Journal of Global Information Management
Volume 31 • Issue 9

Personal Information Protection in Government Data Openness Using Decision Tree Model

Bo Gao, School of Government and Public Affairs, Communication University of China, China*
Jun Sun, School of Government and Public Affairs, Communication University of China, China
Bochun Wang, Institute of Education, University College London, UK

ABSTRACT

To address the issues of personal privacy and information security in the current digital age, this study first collects and analyzes relevant government data openness policies and personal information protection laws and regulations to understand the current policy and legal environment. Secondly, a complexity pruning decision tree model is constructed, which can identify and evaluate potential personal information protection risks in government data openness. Using the Singapore government open dataset, this decision tree model is applied for empirical analysis, and its accuracy and effectiveness are evaluated. The research results demonstrate that the complexity pruning decision tree model performs well in terms of accuracy, recall rate, F1 score, and the area under the ROC Curve (AUC). The model achieves an accuracy of 0.85 on the training and 0.8 on the test sets, indicating its high performance in personal information protection in government data openness.

KEYWORDS

decision tree model, government data openness policies, personal information protection, privacy protection

RESEARCH BACKGROUND AND MOTIVATIONS

The rapid evolution of the digital economy and informatization has ushered in an era of copious data generation, significantly impacting diverse sectors like business, healthcare, and transportation. Simultaneously, data openness has emerged as a prevailing trend, with numerous institutions and organizations sharing their data to foster information exchange and innovation (Park & Woo, 2021). Government agencies, in particular, have amassed extensive data reservoirs encompassing a wide spectrum of domains, including socio-economic statistics, public service records, environmental data, and more (Dai et al., 2020). The openness of government data holds profound implications for advancing innovation, augmenting governmental transparency, fostering public engagement, and enhancing the delivery of public services. Government data openness empowers businesses, the academic community, and the general public to harness these data stores for research and innovation, thereby propelling societal advancement (Löfgren & Webster, 2020; Desborough et al., 2020).

DOI: 10.4018/JGIM.332815 *Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

However, amid the trajectory of data openness, an increasingly conspicuous concern has emerged: the peril of personal information exposure and misappropriation. Government data repositories often harbor a wealth of personal identity information, sensitive data, and private records. Inadequate safeguarding of these data can precipitate issues such as personal information leakage, misutilization, and even identity theft (Mutimukwe et al., 2020; Hasal et al., 2021). Consequently, while championing government data openness, the imperative of securing personal information has assumed paramount importance. Traditional privacy protection methodologies, owing to the evolving sensitivity and intricacy of personal information, have grown increasingly inadequate to cater to the demands of contemporary data processing and sharing (Feng & Chen, 2022). Therefore, personal information protection approaches grounded in decision tree models have gained increasing traction and application in recent years. Decision trees, as prevalent machine learning algorithms, empower users to comprehend data and privacy protection requisites through classification analysis, predictive modeling, and data-driven decision-making processes.

This study introduces a complexity-pruned decision tree model as a robust tool for identifying and evaluating potential risks to personal information protection within the realm of government data openness. Using the Singaporean government's open dataset bolsters government transparency, granting the public deeper insights into government activities and the formulation of policies. By ensuring the security of personal information, this study strives to strike a balance between transparency and individual privacy in the context of data openness. In doing so, it aims to drive social innovation, foster sustainable development, and bolster societal progress.

Research Objectives

The primary objective of this study is to delve into the intricate realm of personal information protection within the context of government data openness, while employing a decision tree model as a pivotal analytical tool. The decision tree model offers a remarkable advantage in this domain, primarily due to its inherent interpretability. This model dissects data into a sequence of straightforward decision steps, constructing a hierarchical tree-like framework. Its structure renders the decision-making process highly intuitive and comprehensible: Each node embodies a distinct feature or attribute, and each branch represents a decision pathway.

In the realm of government data openness, the decision tree model serves as a potent instrument, elucidating clear logical pathways for policymakers, data stewards, and the general public. It facilitates an in-depth understanding of the factors that influence personal information security and privacy protection. Notably, the decision tree model adapts easily to situations where personal information constitutes a minority class amidst a sea of other data, as is often the case. It dynamically adjusts node splits in response to varying data distributions, thus enhancing the precision of decisions concerning personal information protection. This adaptability is crucial in mitigating risks such as personal information leakage or misuse.

Also, since the decision tree model automatically discerns the most salient features of data, it can effectively curtail the interference of irrelevant or redundant data in decisions pertaining to personal information protection. Within the milieu of government data openness, this capacity enables it to the identify data attributes that exert the most profound influence on individual privacy. This, in turn, facilitates the tailored formulation of data openness policies and protective measures.

Furthermore, the decision tree model is especially proficient in handling diverse data types, including numeric and categorical data. Given the eclectic nature of government data, encompassing variables like age, gender, geographical location, socio-economic statistics, and more, the decision tree model proficiently processes this heterogeneous data landscape, furnishing comprehensive analyses for personal information protection.

In addition to this, the decision tree model has expeditious training and prediction speeds, which is an invaluable asset in the context of government data openness. Government data repositories often house extensive and multifaceted information, spanning historical and real-time data. Consequently,

real-time or near real-time assessment of personal information protection risks becomes imperative. The efficiency of the decision tree model empowers swift analysis, facilitating timely data protection recommendations for decision-makers.

Ultimately, the application of the decision tree model within the purview of data openness empowers governments to navigate the delicate equilibrium between data sharing and personal privacy protection, thus propelling sustainable development within the information society. The specific goals of this study include the following:

1. Develop a deep understanding of current government policies on data openness and personal information protection laws. The study will collect and analyze these laws and regulations to understand the current legal and policy environment, thereby providing a foundation for further research.
2. Construct a decision tree model. The study will design and build a framework based on a decision tree model to identify and evaluate potential personal information protection risks in government data openness. The decision tree model is a commonly used machine learning method that can classify and determine data based on different features and conditions, so it can help elucidate the key factors of personal information protection.
3. Apply the decision tree model for empirical analysis. The study will utilize actual government datasets to perform an empirical investigation using the constructed decision tree model. Through analyzing government datasets, the model can identify key factors that may lead to personal information leakage or misuse, and it can also evaluate the accuracy and effectiveness of the decision tree model in personal information protection.

Research Framework

This study unfolds in five sections:

1. The first section is the introduction, which elucidates the study's background, motivation, objectives, and significance.
2. The second section provides a literature review, which encompasses academic research conducted in the field. This review critically examines the strengths and weaknesses of existing studies.
3. The third section introduces the process of personal information protection in government data openness using the complexity-pruned decision tree algorithm.
4. In the fourth section, the experimental design and performance evaluation are outlined. This section includes empirical assessments of model performance, including comprehensive comparisons and analyses with other methods.
5. The fifth and concluding section summarizes the study findings and points out identified limitations as well as potential avenues for future investigation.

LITERATURE REVIEW

There is an extensive body of research in the field of personal information protection in government data openness, and this showcases a rich tapestry of insights and perspectives:

- Data Governance Framework: Janssen et al. (2020) introduced a data governance framework tailored for big data algorithms. This framework streamlines the management of data, processes, and algorithms, enabling controlled data and algorithm openness for external scrutiny. It also facilitates trusted information sharing within and between organizations.
- Political Dynamics of Open Government Data: Ruijer et al. (2020) delved into the political dimensions of open government data frameworks. Their research sheds light on how open data

practices are influenced by institutional pressures, and it also examines the resultant impact on government transparency.

- Determinant Factor Model: Matheus and Janssen (2020) harnessed the determinant factor model to scrutinize public administration and information systems literature. Their work unveiled the determinant factors that either facilitate or hinder transparency and the expected outcomes associated with a government's open data initiatives.
- Challenges in Personal Health Information: Staunton et al. (2021) conducted a survey via semi-structured interviews to unearth the challenges confronting the processing of personal health information in South Africa. Their findings underscored the delicate balance between enhancing data protection and utilizing personal information for individual patient care, treatment plans, and research.
- Privacy Disasters in the Age of Big Data: Culnan (2019) illustrated common characteristics observed in privacy disasters, drawing from case studies. The research emphasized the transformative impact of big data on privacy policies.
- Legal and Ethical Challenges in Healthcare Data: Price and Cohen (2019) highlighted the legal and ethical conundrums associated with patient privacy in the context of big data. Their research underscored the importance of navigating these complexities while harnessing the potential of healthcare data.
- Regulatory Approaches and GDPR: Hoofnagle et al. (2019) delved into strategic approaches for regulating personal data, with a particular focus on the EU General Data Protection Regulation (GDPR). Their work advocated for the development of robust information governance frameworks, the promotion of internal data use, and increased human involvement in decision-making processes within organizations.
- Privacy Issues in Genetic Testing: Hendricks-Sturrup and Lu (2019) conducted research on privacy concerns in direct-to-consumer genetic testing. Their findings highlighted the persistent privacy issues in genetic testing and spurred discussions regarding consumer privacy expectations and protections.
- Information Security and Individual Behavior: Getman et al. (2020) elucidated the multifaceted nature of information security in modern society. They underscored the role of individual behavior in various contexts and under diverse pressures.

