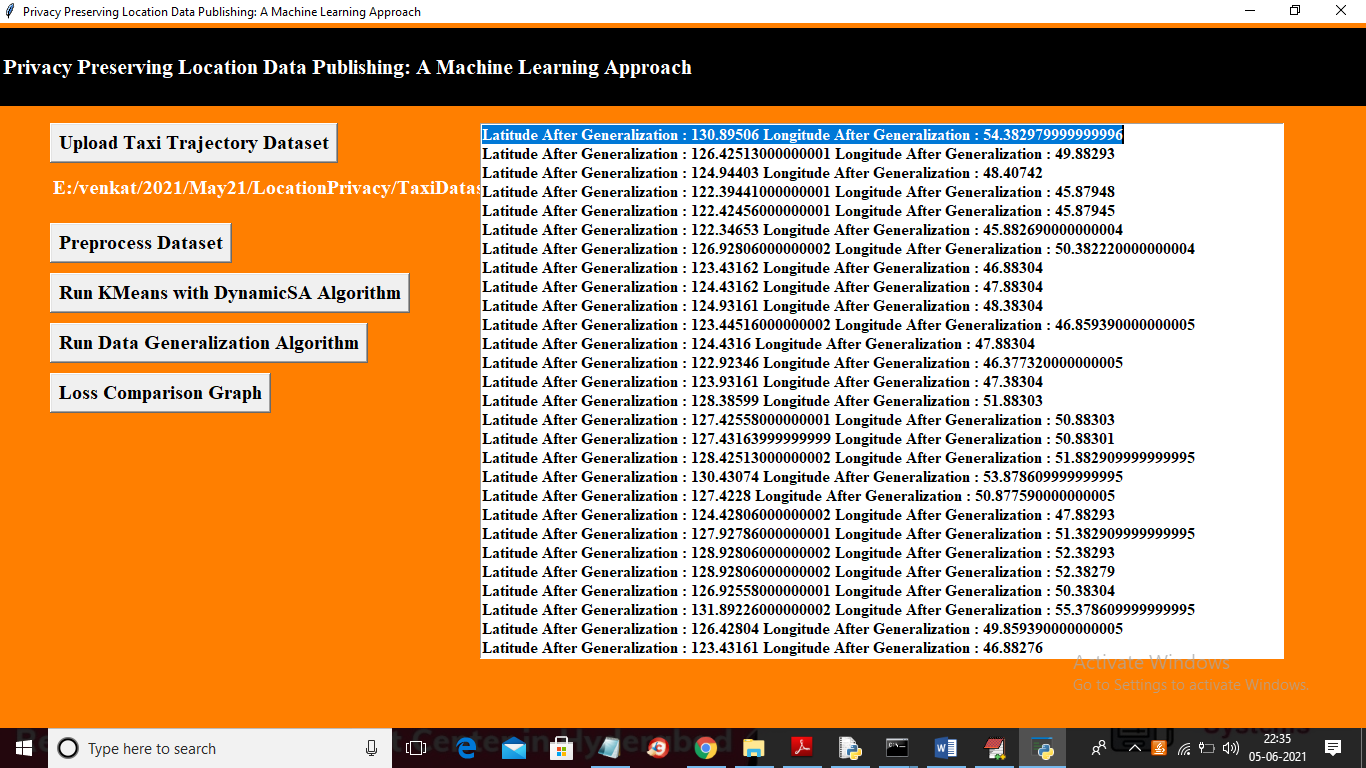
In above screen each location is processed and then calculating loss value with dynamic sequence alignment which align two locations by choosing minimal loss location.



