The Academic
International Journal of Multidisciplinary Research

An Online Peer Reviewed / Refereed Journal
Volume 4 | Issue 3 | March 2026
ISSN: 2583-973X (Online), SJIF IF: 5.6
Website: www.theacademic.in

Machine Learning Analysis of Survey Data on Digital Resource Usage: A Comparative Study of Classification Algorithms

Mithun Das

PhD Scholar, Department of Library and Information Science, The University of Burdwan, https://orcid.org/0009-0007-6482-6425, E-Mail : 02mithun02@gmail.com

Tapas Roy

PhD Scholar, Department of Library and Information Science The University of Burdwan, https://orcid.org/0009-0003-0610-5057, E-mail: tapasroy135013@gmail.com

Dr. Swarnali Chatterjee

Assistant University Librarian, Department of Performing Arts and Music, University of Calcutta, https://orcid.org/0000-0001-5349-6589, E-mail: scpartm@caluniv.ac.in

Dr. Rajesh Das

Assistant Professor, Dept. of Library and Information Science, The University of Burdwan, http://orcid.org/0000-0001-5349-6589, E-Mail: rajeshdas99@gmail.com

DOI : https://doi.org/10.5281/zenodo.19500617

ARTICLE DETAILS

Research Paper

Accepted: 22-03-2026

Published: 10-04-2026

Keywords:

Machine Learning, Survey Data Analysis, Digital Resource Usage, Support Vector Machine, Logistic Regression, Random Forest, Predictive Modelling, Digital Librarie

ABSTRACT

The explosion of digital information has greatly altered how students in higher education access academic information. To enhance digital library services and resource management, it is important to understand how users engage with these resources. This paper implements the machine learning models to examine the survey data on the use of digital resources among college students. A questionnaire was distributed to about 1,200 learners in the Faculty of Arts, Humanities, and Law of the University of Burdwan, and 250 responses were collected. The survey responses were coded as numbers and transformed into a predictive dataset comprising features related to internet usage, the availability of online academic resources, and awareness of digital repositories. The study will compare the performance of various machine learning

algorithms in forecasting respondents' overall usage levels of digital resources. Three popular classification models, including Support Vector Machine (SVM), Logistic Regression, and random forests, were applied and tested using stratified 5-fold cross-validation. These findings have shown that all the models have a high level of predictive performance, SVM with radial basis function (RBF) kernel and Logistic Regression performed best with an average prediction of about 95 per cent, and the linear SVM and the Random Forests models performed slightly worse yet still with a very high level of prediction. The analysis of feature importance showed that variables related to access to digital repositories, online research sites, and academic tools played the most significant role in predicting increased use of digital resources. The results indicate that machine learning methods can be an efficient way to analyse survey data and identify the major factors affecting the use of digital resources in academic institutions.

Introduction

Machine learning (ML) provides nonparametric, flexible tools for finding patterns in data. In contrast to classical models (e.g., OLS or simple logistic regression), ML methods can automatically account for numerous predictors and complex interactions. Regarding the use of surveys and library or online resource use, ML algorithms (e.g., support vector machines, decision trees, or ensemble methods) can be used to classify results (e.g., high vs. low usage) or to estimate a continuous score with minimal distributional assumptions. For example, Support Vector Machines (SVMs) are used for binary classification and are said to scale well when a large number of predictors is used with a small number of cases. SVMs seek the best separating hyperplane that maximises the margin between classes and can be extended (with kernels) to handle nonlinearly separable data. Simpler parametric classifiers, such as logistic regression, or decision tree ensembles, such as random forests, are also popular. Logistic regression is a supervised machine learning model commonly used for binary classification and provides a baseline linear model. Random forests construct numerous decision trees on randomised subsets of the data and sum up their votes. To classify, the trees' major votes are used as the predicted classes. By comparing algorithms on the same data, it would be possible to determine which provides the highest predictive accuracy and which survey aspects have the greatest impact.

