Jurnal Minfo Polgan
Volume 13, Nomor 1, Februari 2024
DOI : https://doi.org/10.33395/jmp.v13i1.14345
e-ISSN : 2797-3298
p-ISSN : 2089-9424

Terbit : 12 Februari 2024

K-Nearest Neighbor and Random Forest Algorithms in Loan Approval Prediction

Syafitri Ramadhani[1], M. Rhifky Wayahdi[2]
[1,2]Program Studi Sistem Informasi, Fakultas Teknologi, Universitas Battuta
[1]syafitriramadhani26@gmail.com, [2]muhammadrhifkywayahdi@gmail.com

ABSTRACT

Loan approval prediction is an important task in the financial sector, which helps banking institutions and lenders make informed decisions regarding loan applications. This research compares the performance of two machine learning algorithms, namely K-Nearest Neighbor (KNN) and Random Forest (RF), in the context of loan approval prediction. The research methodology includes data collection, pre-processing, modeling, and evaluation. The analysis results showed that the Random Forest model performed better overall than KNN, with more true positives and true negatives, and fewer false positives and false negatives. In addition, Random Forest recorded higher accuracy, precision, recall, and F1-score values. These findings provide valuable insights for financial institutions in improving credit risk management strategies and decision-making regarding loan applications.

Keyword: Machine Learning, K-Nearest Neighbor, Random Forest, Prediction, Loan.

INTRODUCTION

The prediction of loan approval is a critical task in the financial sector, as it helps banks and lending institutions make informed decisions about loan applications. In this context, the use of machine learning algorithms, such as K-Nearest Neighbor (KNN) and Random Forest (RF), has gained significant attention due to their ability to accurately predict loan approval (Aditya & Nagaraju, 2022; Sandeep & Devi, 2022). The KNN algorithm is a non-parametric method that classifies an object based on the majority vote of its k nearest neighbors. Several studies have compared the performance of KNN and RF algorithms in loan approval prediction, and the results have been mixed. Some studies have found that RF outperforms KNN in terms of accuracy (Aditya & Nagaraju, 2022; Sandeep & Devi, 2022), while others have reported that KNN performs better (Anannya et al., 2023). These conflicting findings highlight the need for a more comprehensive analysis to determine the most suitable algorithm for loan approval prediction.

The KNN algorithm is a widely used machine learning technique for both classification and regression tasks (Dondekar & Sonkamble, 2020; Gavagsaz, 2022). It is a non-parametric, instance-based learning algorithm that classifies an object based on the class of its k nearest neighbors in the training data (Dondekar & Sonkamble, 2020; Ren et al., 2021; Rosdiana et al., 2021). The working principle of KNN is to find the shortest distance between the data to be evaluated and the closest k-nearest neighbors in the training data (Rosdiana et al., 2021; Paramita et al., 2022). The class of the new data point is then determined by a majority vote of its k nearest neighbors (Dondekar & Sonkamble, 2020). KNN has several advantages, such as simplicity, clarity, and high performance (Gavagsaz, 2022; Wayahdi et al., 2020). It is also robust to noisy training data and easy to implement. However, it also has some drawbacks, such as high computational complexity, large memory requirement for large training datasets, and the curse of dimensionality (Wayahdi & Ruziq, 2022). KNN algorithm can be applied to student thesis subjects

Jurnal Minfo Polgan
Volume 13, Nomor 1, Februari 2024
DOI : https://doi.org/10.33395/jmp.v13i1.14345
e-ISSN : 2797-3298
p-ISSN : 2089-9424

(Paramita et al., 2022), non-linear indsutrial processes (Ren et al., 2021), image classification (Dondekar & Sonkamble, 2020); (Wayahdi et al., 2020), cancer prediction (Wayahdi & Ruziq, 2022), and heart disease (Rosdiana et al., 2021).

The random forest algorithm is a powerful machine learning technique that has been widely used in various fields, including medicine, computer science, and geology. It is an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of the predictions (Devi et al., 2021; Liu et al., 2020; Wayahdi et al., 2024). The random forest algorithm works by creating a large number of decision trees, each trained on a random subset of the data and a random subset of the features. One of the key advantages of the random forest algorithm is its ability to handle high-dimensional data without the need for feature selection (Shin et al., 2021). The random forest algorithm has been successfully applied in various domains, such as medical diagnosis (Devi et al., 2021; Liu et al., 2020), environmental monitoring (Akumu et al., 2021), and financial forecasting (Hota & Dash, 2021; Tong & Duan, 2022)

The present study aims to contribute to this ongoing research by providing a detailed comparison of KNN and RF algorithms in the context of loan approval prediction. The findings of this work will be valuable for financial institutions and lending organizations in making informed decisions about loan applications and improving their credit risk management strategies.

