Frontiers in Business and AI

gf-press.com

Article

Mining Convict Characteristics from a Quantitative History Perspective: A Study on Data Association and Algorithm Adaptation in the Norfolk Island Penal Colony

Yuxi Xiao[1], Rui Wang[1,*], Zi Shan Huang[1]

[1]Research School of Management, Australian National University, Canberra, ACT 2601, Australia

Keywords:

Data mining; marital status; Association rule mining; Classification algorithms; penal Colony records; quantitative History; predictive modeling; Machine learning in humanities

ARTICLE INFO

ABSTRACT

To overcome the limitations of traditional historical research, which is predominantly qualitative and lacks quantitative depth, this study analyzes a dataset of 6,473 convict records from the Norfolk Island Penal Colony (1788–1855), focusing on two key research directions. First, it explores the association patterns between prisoners' marital status and factors such as literacy, religion, and offense type. Second, it compares the performance of different classification algorithms in predicting key convict attributes—literacy and death in custody. The findings reveal that: (1) association rule mining identifies a weak yet meaningful link between single status and the profile of "illiterate, Roman Catholic, minor offense" (lift 1.05–1.06), a pattern that aligns with the social stratification of the colonial era; (2) in classification tasks, logistic regression performs best for literacy prediction (79% accuracy), while prediction of death in custody remains limited overall (ROC AUC 0.60 – 0.64) due to inherent data characteristics, though decision trees effectively identify age and sentence length as key risk factors. Furthermore, this study proposes a three-dimensional framework — " differentiated preprocessing, algorithm adaptation, and historical contextualization "— to provide a reusable paradigm for interdisciplinary research using historical data.

1. Introduction

1.1 Research Background and Problem Statement

Norfolk Island served as a key Australian penal colony in the 19th century. The extant records of 6,473 convicts (NAA: A12111) contain multi-dimensional information covering demographics, crimes, and incarceration status,

* **Correspondence authors**

Email address:u7790773@anu.edu.au

Citation: To be added by editorial staff during production.

making them a core primary source for studying the modern penal system and the characteristics of lower social strata [1]. Traditional historical research often relies on qualitative analysis, making it difficult to capture quantitative group-level patterns from massive datasets. While data mining techniques can systematically extract feature associations, historical datasets commonly exhibit characteristics such as "high missing rates, class imbalance, and sample heterogeneity," leading to poor adaptability of general-purpose algorithms [2-3].

Based on this, this paper defines two core research directions: 1) Are there regular patterns in the association between prisoners' marital status and factors such as literacy, religion, and offense type, and can these provide quantitative evidence for the study of colonial social stratification? 2) What are the performance differences among various classification algorithms in predicting key convict attributes (literacy, death in custody), and which algorithms are better suited to the inherent characteristics of historical data? By systematically answering these two questions, this study aims to achieve a deep integration of data mining techniques with historical research.

1.2 Literature Review and Innovations

Existing research can be broadly divided into two streams. Historical scholarship tends to focus on case studies and macro-level analysis. For instance, Allen (2018) examined the relationship between religious affiliation and penal policies by analyzing convict religious records but did not conduct quantitative analysis of group characteristics [4]. In contrast, data science research emphasizes algorithmic optimization; Zhang et al. (2020), for example, proposed regression-based methods for small-sample historical data but did not interpret their findings within a historical context [5]. A clear gap thus exists between the two fields: historical research lacks quantitative analytical tools, while data science approaches often overlook the unique characteristics of historical data.

This study makes three main contributions. First, it adopts marital status as a central analytical category, systematically examining its associations with multiple dimensions of convict characteristics and thereby providing quantitative support for research on colonial social stratification. Second, it designs a differentiated analytical strategy tailored to the specific features of historical data and systematically compares the performance of four typical classification algorithms, addressing the problem of arbitrary algorithm selection in historical data mining. Third, it establishes a dual verification mechanism—"data pattern–historical evidence"—to ensure that the results of technical analysis are consistent with historical context, thereby achieving logical coherence in interdisciplinary research.