These studies collectively contribute to a nuanced understanding of the intricate landscape of personal information protection within the purview of government data openness, encompassing legal, ethical, regulatory, and practical dimensions. In further research on the subject, J. Liu et al. (2022) combined membership qualification from Hyperledger with encryption based on ciphertext policies, ensuring the security of users' private keys through membership identity in the Hyperledger network rather than traditional key distribution centers. Davidson et al. (2023) addressed societal issues through data governance and digital innovation. Islam et al. (2023) employed a structural equation modeling approach, integrating the Technology Acceptance Model and the Information System Continuance Model to investigate the continuous usage intention of open government data technology.

There has also been a significant amount of research in the subject with using machine learning concepts for information protection. Ferrag et al. (2020) combined different classifier methods based on decision trees and rule-based concepts for novel intrusion detection in IoT networks in their study on data protection using decision trees and other machine learning algorithms. Experimental results demonstrated advantages in accuracy, detection rate, false alarm rate, and time overhead compared to existing techniques. Furthermore, B. Liu et al. (2021) investigated privacy protection issues in the context of machine learning compared to traditional data privacy protection problems, emphasizing the need for comprehensive research on privacy protection and machine learning. Ramya et al. (2019) used an improved decision tree algorithm to handle interference and disruption attacks in power grids.

Shen et al. (2019) proposed a support vector machine (SVM) training scheme for privacy protection of blockchain-based encrypted IoT data, which only requires two interactions in one iteration without needing a trusted third party. Feng et al. (2020) addressed the privacy issues of models by proposing a privacy-preserving mobile prediction framework through federated learning, aiming to solve this problem without significantly sacrificing prediction performance. Y. Liu et al. (2020) proposed a privacy-preserving machine learning model called Federated Forest, which achieved the same level of accuracy as non-privacy-preserving methods. Goncalves et al. (2020) applied three categories of synthetic data generation methods, including probabilistic modeling, classification-based imputation models, and generative adversarial networks, in the context of medical privacy protection. Their study provided guidance on considerations for generating and using synthetic medical data.

Wylde et al. (2022) argued that big data, machine learning, and visualization methods and technologies could inform the level of cybersecurity and privacy-related activities, and that they could ensure secure, intelligent, and effective transparency. Several researchers addressed the computational cost, communication overhead, and privacy issues of smart transportation infrastructure using federated learning and blockchain (Aderonke et al., 2020; Huang, 2020; Javed et al., 2022; Naresh et al., 2020). Shen et al. (2022) analyzed federated learning from the perspective of data privacy and security, identifying vulnerability areas and strategies adversaries use to infiltrate the federated system.

Jupalle et al. (2022) introduced a novel technique using Ludwig classifiers to achieve human behavior automation and prediction. Vasa and Thakkar (2023) proposed various methods to protect deep learning privacy in big data analytics, presenting an effective solution to enhance privacy preservation in deep learning models. Gupta et al. (2023) introduced a data privacy protection model based on differential techniques and deep neural networks, leveraging the properties of Laplace transform to inject noise at appropriate locations for maintaining accuracy levels.

The aforementioned studies provide a systematic data governance framework for managing and protecting personal information in government data. They emphasize the institutional pressures and political context of open data, which impact the protection of personal information in government data openness. These studies offer valuable insights into the practical issues of protecting personal information in government data openness.

However, there are still some unsolved challenges in this field. For instance, one obstacle is integrating this framework with decision tree models in order to identify and assess risks to personal information protection. Another problem is the lack of effective privacy protection methods and technologies for certain data types or specific scenarios. There is also an absence of decision models that comprehensively consider various factors and stakeholder perspectives. Therefore, this study delves into aspects related to these challenges, such as privacy protection methods and decision models, to enhance the effectiveness and sustainability of personal information protection.

RESEARCH METHODOLOGY

Government Data Openness Policies and Personal Information Protection Laws and Regulations

Government data openness policies and personal information protection laws and regulations are crucial cornerstones for the lawful execution of government data openness while safeguarding the privacy of individuals' information. Nations and organizations must proactively advocate the development and reinforcement of these policies and laws. They should also bolster data and personal information protection through a combination of technological innovations and effective management strategies. These measures are essential for harnessing the full spectrum of social benefits and fostering economic development.

Table 1 provides an overview of government data openness policies:

Table 1. Government data openness policies

Government Data Openness Policies	Specific Contents
Regulations on the Disclosure of Government Information in China (Qian & Chen, 2021)	These regulations define the scope, procedures, and methods for government information disclosure, requiring the proactive disclosure of information by the government and ensuring the public's right to know and supervise.
Regulations on Government Data Openness in China (H. Liu et al., 2022)	These regulations sets out the principles and procedures for government data openness, requiring the government's proactive release of data resources to promote information sharing and social innovation.
Open Government Partnership (Muhmad & Muhamad, 2021)	This policy encourages governments to be more transparent, accountable, and open. Member countries commit to taking concrete measures to promote the openness and utilization of government data.
EU Open Data Strategy (Chen et al., 2022)	The European Union encourages member states' governments to proactively open data resources and promote data reuse and innovation while protecting personal privacy and data security.

Introducing these government data openness policies provides a legal basis and guidance for government data openness, emphasizing the importance of information disclosure and personal privacy protection. The laws and regulations on personal information protection are presented in Table 2:

Laws and regulations governing the protection of personal information are designed to enhance the security and responsible utilization of individual data, with a focus on preserving personal privacy and information rights. These regulations are underpinned by key principles, including legality, purpose limitation, data minimization, transparency, and security. They mandate that the processing of personal information must conform to legal requirements and necessitates explicit consent from individuals, ensuring the legitimacy and compliance of data handling practices.

Moreover, these regulations stress that the collection and utilization of personal information must serve clear and lawful purposes and must operate within boundaries prescribed by the law to prevent any form of misuse. The processing of personal information should be confined to the

Table 2. Laws and regulations on personal information protection

Laws and Regulations on Personal Information Protection	Specific Contents
Personal Information Protection Law (Zhang, 2021)	This law is China's first comprehensive legislation specifically targeting the protection of personal information. It aims to strengthen the security and reasonable use of personal data, specifying obligations and responsibilities regarding collecting, storing, processing, and transmitting confidential information.
Administrative Measures for Internet Information Services (Berdik et al., 2021)	This regulation requires internet companies to lawfully, properly, and necessarily handle collected personal information while safeguarding users' rights to choose and be informed about how their data will be used.
Technical Specification for Personal Information Security (Georgiadou et al., 2022)	This specification outlines the technical requirements and management measures for personal information security, guiding various organizations in implementing effective protection measures while handling personal information.
General Data Protection Regulation in Europe (White et al., 2022)	This regulation emphasizes the principles and obligations of personal information protection, specifying the processing of personal information, cross-border transfers, and the protection of individual rights.
Personal Privacy Protection and Electronic Document Act (Boerman et al., 2021)	This act defines principles and rules for personal information processing, highlighting individual control and data protection.

extent necessary for achieving the stated lawful purpose, eschewing any unnecessary collection of data. Governments and businesses are obligated to communicate the purposes, methods, and scope of information collection and usage, thereby ensuring transparency throughout the information processing cycle. These regulations also necessitate the reinforcement of security measures for personal information, including appropriate technical and managerial measures, to mitigate risks such as information leakage and misuse.

At the same time, data openness policies play a pivotal role in enhancing government transparency and facilitating information sharing, thereby fostering social innovation and development. These policies underscore the principles of proactive disclosure, information sharing, and privacy protection. In the course of data openness initiatives, governments should actively make data resources accessible to the public, since this would elevate the transparency of government information and bolster public oversight of governmental activities. Governments are encouraged to openly share data resources with businesses, academia, and the public, thereby promoting innovation and facilitating research through data utilization. Importantly, during the process of data openness, it is imperative to safeguard personal privacy and sensitive information, ensuring that data openness initiatives do not compromise individual rights.

The synergy between these policies and regulations profoundly influences government data openness and personal information protection. Government data openness policies stimulate information sharing and drive social innovation while augmenting government transparency. However, striking a balance is crucial to safeguard privacy and prevent the inadvertent disclosure of personal information. Conversely, laws and regulations governing personal information protection establish the framework for the lawful and secure processing of personal data, thus protecting individual information rights. In the context of government data openness, these regulations serve as guiding principles and standards, ensuring that privacy protection principles are upheld even as data is made open to the public. In summation, government data openness policies and personal information protection laws and regulations jointly provide a comprehensive framework and guidance for the confluence of government data openness and personal information protection, thus promoting a harmonious development of information sharing and privacy preservation.