METHOD

This research was conducted with several stages of a systematic methodology to ensure that the model built can provide accurate and reliable results in prediction. The stages include data collection, data pre-processing, modelling, and evaluation. The research stages can be seen in Figure 1.

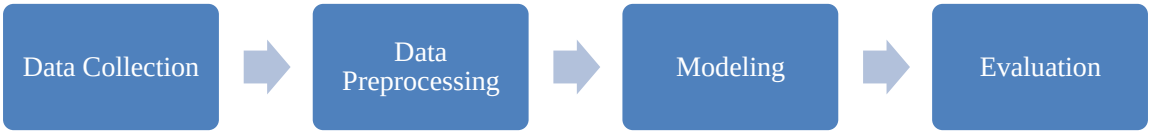

Figure 1. Research Method

Figure 1 shows the stages of the research carried out, namely:

1. **Data Collection**

   The first stage in this research is data collection. The datasets used in this research were obtained from published datasets. This dataset includes various relevant features, such as age of the person, gender of the person, highest education level, and so on as many as 13 variables with 1 label (target). The data collected includes 45,000 records reflecting various borrower characteristics. This comprehensive data collection aims to ensure that the model can learn from various scenarios and conditions that may be encountered in the loan approval process. Figure 2 shows a dataset of the top five data.

	person_age	person_gender	person_education	person_income	person_emp_exp	person_home_ownership	loan_amnt	loan_intent	loan_int_rate	loan_percent_income	cb_person_cred_hist_length	credit_score	previous_loan_defaults_on_file	loan_status
0	22.0	female	Master	71948.0	0	RENT	35000.0	PERSONAL	16.02	0.49	3.0	561	No	1
1	21.0	female	High School	12282.0	0	OWN	1000.0	EDUCATION	11.14	0.08	2.0	504	Yes	0
2	25.0	female	High School	12438.0	3	MORTGAGE	5500.0	MEDICAL	12.87	0.44	3.0	635	No	1
3	23.0	female	Bachelor	79753.0	0	RENT	35000.0	MEDICAL	15.23	0.44	2.0	675	No	1
4	24.0	male	Master	66135.0	1	RENT	35000.0	MEDICAL	14.27	0.53	4.0	586	No	1

Figure 2. Top Five Data from Dataset

2. **Data Pre-processing**

   Once the data has been collected, the next step is data pre-processing. At this stage, the data obtained is cleaned and prepared for further analysis. This process includes several steps, such as addressing missing values, removing duplicates, and converting categorical variables into numerical formats using encoding techniques, such as Label Encoding. In addition, irrelevant

Jurnal Minfo Polgan
Volume 13, Nomor 1, Februari 2024 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v13i1.14345 p-ISSN : 2089-9424

or redundant features are also removed to improve model efficiency. Data normalization is also performed to ensure that all features are on the same scale, so that no feature dominates the model training process. After the pre-processing stage is completed, the dataset is divided into two parts: training data and test data, with a proportion of 80% for training data and 20% for test data. Figure 3 shows the splitting of training and test data.

```
Training set shape: X_train=(36000, 13), y_train=(36000,)
Test set shape: X_test=(9000, 13), y_test=(9000,)
```

Figure 3. Data Splitting

3. **Modeling**
   In the modeling stage, the K-Nearest Neighbor (KNN) and Random Forest (RF) algorithms were applied to build a loan approval prediction model. The KNN model is used due to its simplicity and ability to handle unstructured data, while Random Forest is chosen for its ability to overcome overfitting and improve accuracy through combining multiple decision trees. The training process was conducted using training data, where the model was trained to recognize patterns and relationships between features and loan approval status. The model parameters are optimized using cross-validation techniques to ensure that the model can generalize well on data that has never been seen before.
4. **Evaluation**
   After the model is trained, the last stage is evaluation. The model that has been built is evaluated using test data to measure its performance. Some of the evaluation metrics used in this research include accuracy, precision, recall, and F1-score. These metrics provide a comprehensive overview of how well the model predicts loan approval. In addition, the evaluation stage is also conducted to assess the model's ability to distinguish between positive and negative classes. The results of this evaluation will be used to compare the performance of the two algorithms and determine which algorithm is more effective in predicting loan approval.