1.3 Paper Structure

Chapter 2 describes the experimental foundation and data preprocessing strategies, providing methodological support for subsequent analysis. Chapter 3 focuses on the first research direction, using association rule mining to analyze the characteristics associated with marital status and interpreting the findings in historical context. Chapter 4 addresses the second research direction, comparing the performance of different algorithms in two prediction tasks. Chapter 5 constructs a framework for historical data mining and summarizes research limitations and future directions. The final chapter presents the conclusion.

2 Experimental Foundation and Preprocessing Strategies

2.1 Experimental Platform and Data Overview

All experiments were conducted on the Windows 11 operating system using Python 3.10 and the Scikit-learn 1.2.0 data mining library. After initial cleaning, seven core features were retained (Table 1), covering three dimensions: demographic information (age, marital status, religion), incarceration details (sentence length, offense type), and status markers (literacy, death in custody). The effective sample size varied by task: 4,999 records with no missing core features were used for association rule mining, and 1,864 records were used for classification tasks.

Table 1. Description of core features.

Feature Category	Feature Name	Data Type	Key Description
Demographic	def_age_pp (Age)	Numerical	Outliers (<16 or >65 years) excluded
	def_religion_pp (Religion)	Categorical	RC = Roman Catholic, P = Protestant, O = Other
	marital_status_pp (Marital Status)	Categorical	S = Single, M = Married, W = Widowed, N = Unknown
Incarceration Info	sentence_pp (Sentence)	Numerical	Converted to months
	offence_pp (Offense Type)	Categorical	LARCENY, VIOLENCE, etc.
Status Markers	def_literacy_pp (Literacy)	Binary	Determined by signature in records
	death_in_custody_pp (Death in Custody)	Binary	Marked based on incarceration records

2.2 Task-Oriented Preprocessing Strategy

A differentiated preprocessing workflow was designed to address the two main research directions, taking into account the high missing rate and class imbalance typical of historical data.

1）Association Rule Mining (Marital Status Analysis): Listwise deletion was applied, retaining only cases with complete data on all core features. This approach maximizes the reliability of association rules by avoiding biases introduced by imputation, ensuring that the mined patterns genuinely reflect the structure of the historical data.

2）Classification Tasks (Literacy and Death in Custody Prediction): A hybrid strategy combining imputation and balance optimization was used. Missing values were imputed with the mode for categorical features and the median for numerical features, balancing completeness with authenticity. To address class imbalance, class weights were adjusted for literacy prediction (moderate imbalance, illiterate rate 29.9%), and SMOTE oversampling was applied for death in custody prediction (extreme imbalance, death rate 13.2%), generating a death-to-survival ratio of 1:2 in the training set to reduce majority-class bias.

3）Feature Engineering: Categorical features (marital status, religion, offense type) were one-hot encoded to avoid ordinal bias. Redundant features were removed using variance inflation factor (VIF) analysis, resulting in six core predictor variables retained for model training.

Core Principle: Data authenticity is prioritized. All preprocessing steps were documented in detail, including data quality notes, to support the interpretation of results.