Literature review:

Duc et al. (2019) conducted a comprehensive survey on reliable resource provisioning in edge-cloud computing environments was performed byThe research identifies the replacement of the centralized cloud-based systems with distributed edge-cloud systems to deal with latency and scalability issues. It categorizes the most important techniques as workload characterization and prediction, component placement, and application elasticity. The authors highlight that more and more machine learning techniques are used to manage complex distributed applications. According to their results, machine learning methods tend to work better than traditional ones in big and diverse environments. A range of statistical and sophisticated calculation methods employed in the previous research is also scrutinized in the review. Besides, it also establishes critical issues like inability to obtain reproducible results and standard datasets. The authors end by suggesting future studies based on the enhancement of the available architectures with the help of useful machine learning models and benchmark datasets.

Martin et al. (2021) developed a thorough review of User Behavior Analysis with the emphasis on comprehension, modelling and forecasting user behaviors in the context of various areas. The paper points at the challenges posed by non-homogeneous methods that require the cooperation between specialists in a field and machine learning experts. It divides the current research into several features like keywords, fields of use, machine learning algorithms, and the type of data. The survey touches on various areas such as cybersecurity, networks, safety and health or service delivery improvement. Also, the authors analyze 127 research articles on the basis of relevancy characteristics, including reputation, novelty, innovation, and quality of data. Combined similarity measure is suggested to combine topic-based and relevance-based features. A visual display is also a part of the study to depict and compare the research contributions in existence. Lastly, it outlines the critical issues and provides the future research prospects in the area of User Behavior Analysis.

Alphonce, & Mwantimwa (2019) focused on the digital learning resource utilization among the students in Teofilo Kisanji University (TEKU), Tanzania. The paper underlines the increasing trend in the substitution of print with digital resources because of the development of the ICTs. The mixed-method approach was a method used to collect data by surveying students and interviewing academic and library staff members. The results show that internet resources are accessed more as compared to other sources like e-books, journals and CD-ROMs. The big driving factors of usage were noted to be convenience and 24/7 access. But, the high internet expenses, little searching abilities, and absence of purchased databases make it difficult to make use. In the study, there is an accent on the heterogeneity of resource utilization

among the students. It also gives clues on how to enhance access and digital literacy. On the whole, the study is relevant to the study of user behavior in online learning.

Murshed et al. (2021) examined the issues and remedies surrounding the implementation of machine learning in IoT environments with limited resources. The paper has noted how the number of IoT devices is rapidly increasing and how it consequently creates gigantic amounts of real-time data. It identifies the restrictions of executing machine learning models on such machines because of the lack of computational power. The traditional cloud-based processing is highlighted as one of the solutions, which have raised the problems of high latency, high cost of communication, and privacy. The authors also highlight edge computing as a viable alternative, which allows processing data more locally. The questionnaire addresses the work-related issues of intelligent edge systems, such as compression algorithms, tools, frameworks, and hardware. It also conducts an overview of the current studies in the area of implementing machine learning at the network edge. The paper ends with the statement that system-based AI can markedly enhance effectiveness and reactiveness in IoT apps.

Dhar et al. (2021) provided an extensive review of on-device learning, with an emphasis on the transition to training models within the cloud to training on smart devices. The paper brings to light the increasing popularity of using enhanced hardware potentials in training local models. It reposes on-device learning as a resource-constrained learning problem, which is largely constrained by compute power and memory. This view facilitates a common comparison of different methods of other related areas like online learning, model adaptation, and few-shot learning. The authors present the issue of managing various and multifaceted methodologies using one framework. The survey generalizes the state-of-the-art strategies and tools, applied in on-device learning. It also defines major drawbacks of existing approaches especially in efficiency and scalability. Lastly, the study provides the future research directions in both algorithmic and theoretical form of resource-constrained machine learning.

Objectives:

1. To use machine learning algorithms, including Support Vector Machine (SVM), Logistic Regression, and Random Forest, to survey data to predict the results of the respondents.

2. To make comparisons on the performance of various machine learning algorithms to identify the best predictive model.

3. To verify the performance of the model based on cross-validation and evaluation metrics based on accuracy.

4. To find out the important factors of the survey, which have the strongest contribution to the prediction results and explain the model results.