Decision Tree Model

The decision tree model is a machine learning algorithm that is utilized for classification and regression problems. It uses a tree-like structure to process simple decision rules from the data, enabling predictions or decisions on new data (Idrees et al., 2021). When constructing a decision tree model, the following steps are involved:

1. Feature Selection: In the context of government data openness, metrics like information gain and the Gini coefficient are used to assess the importance of each feature. These metrics help identify which features are crucial for identifying and protecting personal information.
2. Tree Growth: The decision tree model generates a hierarchical tree structure by recursively partitioning the data. A feature is chosen at each node to split the data into different branches. This process determines the optimal splitting condition by calculating impurities on various branches. As a result, the decision tree gradually grows, partitioning the data into different categories.
3. Pruning Process: Decision trees are prone to overfitting, so pruning is used to enhance their generalization ability. Pruning involves removing some branches of the decision tree through cross-validation, which simplifies the tree while maintaining good classification performance.
4. Rule Generation: The final decision tree model can be transformed into a set of rules that clearly describe the data's classification process. These rules are very useful for government agencies and relevant stakeholders since they provide an intuitive understanding of the logic behind data classification.

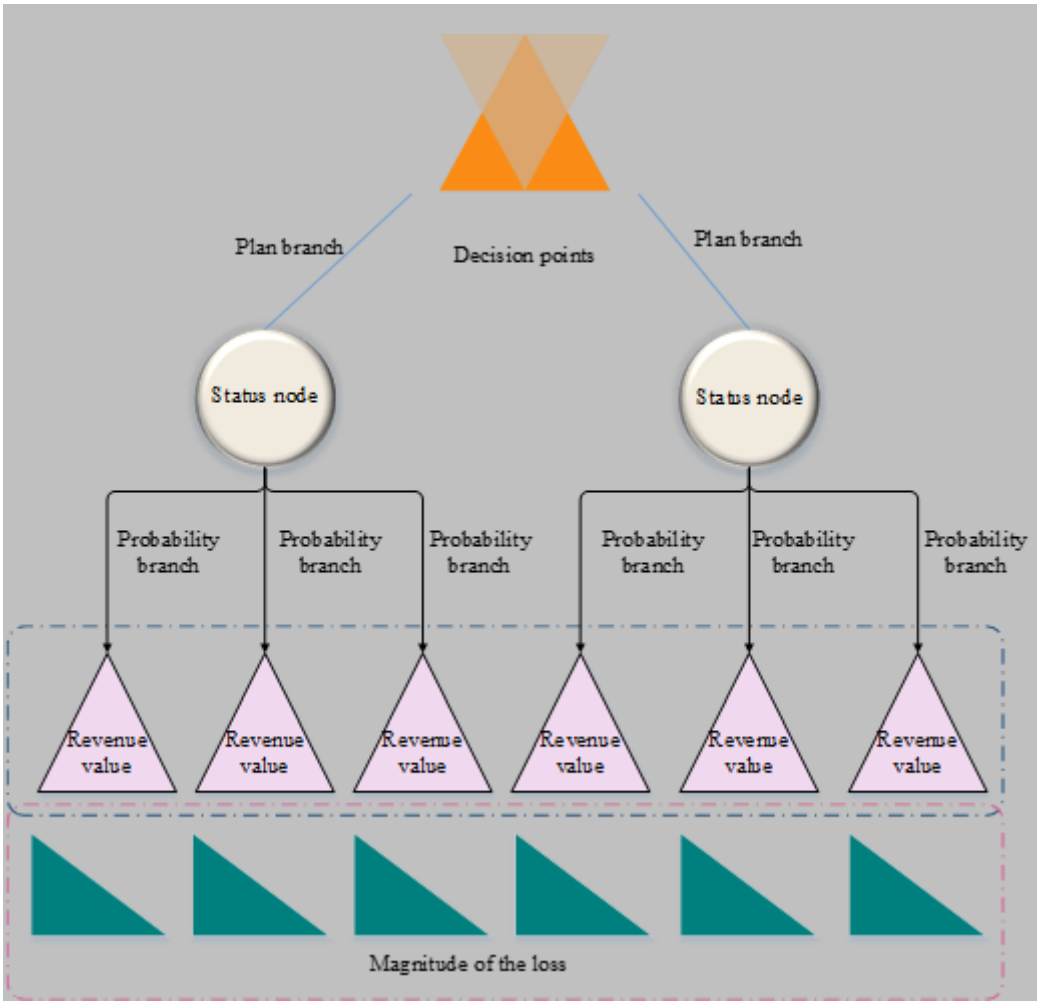
In summary, constructing a decision tree model involves key steps such as feature selection, tree growth, and pruning. These steps help to effectively identify and protect personal information in government data openness by achieving a balance between data openness and privacy protection. The decision tree algorithm process is depicted in Table 3 and the structure of the decision tree model is illustrated in Figure 1:

The structure of the decision tree model, as illustrated in Figure 1, is built based on the partitioning of the feature space. It divides the feature space into different regions, each of which corresponds to a decision rule or prediction outcome. Each internal node of the decision tree represents a feature or attribute, each branch represents a feature value, and each leaf node represents a decision result or category. The decision tree recursively partitions the features of the training data to minimize sample

Table 3. Decision tree algorithm process

Enter:	training set: $D = \left\{\left(x_1, y_1\right), \left(x_2, y_2\right), \cdots, \left(x_m, y_m\right)\right\}$ attribute set: $A = \left\{a_1, a_2, \cdots, a_d\right\}$
Process:	Function: TreeGenerate(D,A)
	1: generate node node;
	2: If in *D*, the samples all belong to the same category *C* then
	3: Mark the node as a *C*-type leaf node; return
	4: end if
	5: If $A = \varnothing OR$ *D*, the sample takes the same value on A, then
	6: Mark the node as a leaf node, and its category is marked as the class with the largest number of samples in *D*; return
	7: end if
	8: Select the optimal partition attribute a_* from *A*;
	9: for each value of a_* a_*^v do
	10: Generate a branch for the node; let D_v represent the sample subset in *D* whose value is a_*^v on a_*;
	11:if D_v is empty then
	12: Mark the branch node as a leaf node, and its category is marked as the class with the most samples in *D*; return
	13: else
	14: Take $TreeGenerate\left(D_v, \frac{A}{\left\{a_*\right\}}\right)$ as the branch node
	15: end .if
	16: end for
Output	A decision tree with a node as the root node

Figure 1. Structure of the decision tree model

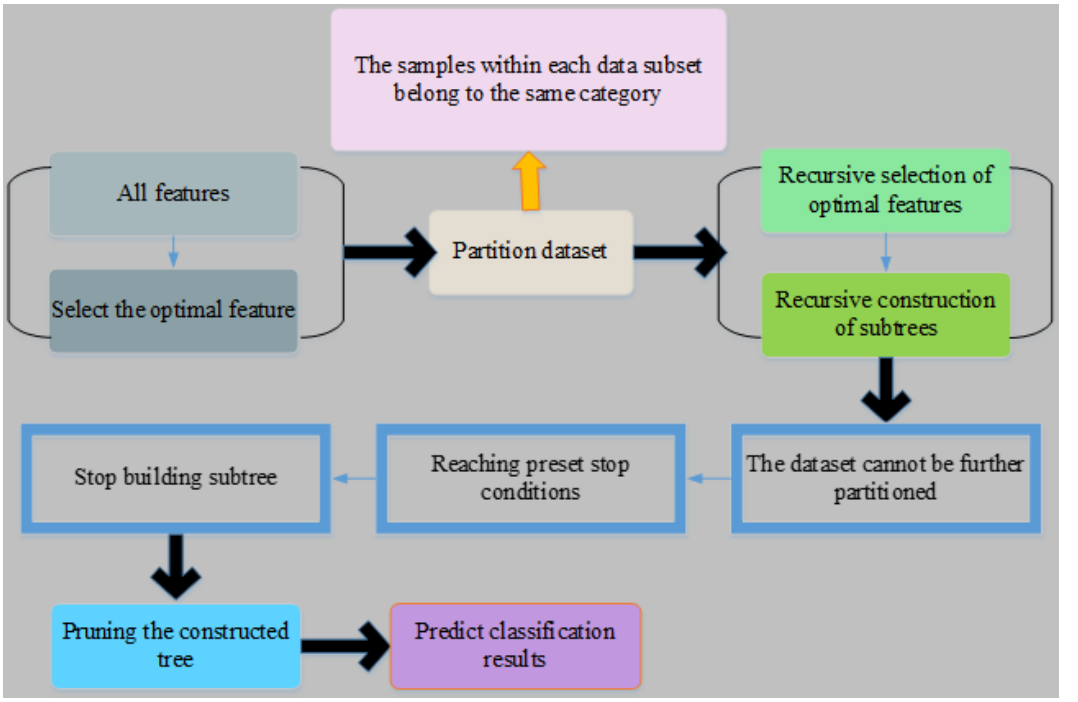


impurity or maximize information gain in generating the tree's structure (Guzek et al., 2020). The construction process of the decision tree is depicted in Figure 2.