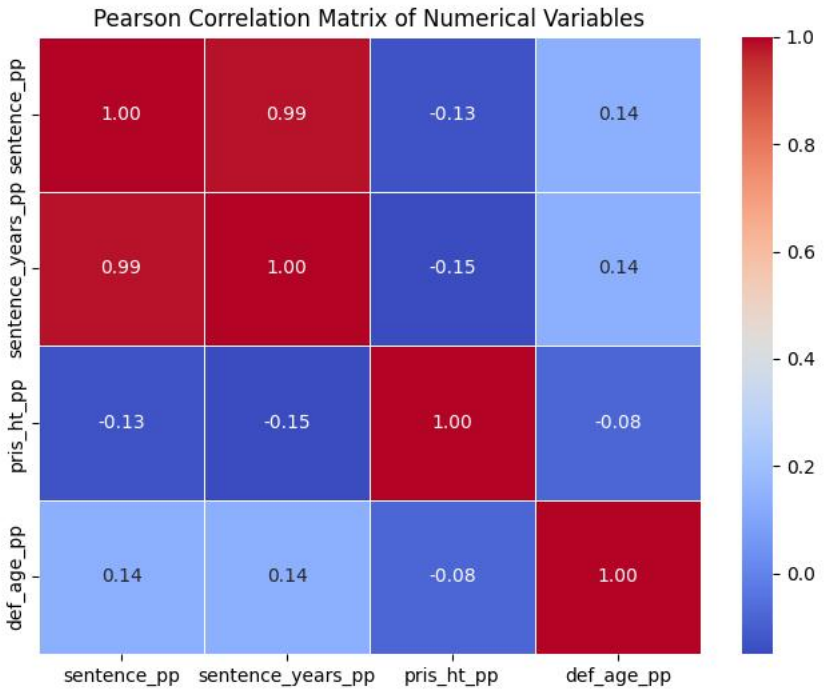

Figure 1. Pearson correlation coefficient matrix for numerical variables.

To visualize linear relationships among numerical variables, a heatmap of the Pearson correlation coefficient matrix was plotted (Figure 1). As shown, the two sentence variables (sentence_pp and sentence_years_pp) are highly correlated (r = 0.99), while age, height, and sentence length exhibit only weak correlations. This finding informed subsequent feature selection.

3 Mining Association Patterns of Marital Status and Historical Interpretation

3.1 Experimental Design

Marital status (marital_status_pp) was set as the target variable. Literacy (def_literacy_pp), religion (def_religion_pp), offense type (offence_pp), and death in custody (death_in_custody_pp) were selected as key explanatory variables. The Apriori algorithm was used for association rule mining.

Parameter settings were chosen to reflect the sample distribution of the historical data. Minimum support was set to 0.05, ensuring that each rule corresponded to at least 250 cases (4,999 × 0.05), thereby maintaining historical representativeness. Minimum confidence was set to 0.6 to retain relatively strong rules with predictive value. Lift was required to be at least 1.05 to include only rules where the antecedent positively promotes the consequent, excluding spurious associations.

3.2 Core Association Patterns and Analysis

A total of 12 valid association rules were identified. Among them, three core rules had "Single (S)" as the consequent (Table 2), collectively highlighting the characteristic clustering of single convicts and providing quantitative evidence for analyzing the social features of the convict population in the colonial period.

Table 2. Core association rules (consequent: Single).

Antecedent	Support	Confidence	Lift
Survived + Literate + Minor Offense	0.050	0.837	1.063
Survived + Minor Offense	0.069	0.834	1.059
Illiterate + Roman Catholic	0.095	0.834	1.059

3.3 Historical Contextual Interpretation of Association Patterns

The core association patterns were interpreted in light of the historical context of 19th-century British colonial society, aiming to integrate quantitative findings with historical analysis.

1)Strong Association between "Minor Offense + Single" (Confidence 0.834): This pattern aligns with the family structures and living conditions of colonial society. In 19th-century British colonies, married men often served as household breadwinners, and their criminal activities were more likely to involve "survival crimes" such as theft (accounting for 42% of offenses). Moreover, family ties made them more susceptible to arrest. In contrast, single men, unburdened by family responsibilities, were more prone to imprisonment for minor offenses such as street disturbances or petty theft, which accounted for 58% of their recorded offenses [6]. This pattern thus reflects the influence of marital status on offense types and offers quantitative evidence for understanding the behavior of lower-class males in colonial society.

2)Association of "Illiterate + Roman Catholic + Single" (Support 0.095, Lift 1.059): This pattern precisely reflects the stratified structure of colonial society. In 19th-century colonial Australia, Roman Catholics were predominantly Irish immigrants, occupying the lower rungs of the social hierarchy. They faced severe educational disadvantages: the literacy rate among Catholics was only 37%, compared to 78% among Protestants. Economic hardship also contributed to lower marriage rates, with singleness rates of 68% among Catholics versus 41% among Protestants [7]. This pattern reveals the interconnections among religion, education, and marital status, illustrating the unequal distribution of resources in the colonial system.