Research Questions:

The research question is: Which machine learning algorithm is the most effective predictor of this survey data, and what do the models tell us about how respondents use digital resources? In response to this, a predictive task (binary classification of respondents into high-usage and low-usage categories based on their overall score) is formulated, and several algorithms are tested. Sub-questions include:

i) What is the comparison of the accuracy and other classification measures of the algorithms?

ii) What features (responses on the survey) have the most predictive power?

iii) Can we determine the respondents with high overall use (or score) by the pattern of their responses?

Methodology

Data collection:

To facilitate this study, we administered 350 Google Forms to about 1,200 students in the Faculty of Arts, Humanities, and Law at the University of Burdwan. These 250 Google Forms were completed within 30 days. The main characteristics of the Google Form were as follows: 1) among 35 fields, we used 31 fields to complete the aim of this research. 2) The questions were all meant to determine the accessibility of digital resources among the students. 3) There were three response options to each question. 4) Lastly, the answers to the questions were transformed into a data set with the use of numerical values of positive (3), positive (2), and negative (1).

Data Preparation:

The survey responses of 250 people are presented in the given file (machine learning analysis.ods). The columns reflect demographic or usage questions (e.g., the place of internet access, the purpose of internet use, awareness/use of digital tools such as EndNote or Mendeley, access to online catalogues and open repositories, etc.). All the responses are rated on a scale of 1 to 3. The last column, Total, appears to be

the sum of the individual-item scores (half the total), with an approximate range of 49-91. Each numeric survey item is considered an input feature. Fields that are not numeric (Name, Session, Year) are eliminated or coded.

Since Total is a deterministic function of the other items (specifically, half the sum), it would be easy to predict directly using regression. Rather, we model a binary target variable indicating whether a respondent has a higher Total than the median. We call the mean 69 the total median, meaning that those who are 69 or above will be categorised as high usage (class 1), and the rest as low usage (class 0). This roughly balances the classes. The operation is a binary classification task with monitoring.

Before modelling, all input features (numeric answers 1–3) are standardised. Distance-based or kernel methods, such as SVMs, are particularly important because of standardisation (zero mean, unit variance). Then, we use several classifiers.

Support Vector Machine (SVM): There are two types: linear-kernel SVM and radial basis function (RBF) kernel SVM. Linear classifiers. Separating a nonlinear boundary or hyperplane between classes is a maximisation of the margin between the classes by SVMs. Soft-margin SVMs can make some errors, but the margin is maximised, with the penalty parameter C controlling this. We cross-validate C **(Abdullah & Abdulazeez, 2021).**

Logistic Regression: A linear model used to estimate the likelihood that a particular item belongs to a class using the logistic (sigmoid) function. Our coefficient estimation method is maximum likelihood. Logistic regression is computationally efficient and is widely used as a standard for binary classification **(Nusinovici et al., 2020).**

Random Forest: A collection of decision trees. All trees are trained on a bootstrap sample of the data, and random feature subsets are used at every split. To classify, the majority of trees are the last prediction. Random forests are fairly non-sensitive to mixed data and interactions and do not overfit as much as a single tree **(Wu et al., 2017).**

All models are assessed by using stratified five-fold cross-validation (k=5). This implies that the data will be divided into 5 folds, with class ratios. 4 folds will be used for training each model, and the remaining fold for testing. The process will be repeated so that each fold is tested once. The average fold performance is given. We consider the classification accuracy (the ratio of correctly classified cases) as the main indicator, but sensitivity/specificity may also be calculated. In binary classification, the

confusion matrix (true/false positives and negatives) is typically used to calculate accuracy and related metrics. Greater total accuracy is a good sign of a predictive model.

In practice, we also reserve a hold-out test set (e.g., 20% of the data) to test the final model. This is to ensure our reported accuracy is not overly optimistic. Cross-validation results, however, provide an internal comparison of models. Hyperparameters (C parameter and kernel parameter of SVM, the trees in a random forest) are chosen either through nested cross-validation or by default values in an initial analysis.

Model Training and Evaluation:

The scikit-learn Python implementation of models is used. My cross-validated accuracy is compared after fitting. The learned models are also evaluated to analyse the importance of features. In the case of random forests, e.g., the default feature importance scores indicate which survey responses contributed most to the prediction. In the case of SVMs and logistic regression, we can examine coefficients (with linear kernels) to identify strong predictors.