As shown in Figure 2, the construction process of the decision tree includes feature selection, tree growth, and pruning. Information gain is used a metric in order to construct a decision tree to select the optimal feature. It measures the reduction of uncertainty in a sample set under a certain feature condition. By calculating the information gain for each feature, the feature with the highest information gain can be chosen as the splitting feature for the current node. Information entropy is a measure of uncertainty in a sample set. Higher information entropy indicates greater uncertainty in the sample set. In the decision tree model, uncertainty in the sample set is reduced by selecting the splitting feature, leading to lower information entropy in the subsets.

There is a close relationship between information gain and information entropy. Information gain can be seen as the difference between the original set's information entropy and the subsets' average information entropy obtained by splitting while using a certain feature. In other words, information gain quantifies how much uncertainty in the original set is reduced when a certain feature is chosen. Therefore, in the process of model establishment, combining information gain and information

Figure 2. Construction process of the decision tree



entropy when making decisions to split tree nodes helps select the optimal splitting feature, which leads to a better assessment of risks related to the protection of personal information. This approach contributes to enhancing the model's performance and accuracy.

Based on the metrics for feature selection, such as information gain, Gini index, or variance, the optimal feature is chosen as the splitting attribute for the current node. The training data is divided into different subsets, and a subtree is recursively constructed for each subgroup using the selected feature. To avoid overfitting, pruning can be performed to simplify the decision tree by removing unimportant branches or merging similar leaf nodes (Wang & Govindarasu, 2020). Information entropy can be calculated as shown in Equation (1):

$$Ent\left(D\right) = -\sum_{k=1}^{|y|} p_k log_2 p_k \tag{1}$$

In Equation (1), D represents the dataset, k is the number of possible classes in the dataset, and p_k represents the probability of the k-th class occurring.

Information gain can be calculated as shown in Equation (2):

$$Gain\left(D,a\right) = Ent\left(D\right) - \sum_{v=1}^{V} \frac{\left|D_v\right|}{\left|D\right|} Ent\left(D_v\right) \tag{2}$$

In Equation (2), a represents a specific feature attribute, V represents the number of values for that attribute, D_v represents the subset of samples corresponding to the v-th attribute value, and $\left|D\right|$ represents the size of the dataset.

The Gini index can be calculated as shown in Equation (3):

$$Gini\left(D\right) = \sum_{k=1}^{|y|}\sum_{k' \neq k} p_k p_k^{'} \tag{3}$$

In Equation (3), D represents the dataset, y represents the possible number of classes in the dataset, and p_k represents the probability of the k-th class occurring (Jiang et al., 2020).

The Process of Personal Information Protection in Government Data Openness Based on the Complexity-Pruned Decision Tree

There may be many discrete or continuous attributes in personal information protection in government data openness. It is typically desirable for the decision tree model to have good generalization ability, meaning a high accuracy in predicting unknown data. This is why the complexity pruning decision tree algorithm is employed. It is an optimization method based on complexity as a cost function, which constrains the complexity of the model to avoid overfitting and improve the model's generalization ability (Sajja et al., 2021). This algorithm can handle discrete and continuous attributes simultaneously, making the model more versatile. Figure 3 illustrates the advantages of using the complexity pruning decision tree algorithm to protect personal information.

As shown in Figure 3, traditional decision tree algorithms can generate complex models that adapt to different types and scales of data, exhibiting higher flexibility. Traditional decision trees also show greater accuracy on training data, since they can fit the data's details more efficiently. However, traditional decision tree algorithms are prone to overfitting training data, which results in poor generalization to new data and an increased risk of personal information leakage. Traditional decision trees may also generate large, intricate models that are difficult to interpret and comprehend, limiting their applicability in government data openness.

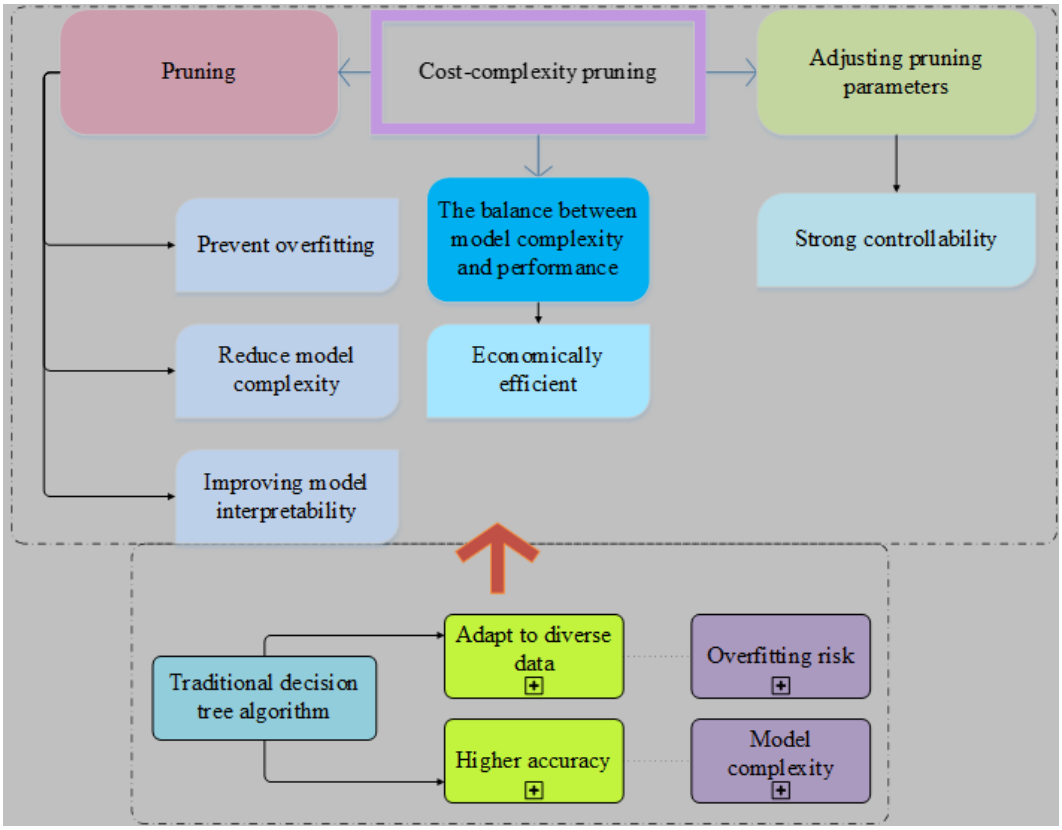
On the other hand, complexity-based pruning decision tree algorithms mitigate the risk of overfitting training data by restricting the depth and branching of nodes during the tree's growth. This is particularly crucial for protecting personal information in government data disclosures, since overfitting can lead to diminished performance on new data, which would then increase the risk of personal information leakage. Pruned decision trees are relatively simpler, generating more intuitive and comprehensible rules. This aids government agencies and other stakeholders in grasping the fundamental principles of personal information protection decisions, enabling better participation in government data openness and protection.

Models generated by complexity-based pruning decision tree algorithms are also more streamlined and require relatively fewer computational resources. In the context of government data openness, where data volumes are substantial, resource consumption is a significant consideration. Pruned models can handle large-scale data more effectively, making them more suitable for processing government data. Taking all these points into account, it is clear that complexity-based pruning decision tree algorithms outperform traditional decision tree algorithms in protecting personal information in government data disclosures.

The steps followed by the complexity-based pruning decision tree algorithm are:

i. Generate a complete decision tree from the training set;
ii. Record the test error rate and complexity of each node along with its corresponding subtree;

Figure 3. Advantages of using the complexity pruning decision tree algorithm



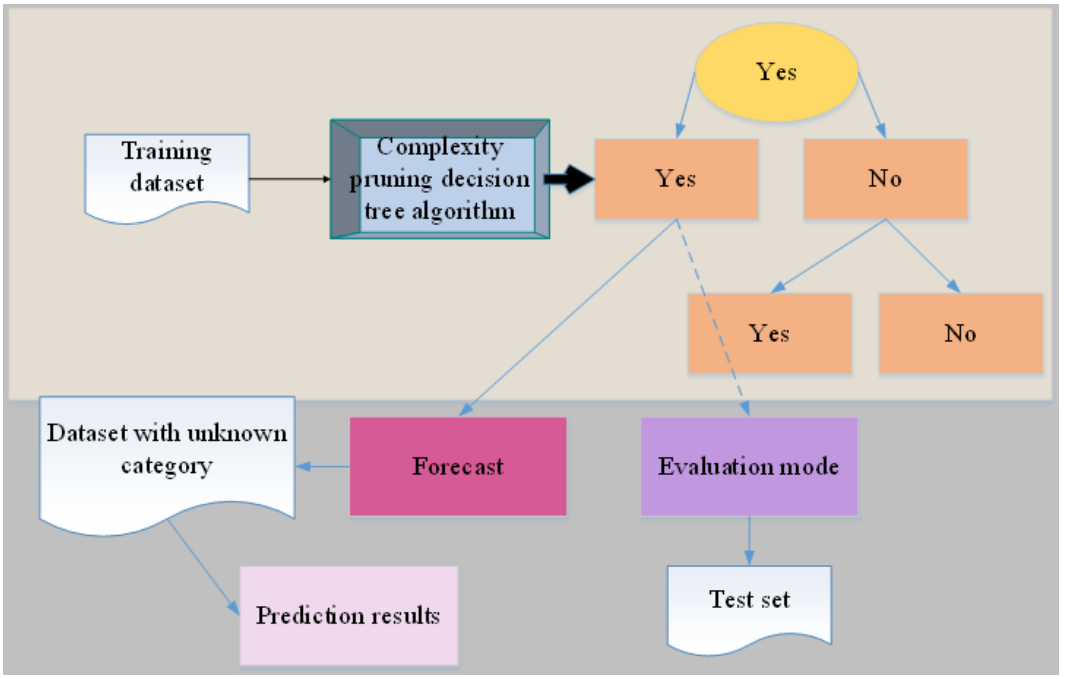
iii. Starting from the leaf nodes, accumulate the test error rate and complexity, creating a curve;
iv. Select the node on the curve with the minimum test error rate as the root node to generate the final decision tree.