RESULT AND DISCUSSION

Figure 4 displays the correlation matrix that illustrates the relationship between the various features in the dataset used for loan approval prediction analysis. This matrix provides important insights into the interactions between features, which can influence decisions in the loan approval process. By analyzing the correlation values, we can identify significant patterns and relationships, and understand how each feature contributes to the prediction results. Correlation values range from -1 to 1, where positive values indicate a direct relationship and negative values indicate an inverse relationship. Let's drill down deeper to understand the dynamics present in this data.

Jurnal Minfo Polgan
Volume 13, Nomor 1, Februari 2024 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v13i1.14345 p-ISSN : 2089-9424

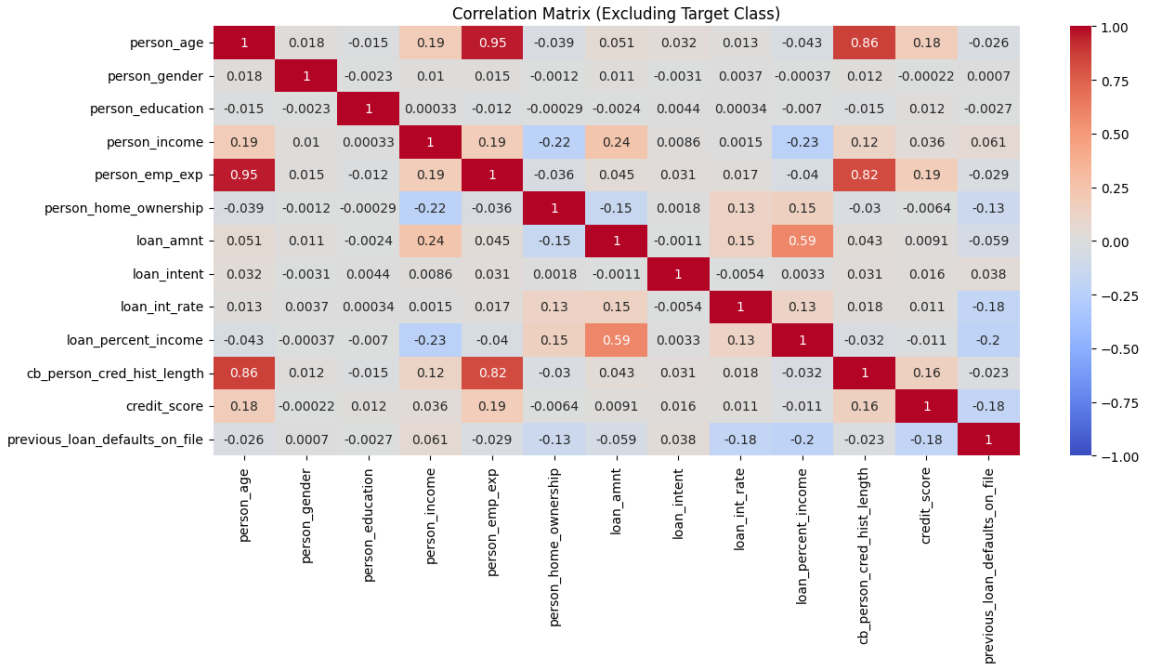

Figure 4. Correlation Matrix Between Variables

Some important findings from the correlation matrix in figure 4 are:

1. **Strong Relationship:**
   a. There is a very strong correlation between person_emp_exp and person_age (0.95), indicating that a person's work experience is closely related to his or her age.
   b. cb_person_cred_hist_length also shows a high correlation with person_age (0.86), which indicates that a person's age has a significant effect on credit history length.
   c. cb_person_cred_hist_length also shows a high correlation with person_emp_exp (0.82), indicating that the length of one's credit history has a significant effect on credit scores.
2. **Negative Relationship:**
   a. person_income has a significant negative correlation with loan_int_rate (-0.23), indicating that the higher a person's income, the lower the loan interest rate they receive.
   b. loan_percent_income also shows a negative correlation with person_income (-0.23), indicating that the proportion of income used for loans decreases as income increases.
3. **Weak Correlation:**
   Some features, such as person_gender and loan_amnt, show very weak correlations, indicating that the gender of the borrower does not have a significant influence on the loan amount applied for.
4. **Color Visualization:**
   The colors in the matrix provide a clear visualization of the strength and direction of the relationship. Red indicates a strong positive correlation, while blue indicates a negative correlation. Lighter colors indicate a weaker relationship.