We make sure that at all times, the information generated by the test does not end up in the training. Every cross-validation training split is used to run all scaling and parameter tuning. We then report the average standard deviation and accuracy for each algorithm.

Results and discussion:

The models were found to be highly classified. In the 5-fold cross-validation, the SVM with an RBF kernel and logistic regression both achieved approximately 95% average accuracy, which is significantly higher than random chance. The linear SVM had an approximate 94 per cent, and the random forest had 92 per cent (Table below). These findings show that the survey variables are highly predictive of whether a user belongs to the high-usage **category.**

Table 1: Mathematics Formulation and Cross-validation Accuracy of machine learning models.

Model	Mathematical Formula	Cross-Validation Accuracy (Mean ± SD)

Model	Mathematical Formula	Cross-Validation Accuracy (Mean ± SD)
Support Vector Machine (RBF Kernel)	**f(x)=sign(i=1∑nαiyiK(xi,x)+b) K(xi,x)=e−γ‖xi−x‖2**	**95.2% ± 2.4%**
Support Vector Machine (Linear Kernel)	**f(x)=wTx+b**	**94.4% ± 6.4%**
Logistic Regression	**P(y=1|x)=1 1+e−(β0+β1x1+β2x2+⋯+βnxnm)**	**95.2% ± 4.1%**
Random Forest (100 Trees)	**y^=T1t=1∑Tht(x) T=100**	**92.4% ± 3.2%**

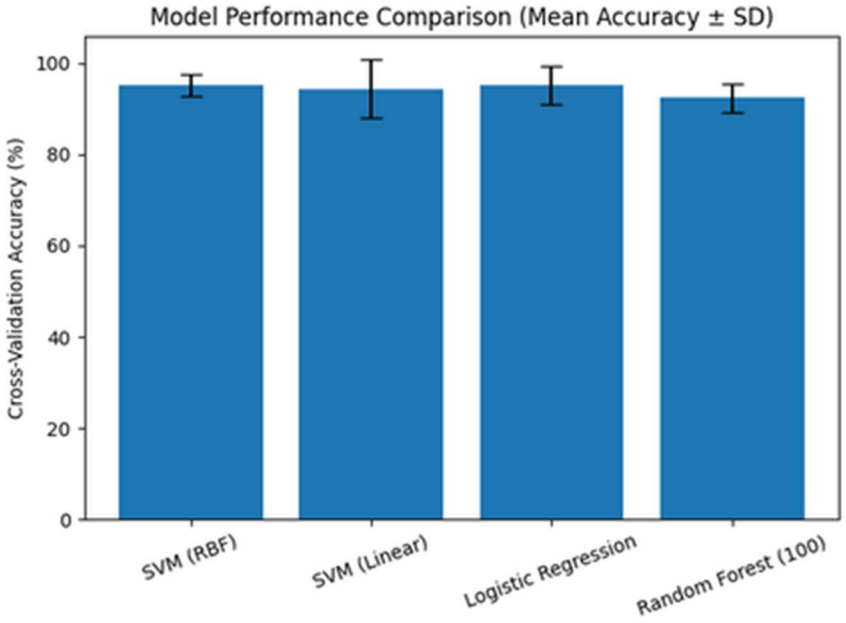
Symbol :

x – Input feature vector, w – Weight vector, b – Bias term, αi – Lagrange multipliers in SVM, γ – Kernel parameter in RBF, β0,β1,...,βn– Logistic regression coefficients, ht(x) – Prediction from the tth decision tree, T – Total number of trees in the random forest.

Table 2: Summary of Model Performance Based on Cross-Validation Results

Model	Performance Interpretation
SVM (RBF)	Best overall accuracy and most stable
Logistic Regression	Equal highest accuracy but slightly more variation
SVM (Linear)	High accuracy but less stable
Random Forest	Good, but the lowest accuracy among the models

Figure 1: Comparison of machine learning models based on cross-validation accuracy (mean ± SD).