The process of utilizing this algorithm for personal information protection in government data openness is illustrated in Figure 4.

Figure 4 shows that within the complexity-based pruning decision tree model, government agencies undertake feature selection. This selection is based on the characteristics and objectives of the data, ensuring that the chosen features contain meaningful information for the model's predictions. After the feature selection phase, government agencies utilize the complexity-based pruning decision tree algorithm for data training. This algorithm automatically determines the optimal pruning parameters, striking a balance between the tree's complexity and its performance. This results in a streamlined decision tree model that maintains high performance.

Government agencies begin the process of feature selection by carefully evaluating the various features present in the data. They select features based on the data's characteristics as well as the predictive objectives of the model. This meticulous process ensures that the chosen features contribute valuable information to the model's predictions. Following the completion of feature selection, government agencies apply the complexity-based pruning decision tree algorithm to train the data. The algorithm's key feature lies in its ability to automatically select the optimal pruning parameters while preserving the decision tree's performance (S. Ye & Zhao, 2023; S. Ye, Yao, et al., 2023). This

Figure 4. Process of utilizing a complexity-based pruning decision tree model for personal information protection in government data openness



selection process effectively manages the trade-off between complexity and performance, resulting in a simplified decision tree model that performs admirably.

EXPERIMENTAL DESIGN AND PERFORMANCE EVALUATION

Data Collection

The data used in this study is sourced from the Singapore government's open dataset, which offers access to 1,700 datasets spread across nine data directories that categorize and organize the data. Each directory represents a distinct theme or domain, encompassing demographics, education, transportation, and more. For our experiments, we partition the datasets into training and testing sets. Using a predetermined ratio, 70% of the data is allocated to the training set, while the remaining 30% forms the testing set. This division allows us to construct the decision tree model using the training set and assess the model's performance and generalization ability on the testing set.

Experimental Environment

The environment and parameters used for the experiment are as follows:

- Operating System: Windows 10
- Software Environment: Python 3.7
- Data Analysis and Modeling Library: Scikit-learn
- Interactive Programming Environment: Jupyter Notebook
- Hardware Configuration: Intel Core i5 CPU, 8 GB RAM

Throughout the experiments, we employ the Pandas data processing library for various tasks, including data preprocessing, cleansing, and feature engineering.

Parameter Settings

When utilizing the complexity-based pruning decision tree algorithm for model training, it is essential to configure the following key parameters:

1. Maximum Depth: This parameter restricts the maximum depth of the generated decision tree to prevent overfitting. We set the maximum depth to 5.
2. Minimum Samples per Leaf Node: To avoid an overly fragmented tree with too few leaf nodes and, subsequently, reduced complexity, we establish a minimum of 5 samples per leaf node.
3. Minimum Samples for Node Splitting: To prevent premature splitting of non-leaf nodes due to insufficient data samples, we specify a minimum of 10 samples for node splitting.
4. Parameter Controlling Complexity-Performance Balance: We set this parameter to 0.001 to govern the trade-off between complexity and performance during the pruning process.

This study employs metrics such as accuracy, recall, F1 score, and Area Under the Curve (AUC) to assess the model's performance, as shown in Equations (4)–(8):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{\left(Precision * Recall * 2\right)}{\left(Precision + Recall\right)} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

In Equations (4)–(8), TP represents the number of samples correctly predicted as positive. TN represents the number of samples correctly predicted as negative. FP represents the number of samples incorrectly predicted as positive. FN represents the number of samples incorrectly predicted as negative. The total number of samples is the sum of true positives, true negatives, false positives, and false negatives ($TP + TN + FP + FN$). The AUC value is the sum of the areas under different parts of the Receiver Operating Characteristic (ROC) curve, where the ROC curve's horizontal axis represents the False Positive Rate (FPR), and the vertical axis represents Recall.

Performance Evaluation

Importance of Personal Information Features in Government Data Openness

The significance of personal information features in government data openness is depicted in Figure 5. It shows that the highest importance score is assigned to the age feature, standing at 0.25. This underscores the substantial impact of subjects' ages on personal information protection, implying a potential risk of revealing individual identities or sensitive data. Likewise, the gender feature receives a noteworthy importance score of 0.12, which indicates that it, too, has great relevance in personal information protection. Similarly, the income feature scores 0.08, suggesting its potential to disclose sensitive information in the context of data openness.

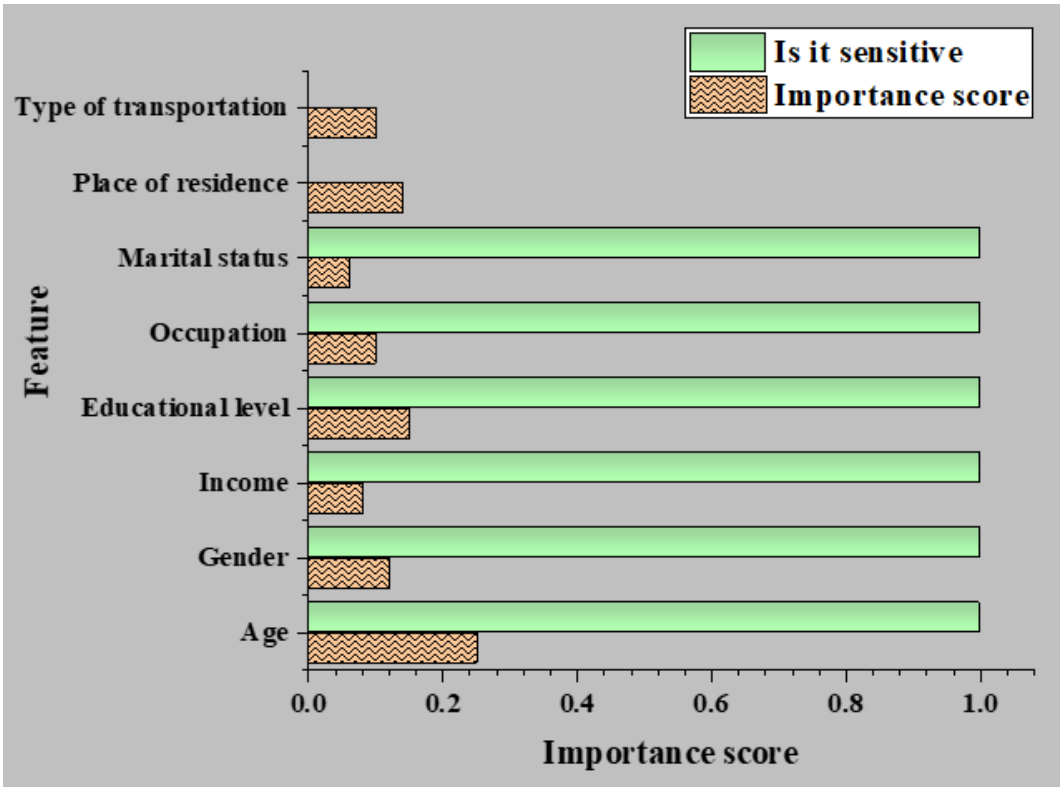
Other features of importance include the education level feature, with an importance score of 0.15, which shares a level of significance comparable to age, highlighting it as a crucial sensitive attribute as well. The occupation feature, scoring 0.1, also signals its potential to expose personal identities or sensitive details. Even the marital status feature, which has a lower importance score of 0.06, still falls within the category of sensitive attributes. The importance scores for residence and transportation mode features are 0.14 and 0.1 respectively. These scores, while not negligible, are considered less likely to raise privacy concerns.

In summary, government data openness necessitates special attention when handling sensitive features such as age, gender, income, education level, and occupation, in order to safeguard individual privacy. Meanwhile, although residence and transportation modes exhibit relatively higher feature importance, they are perceived as less likely to pose privacy risks. The complexity-based pruning decision tree model can enhance personal information protection performance, particularly for these sensitive features.