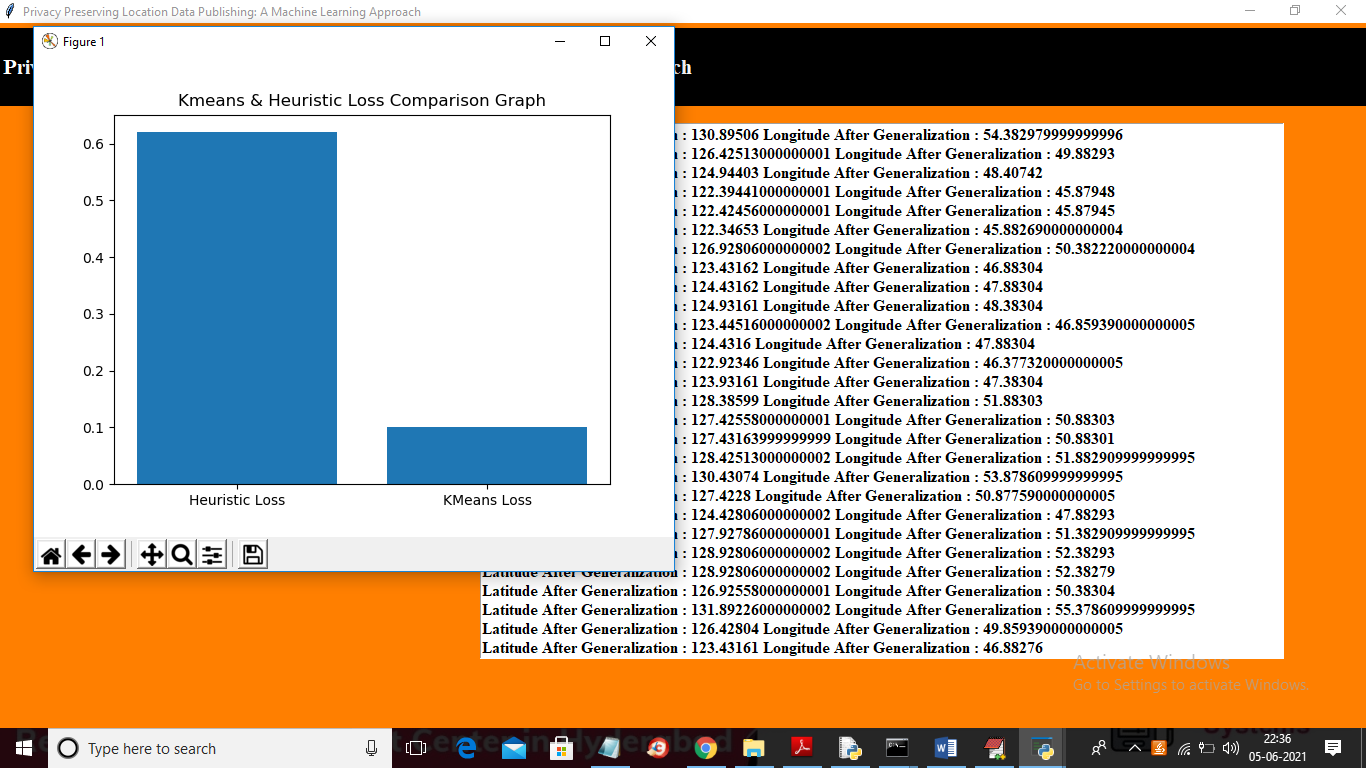
In above screen KMEANS loss is 0.09 and Heuristic Clustering (also known as Dynamic SA) loss is 0.62. Now click on ‘Run Data Generalization Algorithm’ button to generalized data with loss value. In below screen in first record you can see real location values from dataset and in next screen same location was generalized or anonymised with above algorithms



In below screen you can see the same location is generalized with other values



In above screen you can all location values are generalized so no malicious users can understand correct location. Now click on ‘Loss Comparison Graph’ button to get below graph



In above graph x-axis represents algorithm name and y-axis represents loss values generated for that algorithm and in above graph KMEANS got less loss so KMEANS is better in anonymization.

**8. CONCLUSION:**

In this paper, we have proposed a framework to preserve the privacy of users while publishing the spatiotemporal trajectories. The proposed approach is based on an efficient alignment technique termed as progressive sequence alignment in addition to a machine learning clustering approach that aims at minimizing the incurred loss in the anonymization process. We also devised a variation of k 0 -means algorithm for guaranteeing the k-anonymity in overly sensitive datasets. The experimental results on real-life GPS datasets indicate the superior utility performance of our proposed framework compared with the previous works.

**9. REFERENCES:**

[1] S. Shaham, M. Ding, B. Liu, Z. Lin, and J. Li, “Machine learning aided anonymization of spatiotemporal trajectory datasets,” arXiv preprint arXiv:1902.08934, 2019.

[2] A. Government, “New australian government data sharing and release legislation,” 2018.