This correlation matrix provides valuable insights into understanding the interactions between features in the dataset, which can help in the development of more effective loan approval prediction models. Figure 5 shows the correlation of variables with the target class.

Jurnal Minfo Polgan
Volume 13, Nomor 1, Februari 2024
DOI : https://doi.org/10.33395/jmp.v13i1.14345
e-ISSN : 2797-3298
p-ISSN : 2089-9424

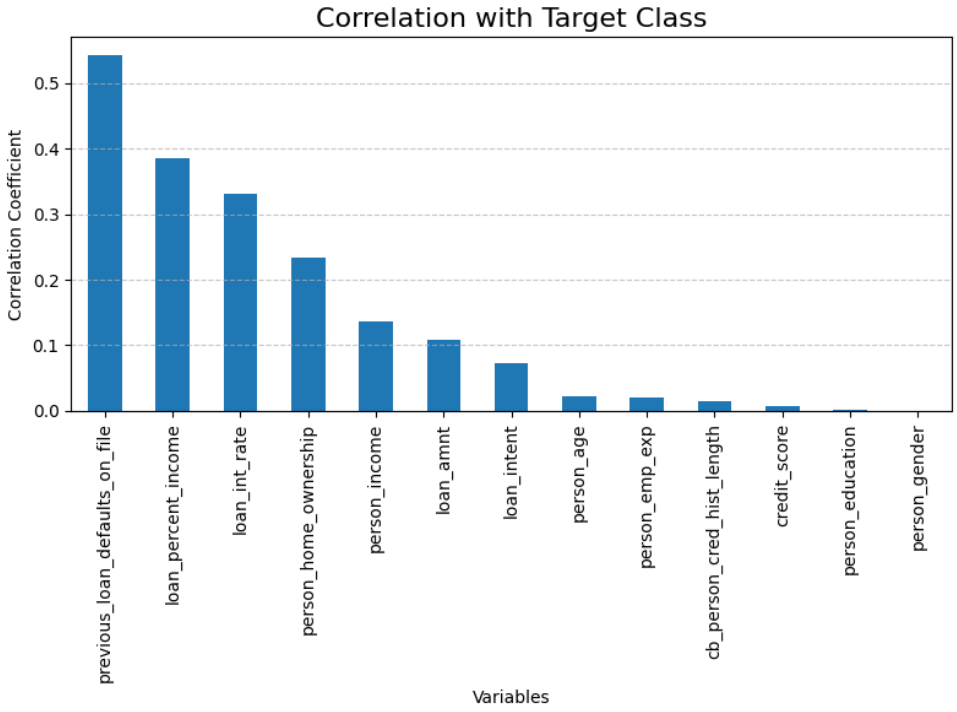

Figure 5. Correlation Of Variable With Target Class

Figure 5 shows a bar graph depicting the correlation coefficient between the various variables in the dataset and the target class, which is the loan approval status. previous_loan_defaults_on_file has the highest correlation coefficient, reaching around 0.5. This indicates that there is a significant positive relationship between the number of previous loan defaults and the likelihood of approval of the current loan. This means that borrowers who have a history of defaults tend to have a harder time getting approved.

loan_percent_income and loan_int_rate also show a fairly strong correlation, at around 0.4 each. This indicates that the proportion of income used for loans and the loan interest rate have a significant influence on approval decisions. On the other hand, variables such as credit_score, person_age, and person_gender show lower correlation coefficients, below 0.1. This suggests that these factors have minimal influence on loan approval decisions in the context of this dataset.

The results of modeling loan approval prediction with K-Nearest Neighbor and Random Forest algorithms show quite good results. Figure 6 shows the results of the analysis of K-Nearest Neighbor and Random Forest algorithms in the form of confusion matrix.