Model Performance for Personal Information Protection in Government Data Openness Using the Complexity-Based Pruning Decision Tree Model

Figure 6 illustrates the model performance for personal information protection in government data openness through the utilization of the complexity-based pruning decision tree model. The model attains an accuracy of 0.85 on the training set and 0.8 on the testing set. These figures indicate commendable performance on the training set but reveal some overfitting concerns on the testing set.

Figure 5. Importance of personal information features in government data openness ("1" indicates sensitive, "0" indicates non-sensitive)



To address this overfitting issue, it may be necessary to take additional measures, such as adjusting model parameters or augmenting the dataset.

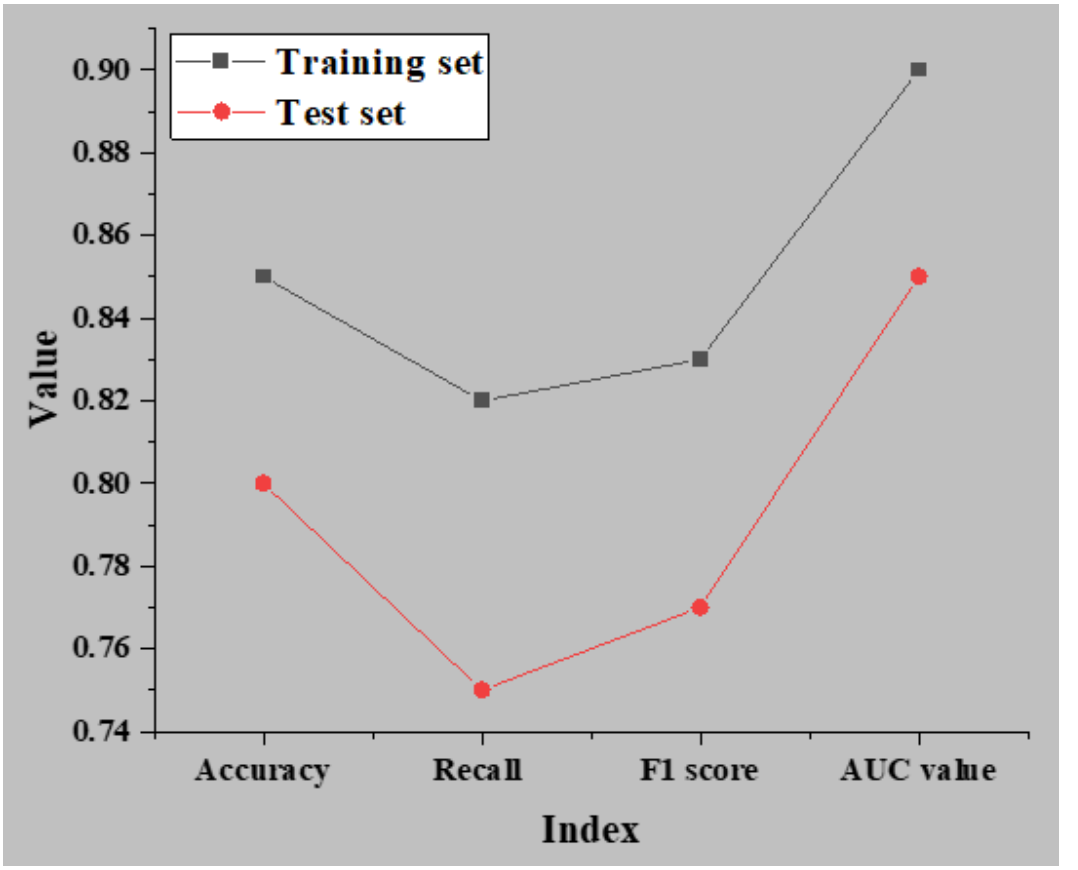
The Area under the ROC Curve (AUC) values are 0.9 for the training set and 0.85 for the testing set. These scores indicate that the model excels in distinguishing between positive and negative samples. Still, there is room for improvement in enhancing the AUC value.

In the context of government data openness, the complexity-based pruning decision tree model strikes a good balance. It effectively maintains data privacy and security while preserving valuable information to the extent possible. The model's performance underscores its ability to filter out extraneous data, thereby reducing data loss.

Government data openness mandates a degree of data access control to prevent unauthorized access and misuse. The application of the complexity-based pruning decision tree model in government data openness establishes an access control mechanism for the data. Through the definition of permissions and access rules within the model, the government can ensure that only authorized users gain access to specific data features. This effectively narrows the scope of data access and mitigates potential data leak risks.

In conclusion, the complexity-based pruning decision tree model plays a pivotal role in safeguarding personal information within the realm of government data openness. By incorporating features such as feature selection and pruning algorithms, this model mitigates data loss to a considerable extent and facilitates data access authorization control.

Figure 6. Model performance of personal information protection in government data openness using a complexity-based pruning decision tree model



Model Performance Analysis of Different Decision Tree Algorithms in Personal Information Protection for Government Data Openness

Figure 7 provides an overview of the model performance of various decision tree algorithms in personal information protection for government data openness. The complexity-based pruning decision tree model demonstrates impressive performance across multiple metrics, including accuracy, recall, F1 score, and AUC value. These results signify that the model excels at safeguarding personal information within the context of government data openness, with an accuracy of 0.85, a recall rate of 0.82, an F1 score of 0.83, and an AUC value of 0.9.

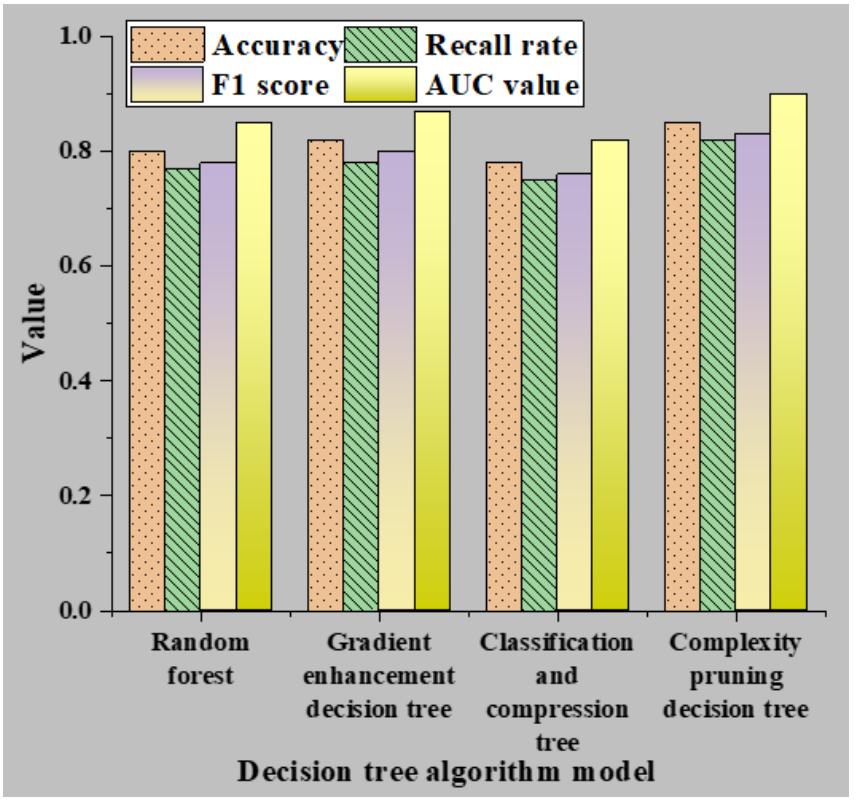
Comparatively, the model performances of Random Forest and Gradient Boosting Decision Trees closely align, achieving accuracies of 0.8 and 0.82 respectively.

In contrast, the Classification and Compression Tree exhibits the lowest precision among the models, registering an accuracy of 0.78.

Discussion

Recent studies have shed light on the potential and effectiveness of applying machine learning models for personal information protection. For instance, Miao et al. (2021) conducted a comprehensive review of machine learning-based attack methods, focusing on controlled information spanning controlled user activities, controlled machine learning model-related information, and controlled authentication information. Their research confirmed that machine learning models can indeed serve as effective safeguards for controlled personal information.

Figure 7. Model performance results of different decision tree algorithm models in personal information protection for government data openness



Harinahalli Lokesh et al. (2021) also delved into the applicability of classical machine learning techniques like random forests, k-nearest neighbors, decision trees, and support vector machines (SVM) in identifying and countering phishing attacks. Their findings revealed both positive and negative aspects, demonstrating that machine learning models can be instrumental in protecting website information data and mitigating the impact of phishing attacks.

Additionally, Yan et al. (2020) contributed a novel hinge-based classification algorithm with adaptive learning rates and momentum for mini-batch gradient descent, designed to minimize a network's vulnerability to security attacks. In comparison to traditional neural networks, decision trees, and logistic regression, their algorithm significantly improved the performance of deep network training in terms of scale and speed.