[3] A. Tamersoy, G. Loukides, M. E. Nergiz, Y. Saygin, and B. Malin, “Anonymization of longitudinal electronic medical records,” IEEE Transactions on Information Technology in Biomedicine, vol. 16, no. 3, pp. 413–423, 2012.

[4] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin, “Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data,” in Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017, pp. 1241–1250.

[5] Y. Dong and D. Pi, “Novel privacy-preserving algorithm based on frequent path for trajectory data publishing,” Knowledge-Based Systems, vol. 148, pp. 55–65, 2018.

[6] M. Gramaglia, M. Fiore, A. Tarable, and A. Banchs, “Towards privacy-preserving publishing of spatiotemporal trajectory data,” arXiv preprint arXiv:1701.02243, 2017.

[7] M. Terrovitis, G. Poulis, N. Mamoulis, and S. Skiadopoulos, “Local suppression and splitting techniques for privacy preserving publication of trajectories,” IEEE Trans. Knowl. Data Eng, vol. 29, no. 7, pp. 1466–1479, 2017.

[8] M. E. Nergiz, M. Atzori, and Y. Saygin, “Towards trajectory anonymization: a generalization-based approach,” in Proc. of the SIGSPATIAL ACM GIS. ACM, 2008, pp. 52–61.

[9] S. Gurung, D. Lin, W. Jiang, A. Hurson, and R. Zhang, “Traffic information publication with privacy preservation,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 5, no. 3, p. 44, 2014. [10] R. Yarovoy, F. Bonchi, L. V. Lakshmanan, and W. H. Wang, “Anonymizing moving objects: How to hide a mob in a crowd?” in Proc. of the 12th International Conference on Extending Database Technology: Advances in Database Technology. ACM, 2009, pp. 72– 83.

[11] B. Liu, W. Zhou, T. Zhu, L. Gao, and Y. Xiang, “Location privacy and its applications: A systematic study,” IEEE Access, vol. 6, pp. 17 606–17 624, 2018.

[12] G. Poulis, G. Loukides, S. Skiadopoulos, and A. GkoulalasDivanis, “Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints,” Journal of biomedical informatics, vol. 65, pp. 76–96, 2017.

[13] T. Takahashi and S. Miyakawa, “Cmoa: Continuous moving object anonymization,” in Proceedings of the 16th International Database Engineering & Applications Sysmposium. ACM, 2012, pp. 81–90.

[14] X. Zhou and M. Qiu, “A k-anonymous full domain generalization algorithm based on heap sort,” in International Conference on Smart Computing and Communication. Springer, 2018, pp. 446–459.

[15] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Incognito: Efficient full-domain k-anonymity,” in Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, 2005, pp. 49–60.

[16] S. Yaseen, S. M. A. Abbas, A. Anjum, T. Saba, A. Khan, S. U. R. Malik, N. Ahmad, B. Shahzad, and A. K. Bashir, “Improved generalization for secure data publishing,” IEEE Access, vol. 6, pp. 27 156–27 165, 2018.

[17] G. Poulis, A. Gkoulalas-Divanis, G. Loukides, S. Skiadopoulos, and C. Tryfonopoulos, “Secreta: A tool for anonymizing relational, transaction and rt-datasets,” in Medical Data Privacy Handbook. Springer, 2015, pp. 83–109.

[18] K. Sreedhar, M. Faruk, and B. Venkateswarlu, “A genetic tds and bug with pseudo-identifier for privacy preservation over incremental data sets,” Journal of Intelligent & Fuzzy Systems, vol. 32, no. 4, pp. 2863–2873, 2017.

[19] M. E. Rana, M. Jayabalan, and M. A. Aasif, “Privacy preserving anonymization techniques for patient data: An overview,” in Third International Congress on Technology, Communication and Knowledge (ICTCK 2016), 2016.

[20] M. Jayabalan and M. E. Rana, “Anonymizing healthcare records: A study of privacy preserving data publishing techniques,” Advanced Science Letters, vol. 24, no. 3, pp. 1694–1697, 2018.

[21] D. Narula, P. Kumar, and S. Upadhyaya, “Privacy preservation using various anonymity models,” in Cyber Security: Proceedings of CSI 2015. Springer, 2018, pp. 119–130.

[22] J. Ding, “Trajectory mining, representation and privacy protection,” in Proceedings of the 2nd ACM SIGSPATIAL PhD Workshop. ACM, 2015, p. 2.