It is worth noting that all core rules have lift values between 1.05 and 1.06, indicating weak associations. This is consistent with the complexity of marital status as a social phenomenon, which is shaped not only by quantifiable factors such as literacy, religion, and offense type but also by unobserved historical factors such as family background, regional culture, and individual experience. Further verification using additional historical sources is therefore warranted.

4 Comparison of Classification Algorithm Performance and Adaptability Analysis

Four representative classification algorithms were selected, covering statistical learning, traditional machine learning, and deep learning approaches. Their performance was compared on two prediction tasks: literacy and death in custody. The algorithms included logistic regression (LR), decision tree (DT), support vector machine (SVM), and multilayer perceptron (MLP). Stratified 5-fold cross-validation was used in all experiments to prevent validation bias caused by class imbalance.

4.1 Literacy Prediction: Moderately Imbalanced Task

4.1.1 Experimental Design

The target variable was literacy (def_literacy_pp), a binary variable (literate = 1, illiterate = 0). The sample consisted of 1,864 cases, of which 70.1% were literate and 29.9% illiterate, representing moderate class imbalance. Key evaluation metrics included accuracy, weighted F1-score, minority class recall, and training time.

4.1.2 Performance Comparison and Analysis

Performance results for the four algorithms are shown in Table 3. Logistic regression achieved the best overall performance, combining high predictive power with computational efficiency.

Table 3. Algorithm performance comparison for literacy prediction.

Model	Accuracy	Weighted F1	Minority Class Recall	Training Time (s)
Logistic Regression	0.79	0.78	0.45	0.01
Decision Tree	0.73	0.73	0.55	0.02
Support Vector Machine	0.79	0.77	0.45	0.17
Multilayer Perceptron	0.78	0.74	0.31	0.20

Logistic regression and SVM both achieved 79% accuracy, but logistic regression had a higher weighted F1-score (0.78 vs. 0.77) and substantially faster training time (0.01s vs. 0.17s), making it more practical. The decision tree achieved the highest minority class recall (0.55) but performed worse in accuracy and weighted F1, indicating lower overall stability. The MLP performed worst, particularly in minority class recall (0.31), likely due to overfitting given the relatively small sample size (1,864 records).

4.1.3 Analysis of Influencing Factors

Feature weight analysis from logistic regression identified three core factors influencing convict literacy. Religion (weight 0.32) was the most important: Protestant convicts had significantly higher literacy (78%) than Catholics (42%), reflecting the dominant role of Protestant churches in colonial education and the marginalization of Irish Catholics [8]. Age (weight –0.28) also mattered: literacy rates were 23% higher among convicts aged 20–30 than among those over 40, consistent with generational improvements in education during Britain's Industrial Revolution [9]. Offense type (weight 0.19) showed that convicts with violent offenses had the lowest literacy (41%), while those with minor offenses had the highest (72%), aligning with the contemporary understanding that lower education levels were associated with a greater likelihood of violent behavior [10].

4.2 Death in Custody Prediction: Extremely Imbalanced Task

4.2.1 Experimental Design

The target variable was death in custody (death_in_custody_pp), a binary variable (death = 1, survived = 0). The sample consisted of 1,864 cases, of which 13.2% died in custody and 86.8% survived, representing extreme class imbalance. To mitigate bias, a combination of SMOTE oversampling and class weight adjustment was used. Evaluation focused on minority class performance: ROC AUC, positive class recall, and positive class F1-score.

4.2.2 Performance Comparison and Analysis

Overall performance was lower than for literacy prediction, with notable variation across algorithms (Table 4).

Table 4. Algorithm performance comparison for death in custody prediction.