Building upon these research findings, alongside the emphasis on feature analysis in personal information protection discussed in this study, the performance results of the complexity-based pruning decision tree model reinforce the pivotal role of machine learning models in this domain. The complexity-pruned decision tree model exhibits impressive performance metrics, boasting an accuracy of 0.85, recall rate of 0.82, F1 score of 0.83, and an AUC value of 0.9.

Furthermore, the advantages of the complexity-pruned decision tree model become apparent when comparing its performance to that of Random Forest, Gradient Boosting Decision Trees, and Classification and Compression Trees. This underscores the importance of ongoing exploration and optimization of related algorithms to enhance these models' predictive capabilities and overall performance.

Moreover, when dealing with data openness, safeguarding sensitive features takes on paramount importance to ensure the preservation of individual privacy. The complexity-based pruning decision tree model serves as a valuable tool in this regard, contributing to effective data access control and personal information protection.

CONCLUSION

Research Contribution

This study's findings and insights underscore the exceptional performance of the complexity-based pruning decision tree model in terms of accuracy, recall, F1 score, and AUC value, highlighting its effectiveness. The research centers on the paramount importance of personal information protection in the context of government data openness, addressing contemporary privacy and information security challenges in the digital age.

The strengths of this study are multifaceted and encompass the following key aspects:

- Comprehensive Perspective: This study meticulously analyzes government data openness policies and personal information protection regulations, and couples this with the practical application of decision tree models. Therefore, this research offers a holistic viewpoint, combining both theoretical and empirical dimensions.
- Innovative Model: The study introduces a novel complexity-pruned decision tree model tailored to address personal information protection concerns in government data openness. This model not only efficiently leverages data while safeguarding individual privacy, but also exhibits outstanding performance in real-world applications. It accomplishes this by seamlessly integrating feature selection and pruning algorithms, and effectively optimizing tree complexity and performance. In doing so, it strikes an ideal balance between privacy preservation and data analysis requirements.
- Practical Analytical Tool: By leveraging the complexity-based pruning decision tree model, this study furnishes an invaluable analytical tool for safeguarding personal information within government data openness. At both policy and implementation levels, it underscores the vital significance of personal information protection. Establishing robust legal regulations and technical

measures to ensure individual privacy protection while promoting government data openness is crucial for enhancing government transparency, fostering social innovation, and bolstering public trust.

Future Work and Research Limitations

While this study has made significant strides in the realm of personal information protection within government data openness, some limitations and opportunities for future research warrant consideration.

Research Limitations

- Methodological Focus: This study primarily focuses on the application and performance assessment of decision tree models. It provides limited comparisons and analyses of alternative methods and technologies for personal information protection. Future research should explore and evaluate a wider array of approaches, including various machine learning algorithms, privacy-preserving techniques, and encryption methods, in order to offer more comprehensive and effective solutions.
- Data Set Specificity: The empirical analysis conducted in this study relies exclusively on the Singapore government's open data set. The extent to which the findings apply to data sets from other countries or regions has not been fully explored. Future research should investigate the adaptability and generalizability of these methods to diverse data sets.
- Parameter Uncertainties: In the experimental process, decisions were made regarding methods and parameter settings for data preprocessing, feature selection, and model optimization. These choices may introduce uncertainties and require further fine-tuning and exploration to enhance the robustness of the analysis.

Future Directions

- Comparative Studies: Future research can delve into comparative studies that rigorously assess the strengths and weaknesses of various personal information protection methods, drawing on a broader array of machine learning algorithms and privacy-enhancing technologies. These studies could aid in identifying the most suitable approaches for specific data openness contexts.
- Cross-Regional Analysis: Expanding the scope of analysis to encompass data sets from different countries and regions will facilitate a more comprehensive understanding of the challenges and opportunities in personal information protection within the context of government data openness.
- Parameter Optimization: Ongoing efforts to refine and optimize parameter settings for data preprocessing, feature selection, and model configuration can contribute to more robust and reliable personal information protection mechanisms.

In conclusion, while this study provides valuable insights and a solid foundation for personal information protection in government data openness, there is ample room for further research to explore a wider range of methods, evaluate cross-regional applicability, and enhance the precision of parameter settings to advance this critical field.

FUNDING

This work was supported by "the Fundamental Research Funds for the Central Universities" (No. CUC230B023) and(No.CUC230B030).

REFERENCES

Berdik, D., Otoum, S., Schmidt, N., Porter, D., & Jararweh, Y. (2021). A survey on blockchain for information systems management and security. *Information Processing & Management*, *58*(1), 102397. doi:10.1016/j.ipm.2020.102397

Boerman, S. C., Kruikemeier, S., & Zuiderveen Borgesius, F. J. (2021). Exploring motivations for online privacy protection behavior: Insights from panel data. *Communication Research*, *48*(7), 953–977. doi:10.1177/0093650218800915

Chen, Z., Azman, A. S., Chen, X., Zou, J., Tian, Y., Sun, R., Xu, X., Wu, Y., Lu, W., Ge, S., Zhao, Z., Yang, J., Leung, D. T., Domman, D. B., & Yu, H. (2022). Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nature Genetics*, *54*(4), 499–507. doi:10.1038/s41588-022-01033-y PMID:35347305

Chi-Cheng, H. (2020). User's segmentation on continued knowledge management system use in the public sector. *Journal of Organizational and End User Computing*, *32*(1), 19–40. doi:10.4018/JOEUC.2020010102

Culnan, M. J. (2019). Policy to avoid a privacy disaster. *Journal of the Association for Information Systems*, *20*(6), 1. doi:10.17705/1jais.00554

Dai, B., Fu, D., Meng, G., Liu, B., Li, Q., & Liu, X. (2020). The effects of governmental and individual predictors on COVID-19 protective behaviors in China: A path analysis model. *Public Administration Review*, *80*(5), 797–804. doi:10.1111/puar.13236 PMID:32836438

Davidson, E., Wessel, L., Winter, J. S., & Winter, S. (2023). Future directions for scholarship on data governance, digital innovation, and grand challenges. *Information and Organization*, *33*(1), 100454. doi:10.1016/j.infoandorg.2023.100454

Desborough, J., Dykgraaf, S. H., Rankin, D., & Kidd, M. (2020). The importance of consistent advice during a pandemic: 'An analysis of Australian advice regarding personal protective equipment in healthcare settings during COVID-19'. *Australian Journal of General Practice*, *49*(6), 369–372. doi:10.31128/AJGP-04-20-5374 PMID:32464735

Feng, J., Rong, C., Sun, F., Guo, D., & Li, Y. (2020). PMF: A privacy-preserving human mobility prediction framework via federated learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(1), 1–21. doi:10.1145/3381006 PMID:35846237

Feng, Z., & Chen, M. (2022). Platformance-based cross-border import retail e-commerce service quality evaluation using an artificial neural network analysis. *Journal of Global Information Management*, *30*(11), 1–17. doi:10.4018/JGIM.306271

Ferrag, M. A., Maglaras, L., Ahmim, A., Derdour, M., & Janicke, H. (2020). Rdtids: Rules and decision tree-based intrusion detection system for internet-of-things networks. *Future Internet*, *12*(3), 44. doi:10.3390/fi12030044

Georgiadou, A., Mouzakitis, S., Bounas, K., & Askounis, D. (2022). A cyber-security culture framework for assessing organization readiness. *Journal of Computer Information Systems*, *62*(3), 452–462. doi:10.1080/08874417.2020.1845583

Getman, A. P., Danilyan, O. G., Dzeban, A. P., Kalinovsky, Y. Y., & Hetman, Y. A. (2020). Information security in modern society: Sociocultural aspects. *Amazonia Investiga*, *9*(25), 6–14.

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, *20*(1), 1–40. doi:10.1186/s12874-020-00977-1 PMID:32381039

Gupta, R., Gupta, I., Saxena, D., & Singh, A. K. (2023). A differential approach and deep neural network based data privacy-preserving model in cloud environment. *Journal of Ambient Intelligence and Humanized Computing*, *14*(5), 4659–4674. doi:10.1007/s12652-022-04367-x PMID:36684481

Guzek, D., Skolmowska, D., & Głąbska, D. (2020). Analysis of gender-dependent personal protective behaviors in a national sample: Polish adolescents' COVID-19 experience (PLACE-19) study. *International Journal of Environmental Research and Public Health*, *17*(16), 5770. doi:10.3390/ijerph17165770 PMID:32785004

Harinahalli Lokesh, G., & BoreGowda, G. (2021). Phishing website detection based on effective machine learning approach. *Journal of Cyber Security Technology*, *5*(1), 1–14. doi:10.1080/23742917.2020.1813396

Hasal, M., Nowaková, J., Ahmed Saghair, K., Abdulla, H., Snášel, V., & Ogiela, L. (2021). Chatbots: Security, privacy, data protection, and social aspects. *Concurrency and Computation*, *33*(19), e6426. doi:10.1002/cpe.6426

Hendricks-Sturrup, R. M., & Lu, C. Y. (2019). Direct-to-consumer genetic testing data privacy: Key concerns and recommendations based on consumer perspectives. *Journal of Personalized Medicine*, *9*(2), 25. doi:10.3390/jpm9020025 PMID:31075859