[23] A. E. Cicek, M. E. Nergiz, and Y. Saygin, “Ensuring location diversity in privacy-preserving spatio-temporal data publishing,” The VLDB JournalThe International Journal on Very Large Data Bases, vol. 23, no. 4, pp. 609–625, 2014.

[24] C. Romero-Tris and D. Meg´ıas, “Protecting privacy in trajectories with a user-centric approach,” ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 12, no. 6, p. 67, 2018.

[25] F. T. Brito, A. C. A. Neto, C. F. Costa, A. L. Mendonc¸a, and J. C. Machado, “A distributed approach for privacy preservation in the publication of trajectory data,” in Proceedings of the 2nd Workshop on Privacy in Geographic Information Collection and Analysis. ACM, 2015, p. 5.

[26] E. Naghizade, L. Kulik, and E. Tanin, “Protection of sensitive trajectory datasets through spatial and temporal exchange,” in Proc. of the 26th International Conference on Scientific and Statistical Database Management. ACM, 2014, p. 40.

[27] K. Jiang, D. Shao, S. Bressan, T. Kister, and K.-L. Tan, “Publishing trajectories with differential privacy guarantees,” in Proc. of the 25th International Conference on Scientific and Statistical Database Management. ACM, 2013, p. 12.

[28] L. Sweeney, “k-anonymity: A model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557–570, 2002.

[29] B. Chowdhury and G. Garai, “A review on multiple sequence alignment from the perspective of genetic algorithm,” Genomics, 2017.

[30] X. Chen, C. Wang, S. Tang, C. Yu, and Q. Zou, “Cmsa: a heterogeneous cpu/gpu computing system for multiple similar rna/dna sequence alignment,” BMC bioinformatics, vol. 18, no. 1, p. 315, 2017.

[31] Q. Le, F. Sievers, and D. G. Higgins, “Protein multiple sequence alignment benchmarking through secondary structure prediction,” Bioinformatics, vol. 33, no. 9, pp. 1331–1337, 2017.

[32] J. MacQueen et al., “Some methods for classification and analysis of multivariate observations,” in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[33] S. K. Pal and P. P. Wang, Genetic algorithms for pattern recognition. CRC press, 2017.

[34] A. Fischer and D. Picard, “Convergence rates for smooth k-means change-point detection,” arXiv preprint arXiv:1802.07617, 2018.

[35] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, “Mining interesting locations and travel sequences from gps trajectories,” in Proceedings of the 18th international conference on World wide web. ACM, 2009, pp. 791–800.

[36] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, “Understanding mobility based on gps data,” in Proceedings of the 10th international conference on Ubiquitous computing. ACM, 2008, pp. 312–321. [37] Y. Zheng, X. Xie, and W.-Y. Ma, “Geolife: A collaborative social networking service among user, location and trajectory.” IEEE Data Eng. Bull., vol. 33, no. 2, pp. 32–39, 2010.

[38] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, “T-drive: driving directions based on taxi trajectories,” in Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems. ACM, 2010, pp. 99–108.

[39] J. Yuan, Y. Zheng, X. Xie, and G. Sun, “Driving with knowledge from the physical world,” in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011, pp. 316–324.

[40] Y. Sun, M. Chen, L. Hu, Y. Qian, and M. M. Hassan, “Asa: Against statistical attacks for privacy-aware users in location based service,” Future Generation Computer Systems, vol. 70, pp. 48–58, 2017.

[41] O. Asan, F. Cooper II, S. Nagavally, R. J. Walker, J. S. Williams, M. N. Ozieh, and L. E. Egede, “Preferences for health information technologies among us adults: Analysis of the health information national trends survey,” Journal of medical Internet research, vol. 20, no. 10, p. e277, 2018.

[42] C. Gates, “Electronic medical record reminder to improve human papillomavirus vaccination rates among adolescents,” 2018.

[43] F. Rahman, P. Karanikas, and T. D. Giles, “Systems and methods for creating contextualized summaries of patient notes from electronic medical record systems,” Aug. 17 2017, uS Patent App. 15/430,401.

[44] J. Colleti Junior, A. B. d. Andrade, and W. B. d. Carvalho, “Evaluation of the use of electronic medical record systems in brazilian intensive care units,” Revista Brasileira de terapia intensiva, vol. 30, no. 3, pp. 338–346, 2018.