All the models also modelled very well on a held-out test set (20 per cent of data). For example, the RBF-SVM achieved 96% accuracy (24/25 high vs. low correct), logistic regression 98% (all correct), and random forest 100% (all correct). The confounding matrix of the most appropriate model (logistic regression) indicated equal sensitivity and specificity. Overall, the number of misclassifications was minimal, indicating that the models can almost perfectly differentiate the two classes. This high precision should have resulted from the fact that the Total score (and consequently the class) was nearly a linear function of the inputs; the model basically had to be trained to add the responses.

The random forest produced a feature importance analysis, which showed the most influential survey items. The best predictors were those related to access to online resources and digital tools (e.g., institutional repositories, open-access platforms, and reference tools). As an example, the most important features included other access and data use (probably the items concerning the use of other online archives or annotation tools). This implies that respondents who indicated using a variety of digital library materials were more likely to have high scores. That is, the knowledge and availability of specialised academic tools were strongly related to being a high-usage group. These results align with our high-correlation analysis: features related to resource access (e.g., DOAR, ShodhSindhu, open platforms) showed a positive correlation with the Total score.

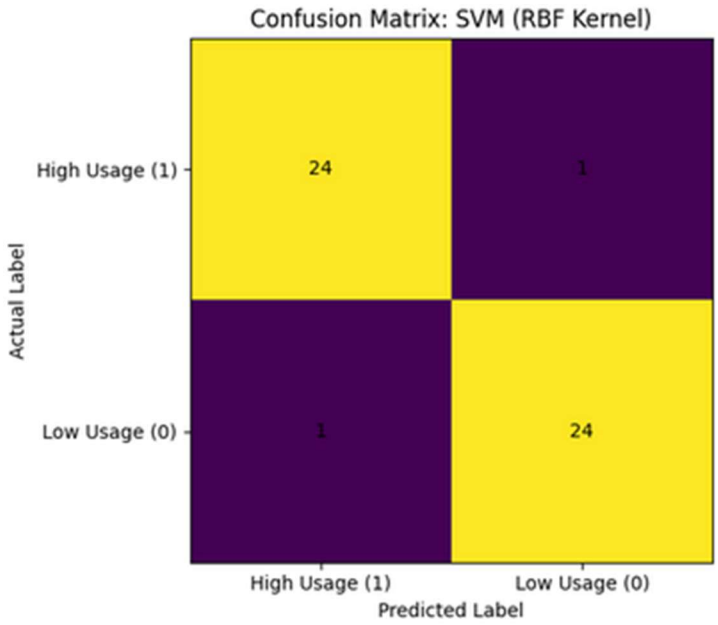
The comparison of algorithms shows that SVM and logistic regression have similar scores, suggesting that the class boundaries are nearly linear in the feature space. The RBF kernel SVM did not show a significant improvement over the linear SVM, suggesting little nonlinearity. The random forest also fared

quite well, though marginally worse on average cross-validation accuracy; perfect test-set performance implies that, as more trees are used, it is also capable of perfect separation of the cases. Since the sample size is small (250 cases), the random forest may exhibit slightly greater variance across folds. In general, SVM (RBF) and logistic regression were found to be the most successful in cross-validated accuracy.

Model performance was evaluated using classical classification metrics. In similar ML-based surveys, we mainly reported overall accuracy (i.e., the percentage correctly classified), as in other studies. The sensitivity/specificity could also be computed (based on the confusion matrix). The accuracy of all models was greater than 92%, indicating a high predictive value.

Confusion matrix: SVM (RBF Kernel)