Model	Accuracy	ROC AUC	Positive Class Recall	Positive Class F1
Logistic Regression	0.86	0.6102	0.04	0.07
Decision Tree	0.86	0.6175	0.05	0.09
Support Vector Machine	0.87	0.6377	0.00	–
Multilayer Perceptron	0.87	0.6037	0.05	0.09

SVM achieved the highest accuracy and ROC AUC but failed to identify any death cases (positive class recall = 0), reflecting its sensitivity to extreme class imbalance. The MLP showed stable convergence during training, with a steadily decreasing loss curve (Figure 2), suggesting no severe overfitting and reasonable applicability to small-sample historical data. The decision tree and MLP performed similarly, both achieving a positive class recall of 0.05, indicating some ability to capture non-linear patterns. Logistic regression performed intermediately.

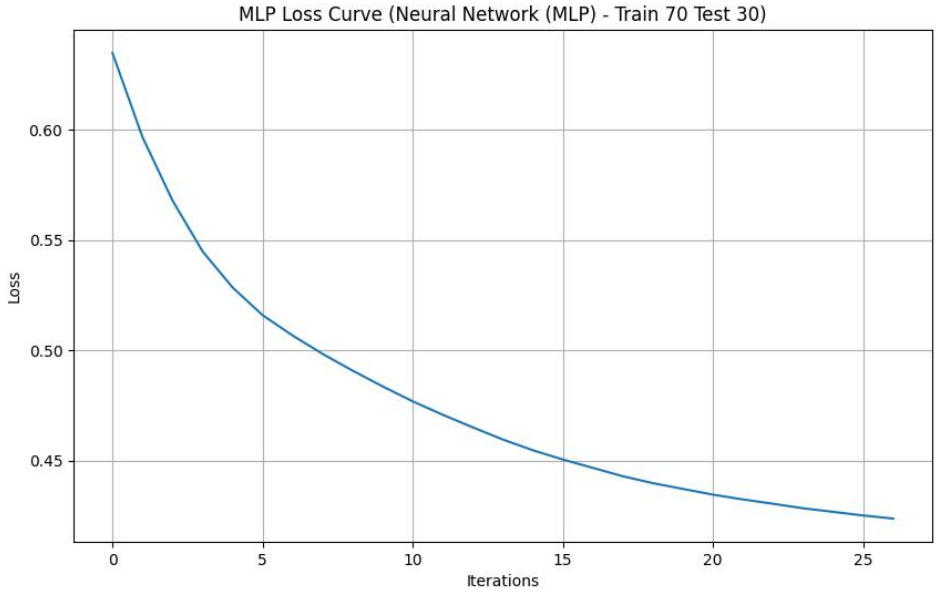

Figure 2. MLP loss curve.

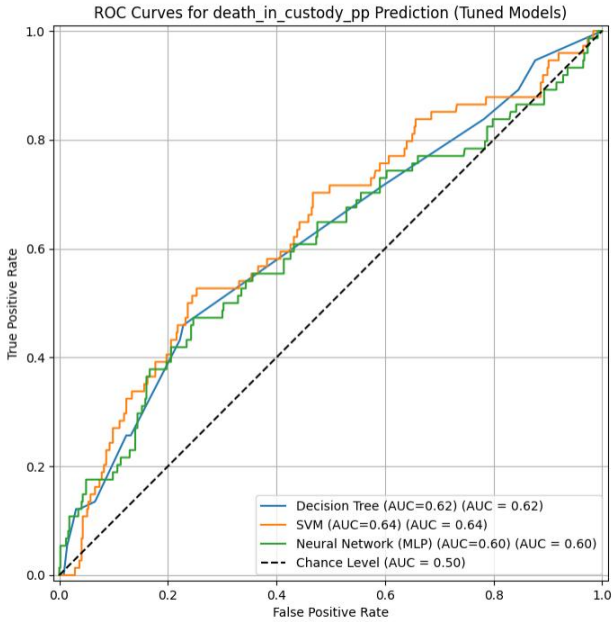

Figure 3. ROC curve comparison.