Hoofnagle, C. J., Van Der Sloot, B., & Borgesius, F. Z. (2019). The European Union general data protection regulation: What it is and what it means. *Information & Communications Technology Law*, *28*(1), 65–98. doi:10.1080/13600834.2019.1573501

Idrees, S. M., Nowostawski, M., & Jameel, R. (2021). Blockchain-based digital contact tracing apps for COVID-19 pandemic management: Issues, challenges, solutions, and future directions. *JMIR Medical Informatics*, *9*(2), e25245. doi:10.2196/25245 PMID:33400677

Islam, M. T., Talukder, M. S., Khayer, A., & Islam, A. N. (2023). Exploring continuance usage intention toward open government data technologies: An integrated approach. *VINE Journal of Information and Knowledge Management Systems*, *53*(4), 785–807. doi:10.1108/VJIKMS-10-2020-0195

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly*, *37*(3), 101493. doi:10.1016/j.giq.2020.101493

Jiang, J., Zhu, X., Han, G., Guizani, M., & Shu, L. (2020). A dynamic trust evaluation and update mechanism based on C4. 5 decision tree in underwater wireless sensor networks. *IEEE Transactions on Vehicular Technology*, *69*(8), 9031–9040. doi:10.1109/TVT.2020.2999566

Jupalle, H., Kouser, S., Bhatia, A. B., Alam, N., Nadikattu, R. R., & Whig, P. (2022, August). Automation of human behaviors and its prediction using machine learning. *Microsystem Technologies*, *28*(8), 1879–1887. doi:10.1007/s00542-022-05326-4

Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2021). When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys*, *54*(2), 1–36. doi:10.1145/3436755

Liu, H., Lei, H., & Zhou, Y. (2022). How does green trade affect the environment? Evidence from China. *Journal of Economic Analysis*, *1*(1), 1–19. doi:10.58567/jea01010001

Liu, J., Zhao, J., Huang, H., & Xu, G. (2022). A novel logistics data privacy protection method based on blockchain. *Multimedia Tools and Applications*, *81*(17), 2386–23887. doi:10.1007/s11042-022-12836-w

Liu, Y., Liu, Y., Liu, Z., Liang, Y., Meng, C., Zhang, J., & Zheng, Y. (2020). Federated forest. *IEEE Transactions on Big Data*, *8*(3), 843–854. doi:10.1109/TBDATA.2020.2992755

Löfgren, K., & Webster, C. W. R. (2020). The value of Big Data in government: The case of 'smart cities'. *Big Data & Society*, *7*(1), 2053951720912775. doi:10.1177/2053951720912775

Matheus, R., & Janssen, M. (2020). A systematic literature study to unravel transparency enabled by open government data: The window theory. *Public Performance & Management Review*, *43*(3), 503–534. doi:10.1080/15309576.2019.1691025

Miao, Y., Chen, C., Pan, L., Han, Q. L., Zhang, J., & Xiang, Y. (2021). Machine learning–based cyber attacks targeting on controlled information: A survey. *ACM Computing Surveys*, *54*(7), 1–36. doi:10.1145/3465171

Muhmad, S. N., & Muhamad, R. (2021). Sustainable business practices and financial performance during pre- and post-SDG adoption periods: A systematic review. *Journal of Sustainable Finance & Investment*, *11*(4), 291–309. doi:10.1080/20430795.2020.1727724

Mutimukwe, C., Kolkowska, E., & Grönlund, Å. (2020). Information privacy in e-service: Effect of organizational privacy assurances on individual privacy concerns, perceptions, trust and self-disclosure behavior. *Government Information Quarterly*, *37*(1), 101413. doi:10.1016/j.giq.2019.101413

Naresh, R., Vijayakumar, P., Jegatha, D. L., & Sivakumar, R. (2020). A novel trust model for secure group communication in distributed computing. *Journal of Organizational and End User Computing*, *32*(3), 1–14. doi:10.4018/JOEUC.2020070101

Oni, A. A., Musa, U., & Oni, S. (2020). E-revenue adoption in state internal revenue service: Interrogating the institutional factors. *Journal of Organizational and End User Computing*, *32*(1), 41–61. doi:10.4018/JOEUC.2020010103

Park, J. Y., & Woo, C. H. (2021). Factors affecting nursing students' practice of medical information protection. *CIN: Computers, Informatics*. *Computers, Informatics, Nursing*, *39*(12), 874–882. doi:10.1097/CIN.0000000000000750 PMID:33941720

Price, W. N. II, & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine*, *25*(1), 37–43. doi:10.1038/s41591-018-0272-7 PMID:30617331

Qian, W., & Chen, X. (2021). Corporate environmental disclosure and political connection in regulatory and leadership changes: The case of China. *The British Accounting Review*, *53*(1), 100935. doi:10.1016/j.bar.2020.100935

Javed, A.R., Hassan, M.A., Shahzad, F., Ahmed, W., Singh, S., Baker, T., Gadekallu, T.R.R. (2022). Integration of blockchain technology and federated learning in vehicular (iot) networks: A comprehensive survey. *Sensors, 22*(12), 4394. 10.3390/s22124394

Ramya, K., Teekaraman, Y., & Kumar, K. R. (2019). Fuzzy-based energy management system with decision tree algorithm for power security system. *International Journal of Computational Intelligence Systems*, *12*(2), 1173–1178. doi:10.2991/ijcis.d.191016.001

Ruijer, E., Détienne, F., Baker, M., Groff, J., & Meijer, A. J. (2020). The politics of open government data: Understanding organizational responses to pressure for more transparency. *American Review of Public Administration*, *50*(3), 260–274. doi:10.1177/0275074019888065

Sajja, G. S., Mustafa, M., Ponnusamy, R., & Abdufattokhov, S. (2021). Machine learning algorithms in intrusion detection and classification. *Annals of the Romanian Society for Cell Biology*, *25*(6), 12211–12219.

Shen, M., Tang, X., Zhu, L., Du, X., & Guizani, M. (2019). Privacy-preserving support vector machine training over blockchain-based encrypted IoT data in smart cities. *IEEE Internet of Things Journal*, *6*(5), 7702–7712. doi:10.1109/JIOT.2019.2901840

Shen, S., Zhu, T., Wu, D., Wang, W., & Zhou, W. (2022). From distributed machine learning to federated learning: In the view of data privacy and security. *Concurrency and Computation*, *34*(16), e6002. doi:10.1002/cpe.6002

Staunton, C., Tschigg, K., & Sherman, G. (2021). Data protection, data management, and data sharing: Stakeholder perspectives on the protection of personal health information in South Africa. *PLoS One*, *16*(12), e0260341. doi:10.1371/journal.pone.0260341 PMID:34928950

Vasa, J., & Thakkar, A. (2023). Deep learning: Differential privacy preservation in the era of big data. *Journal of Computer Information Systems*, *63*(3), 608–631. doi:10.1080/08874417.2022.2089775

Wang, P., & Govindarasu, M. (2020). Multi-agent based attack-resilient system integrity protection for smart grid. *IEEE Transactions on Smart Grid*, *11*(4), 3447–3456. doi:10.1109/TSG.2020.2970755

White, T., Blok, E., & Calhoun, V. D. (2022). Data sharing and privacy issues in neuroimaging research: Opportunities, obstacles, challenges, and monsters under the bed. *Human Brain Mapping*, *43*(1), 278–291. doi:10.1002/hbm.25120 PMID:32621651

Wylde, V., Rawindaran, N., Lawrence, J., Balasubramanian, R., Prakash, E., Jayal, A., Khan, I., Hewage, C., & Platts, J. (2022). Cybersecurity, data privacy and blockchain: A review. *SN Computer Science*, *3*(2), 127. doi:10.1007/s42979-022-01020-4 PMID:35036930

Yan, X., Xu, Y., Xing, X., Cui, B., Guo, Z., & Guo, T. (2020). Trustworthy network anomaly detection based on an adaptive learning rate and momentum in IIoT. *IEEE Transactions on Industrial Informatics*, *16*(9), 6182–6192. doi:10.1109/TII.2020.2975227

Ye, S., Yao, K., & Xue, J. (2023). Leveraging empowering leadership to improve employees' improvisational behavior: The role of promotion focus and willingness to take risks. *Psychological Reports*. Advance online publication. doi:10.1177/00332941231172707 PMID:37092876

Ye, S., & Zhao, T. (2023). Team knowledge management: How leaders' expertise recognition influences expertise utilization. *Management Decision*, *61*(1), 77–96. doi:10.1108/MD-09-2021-1166

Zhang, J. (2021). Research on information protection system of automatic exchange of tax information in China. *Learning & Education*, *10*(4), 252–253.

Bochun Wang received a bachelor's degree in Public Administration from the Department of Public Administration, Communication University of China, Beijing, China in 2023. She is currently a postgraduate student at the Institute of Education, University College London.