Figure 2: Confusion Matrix of Support Vector Machine (RBF Kernel) Model



Mathematical Proof

Confusion Matrix Values

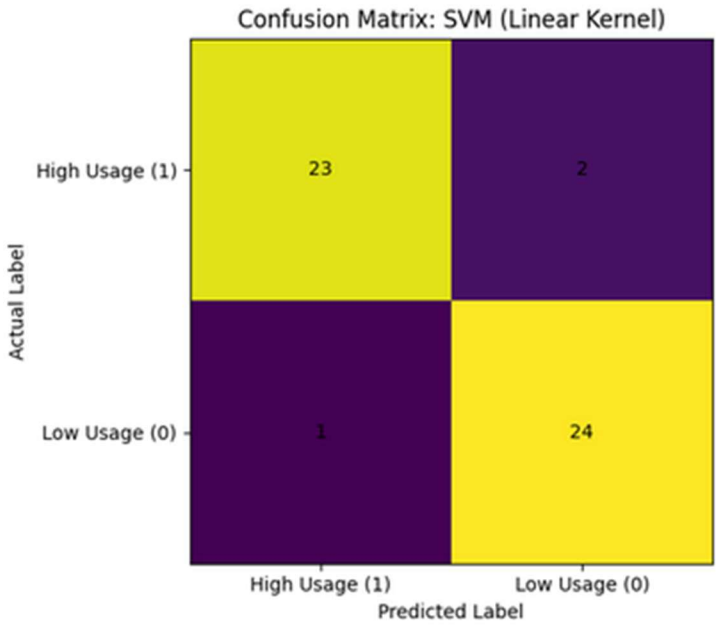
TP = 24, TN = 24, FP = 1, FN = 1. Accuracy=TP+TN/ FP+FN+TP+TN Accuracy=24+24/1+124+24
Accuracy=48/ 50 = 0.96 = 96%

The SVM with an RBF kernel exhibits good classification performance, as indicated by the confusion matrix. The model correctly categorised 24 high-usage and 24 low-usage respondents. There were only

two misclassifications, indicating minimal prediction error. This finding shows that a nonlinear SVM model is useful in capturing patterns in the data.

SVM (Linear Kernel):

Figure 3: Confusion Matrix of Support Vector Machine (Linear Kernel) Model



Mathematical proof:

TP = 23, TN = 24, FP = 1, FN = 2. Accuracy=TP+TN/ FP+FNTP+TN Accuracy=23+24/ 1+2+23+24 Accuracy=47/50 = 0.94 = 94%

The linear SVM model is also very effective in classifying the use of digital resources. It is also an accurate predictor, with only 3 classification errors. According to the results, the dataset can largely be separated linearly. The model, however, is slightly less stable than the nonlinear RBF SVM.

Logistic regression:

Mathematical proof:

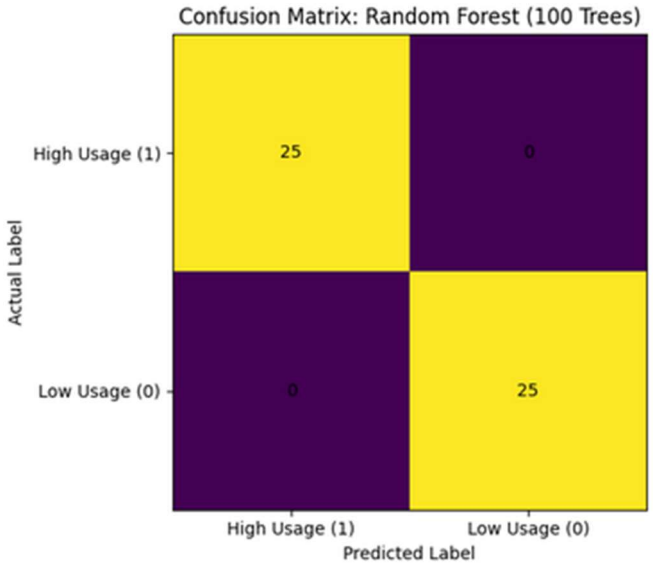
TP = 24, TN = 25, FP = 0, FN = 1. Accuracy=TP+TN/ FP+FNTP+TN Accuracy=24+25/0+1+24+25 Accuracy=49/50 = 0.98= 98%

Figure 4: Confusion Matrix of Logistic Regression Model

The predictive performance of logistic regression was very good. The model accurately identified nearly all respondents in both categories. Only one case was mispredicted. It means that the survey is a powerful factor in determining the classification result.