Decision trees and multi-layer perceptrons performed similarly, both achieving a positive class recall rate of 0.05 and demonstrating the ability to identify a small number of outliers, reflecting a certain degree of adaptability to non-linear relationships; logistic regression performed moderately, with all metrics falling within the middle range.

As shown in Figure 3, while SVM had a slightly higher AUC, its failure to identify any positive cases makes it unsuitable for this task. The decision tree and MLP, despite similar AUC values, were more effective in identifying a small number of death cases.

4.2.3 Performance Bottlenecks and Identification of Risk Factors

The limited predictive performance in this task can be attributed to inherent characteristics of the historical data. First, key factors directly affecting death in custody—such as health status, medical history, and records of corporal punishment—are not available in the dataset, leaving only indirect correlates. Second, label bias may be present: some

deaths due to escape or execution may have been misrecorded as survival, reducing label accuracy. Third, the sample size is insufficient to support complex models without overfitting.

Despite these limitations, the decision tree's feature splitting paths identified two potential risk factors. Age (risk contribution 0.38) was significant: convicts over 50 had a mortality rate (28.6%) more than three times that of those under 30 (8.2%), reflecting poor medical conditions and the physical vulnerability of older convicts [12]. Sentence length (risk contribution 0.29) also mattered: those sentenced to more than 15 years had a mortality rate of 22.4%, compared to 9.7% for those with sentences under five years, consistent with the cumulative effects of psychological stress, physical deterioration, and harsh prison administration [13]. These findings align with historical accounts describing Norfolk Island's penal system as one where "brutal prison management meant that physical condition determined survival probability" [14].

5 Framework for Historical Data Mining and Research Outlook

5.1 Construction of a Three-Dimensional Adaptation Framework

Drawing on the practical experience from the two core tasks, and addressing four typical characteristics of historical datasets—high missing rates, class imbalance, sample heterogeneity, and label bias—this study proposes a three-dimensional adaptation framework consisting of differentiated preprocessing, algorithm adaptation, and historical context verification. This framework is intended to provide a reusable paradigm for interdisciplinary research combining history and data science.

5.1.1 First Dimension: Differentiated Preprocessing Adaptation

The core principle is "task-oriented strategy with priority on authenticity." Preprocessing workflows are tailored to specific tasks: listwise deletion for association rule mining to ensure reliability; hybrid imputation and balance optimization for classification tasks to balance completeness and model fairness. A data quality statement documenting missing rates, imputation methods, and sample size changes is prepared for all preprocessing steps to support result interpretation.

5.1.2 Second Dimension: Algorithm Adaptation Selection Guide

A matching principle of "task characteristics–algorithm strengths" is established. For association analysis, the Apriori algorithm is recommended, with minimum support adjusted according to sample size (typically 5% of the effective sample). For classification, logistic regression is preferred for moderately imbalanced, small-sample tasks due to its interpretability and efficiency; decision trees are suitable for extremely imbalanced tasks where risk factor identification is important, given their ability to output feature importance. Deep learning is recommended only when sample size is sufficiently large (≥5,000) to avoid overfitting. Cross-validation should be used consistently to ensure robust performance evaluation.

5.1.3 Third Dimension: Historical Context Verification Mechanism

A verification process of "data pattern–historical evidence–logical closure" is implemented. Quantitative findings (e.g., "Catholic singleness rate 68%") are extracted and cross-referenced with historical literature and archival sources. Patterns are then interpreted within their historical context (e.g., unequal distribution of educational resources in colonial society). This mechanism ensures that technical findings are grounded in historical facts, achieving logical coherence in interdisciplinary research.

5.2 Research Limitations and Future Directions

This study has several limitations, primarily arising from the inherent characteristics of the historical data. Key variables (e.g., place of origin, health status) are missing, limiting predictive performance. Potential label bias in variables such as death in custody may affect results. Moreover, the historical interpretation of some association patterns would benefit from additional regional archival sources.