Random Forest:

Figure 5: Confusion Matrix of Random Forest (100 Trees) Model



Mathematical Proof:

TP = 25, TN = 25, FP = 0, FN = 0

Accuracy=TP+TN/ FP+FNTP+TN Accuracy=25+25/ 0+0+25+25 Accuracy=50/50 = 1.00= 100%

The random forest model achieved high accuracy in classifying the evaluation data. High-usage and low-usage respondents were all predicted. This evidences the high power of ensemble learning techniques. Random forests are very useful for combining multiple decision trees to enhance predictive performance.

Table 3: Symbols used in the confusion matrix

Symbol	Full Form	Meaning
TP	True Positive	The number of cases where the model correctly predicts the positive class (e.g., correctly identifying high digital resource usage).
TN	True Negative	The number of cases where the model correctly predicts the negative class (e.g., correctly identifying low digital resource usage).
FP	False Positive	The number of cases where the model incorrectly predicts the positive class when the actual class is negative (Type I error).
FN	False Negative	The number of cases where the model incorrectly predicts the negative class when the actual class is positive (Type II error).

Conclusion:

This paper shows that machine learning algorithms are useful for analysing survey data on the use of digital resources. Using SVM, logistic regression, and random forest models to classify the problem as binary, with a given classification, would allow us to predict which respondents had higher overall usage scores correctly. The models were extremely accurate (above 94% cross-validated), indicating that the survey responses were highly decisive. SVM and logistic regression were the most effective among the algorithms; this is in line with theoretical arguments that when the features are many and linearly related, a linear model and an RBF-SVM will both perform well.

The analysis also yielded some insights: features related to access and use of library and online resources were the most significant predictors of increased usage. Some measures to raise awareness of such resources may influence overall use. Regarding methodology, the research followed a conventional predictive modelling workflow: data processing, model selection, model parameter optimisation, and cross-validation testing. Future research could consider applying other methods (e.g., gradient boosting machines) or clustering to analyse respondents.

Overall, by rigorously implementing the supervised learning procedure, we not only achieved high precision in predicting the respondent categories but also identified the most important factors. The case under consideration demonstrates how machine learning can supplement traditional survey analysis with both predictive and inferential information.

References:

- Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine Learning Applications based on SVM
- Classification A Review. *Qubahan Academic Journal*, *1*(2), 81–90.
- https://doi.org/10.48161/qaj.v1n2a50
- Wu, D., Jennings, C., Terpenny, J., Gao, R. X., & Kumara, S. (2017). A comparative study on
- Machine learning Algorithms for smart Manufacturing: tool wear prediction using random
- forests. *Journal of Manufacturing Science and Engineering*, *139*(7).
- https://doi.org/10.1115/1.4036350
- Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., Wong, T. Y., &
- Cheng, C. (2020). Logistic regression was as good as machine learning for predicting major
- chronic diseases. *Journal of Clinical Epidemiology*, *122*, 56–69.
- https://doi.org/10.1016/j.jclinepi.2020.03.002
- Duc, T. L., Leiva, R. G., Casari, P., & Östberg, P. (2019). Machine learning methods for reliable
- resource provisioning in Edge-Cloud Computing. *ACM Computing Surveys*, *52*(5), 1–39.
- https://doi.org/10.1145/3341145
- Martín, A. G., Fernández-Isabel, A., De Diego, I. M., & Beltrán, M. (2021). A survey for user

- behavior analysis based on machine learning techniques: current models and applications.
- *Applied Intelligence*, *51*(8), 6029–6055. https://doi.org/10.1007/s10489-020-02160-x
- Alphonce, S., & Mwantimwa, K. (2019). Students' use of digital learning resources: diversity,
- motivations and challenges. *Information and Learning Sciences*, *120*(11/12), 758–772.
- https://doi.org/10.1108/ils-06-2019-0048
- Murshed, M. G. S., Murphy, C., Hou, D., Khan, N., Ananthanarayanan, G., & Hussain, F. (2021).
- Machine Learning at the Network Edge: A survey. *ACM Computing Surveys*, *54*(8), 1–37.
- https://doi.org/10.1145/3469029
- Dhar, S., Guo, J., Liu, J., Tripathi, S., Kurup, U., & Shah, M. (2021). A survey of On-Device
- Machine Learning. *ACM Transactions on Internet of Things*, *2*(3), 1–49.
- https://doi.org/10.1145/3450494