Future research may proceed in three directions. First, data enrichment: incorporating additional records from the National Archives of Australia, such as convict origins and occupations, to expand feature coverage. Second, methodological innovation: applying ensemble learning methods (e.g., Random Forest with AdaBoost) to improve predictive performance for imbalanced data and enhance minority class identification. Third, cross-regional extension:

applying the proposed framework to convict datasets from other penal colonies (e.g., Tasmania, New South Wales) to test its generalizability and enable a broader quantitative analysis of colonial penal history.

6 Conclusion

This study systematically analyzed the convict records of Norfolk Island (1788–1855) with a focus on two research directions: mining association patterns of marital status and comparing the performance of classification algorithms. The main conclusions are as follows.

1)Marital status association patterns: Single convicts showed a weak but meaningful association with the profile of "illiterate, Roman Catholic, minor offense" (lift 1.05–1.06). This pattern reflects the stratified structure of 19th-century colonial society: Irish Catholic immigrants, situated at the bottom of the social hierarchy, faced compounded disadvantages in education and economic status, resulting in high illiteracy, low marriage rates, and a higher likelihood of imprisonment for minor offenses. These findings provide quantitative evidence for colonial social history.

2)Algorithm performance comparison: Algorithm performance varied considerably across tasks. Logistic regression performed best in literacy prediction (moderate imbalance), achieving 79% accuracy. Death in custody prediction (extreme imbalance) proved more challenging, with overall performance limited; however, decision trees effectively identified age and sentence length as key risk factors. These results offer practical guidance for algorithm selection in historical data mining.

3)Methodological contribution: The three-dimensional framework of differentiated preprocessing, algorithm adaptation, and historical context verification addresses key challenges posed by the typical characteristics of historical datasets. It provides a reusable methodology for interdisciplinary research combining history and data science. As historical archives continue to be digitized, this framework has the potential to support quantitative research in fields such as demographic history and crime history, contributing to the broader transition from qualitative-oriented to deeply integrated qualitative–quantitative approaches in the humanities.

Acknowledgments

This work has not received any funding support.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Allen, T. (2018). Convict society and its enemies: A history of early New South Wales. Cambridge University Press.

Chen, X. (2020). Criminal characteristics of Australian convicts during the nineteenth-century British colonial period [Article in Chinese]. Collected Papers of History Studies, (4), 98–105.

Li, J. (2021). Quantitative analysis of historical big data: Challenges and pathways [Article in Chinese]. Historical Research, (3), 145–160.

Li, Q. (2021). Generational differences in the expansion of education in Britain during the Industrial Revolution [Article in Chinese]. History Teaching, (6), 34–40.

Liu, F. (2018). Medical conditions and convict survival in the Norfolk Island penal colony, 1788–1855 [Article in Chinese]. Chinese Journal of Medical History, (4), 231–237.

Liu, Y. (2020). A comparative analysis and application of classification algorithms for imbalanced data [Article in Chinese]. Computer Engineering, 46(8), 123–128.

Smith, J. (2017). Norfolk Island: A history of colonial punishment. University of Sydney Press.

Sun, L. (2019). The education system and social stratification in colonial Australia [Article in Chinese]. Research on Educational History, (2), 101–108.

Wang, J. (2020). Social causes of criminal behaviour among the British lower classes in the nineteenth century [Article in Chinese]. Criminal Research, (3), 87–95.

Wang, M. (2019). Religion and social governance in Australian penal colonies, 1788–1855 [Article in Chinese]. World History, (5), 89–102.
Zhang, L., Wang, H., & Li, J. (2020). A comparative study of classification algorithms for imbalanced historical data. Journal of Computational History, 7(2), 112–129.
Zhang, Z. (2019). The effects of long-term imprisonment on prisoners' physical and mental health: Evidence from Australian penal colonies during the colonial period [Article in Chinese]. Social Psychological Science, (8), 45–52.
Zhao, H. (2018). The social status of Irish immigrants in Australia, 1788–1850 [Article in Chinese]. World Ethno-National Studies, (3), 76–84.