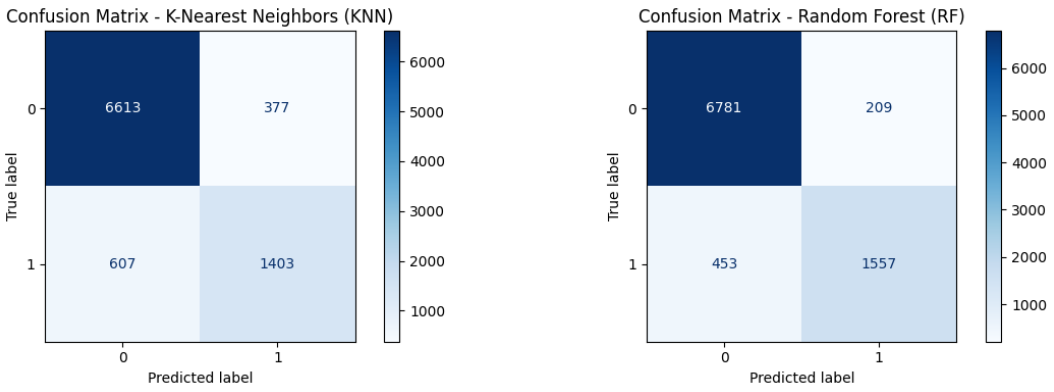

Figure 6. Confusion Matrix Analysis

Jurnal Minfo Polgan
Volume 13, Nomor 1, Februari 2024 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v13i1.14345 p-ISSN : 2089-9424

The performance results of the two classification models, K-Nearest Neighbors (KNN) and Random Forest (RF), based on evaluation metrics commonly used in classification analysis can be seen in Figure 7.

Model	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors (KNN)	1403	377	607	6613	0.8907	0.7882	0.6980	0.7404
Random Forest (RF)	1557	209	453	6781	0.9264	0.8817	0.7746	0.8247

Figure 7. KNN and RF Model Evaluation

From the confusion matrix analysis (Figure 6) and model evaluation (Figure 7), it can be concluded that the Random Forest model shows better performance compared to the K-Nearest Neighbors model in terms of credit approval prediction, where the Random Forest model obtained 93% accuracy while K-Nearest Neighbors obtained 89% accuracy. Random Forest has more true positives and true negatives, and fewer false positives and false negatives. In addition, this model also recorded higher accuracy, precision, recall, and F1-score values. This suggests that Random Forest is more effective in identifying both positive and negative classes, which could have positive implications for decision-making in the financial sector. This analysis provides valuable insights into understanding the strengths and weaknesses of each model, as well as assisting in selecting the most suitable model for a particular application.

CONCLUSION

From this study, it can be concluded that Random Forest algorithm is more effective than K-Nearest Neighbor in predicting loan approval. Confusion matrix analysis and evaluation metrics show that Random Forest has superior performance in identifying positive and negative classes, which has positive implications for decision-making in the financial sector. In addition, correlation analysis revealed significant relationships between some features, such as the number of previous loan defaults and the proportion of income used for loans, and approval decisions. These findings emphasize the importance of appropriate model selection and an in-depth understanding of the interactions between features in the development of more effective prediction models. This research makes an important contribution to financial institutions in improving the loan evaluation process and credit risk management.

REFERENCES

Aditya, B. & Nagaraju, V. (2022). Prophecy of loan approval by comparing Decision Tree with Logistic Regression, Random Forest, KNN for better Accuracy. *Journal of Pharmaceutical Negative Results*, *13*. https://doi.org/10.47750/pnr.2022.13.S03.87

Akumu, C., Smith, R., & Haile, S. (2021). Mapping and monitoring the canopy cover and greenness of southern yellow pines (Loblolly, shortleaf, and virginia pines) in central-eastern tennessee using multi-temporal landsat satellite data. *Forests*, *12*(4). https://doi.org/10.3390/f12040499

Anannya, M., Khatun, M. S., Hosen, M. B., Ahmed, S., Hossain, M. F., & Kaiser, M. S. (2023). Eligible Personal Loan Applicant Selection using Federated Machine Learning Algorithm. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 14, Issue 8). www.ijacsa.thesai.org

Devi, K. R., Pradhan, J., Bhutia, R., Dadul, P., Sarkar, A., Gohain, N. & Narain, K. (2021). Molecular diversity of Mycobacterium tuberculosis complex in Sikkim, India and prediction of dominant spoligotypes using artificial intelligence. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-86626-z

Dondekar, A. D. & Sonkamble, B. A. (2020). Harmonic Mean based Classification of Images using Weighted Nearest Neighbor for Tagging. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 11, Issue 11). www.ijacsa.thesai.org

Gavagsaz, E. (2022). Efficient Parallel Processing of k-Nearest Neighbor Queries by Using a Centroid-based and Hierarchical Clustering Algorithm. *Artificial Intelligence Advances*, *4*(1), 26–41. https://doi.org/10.30564/aia.v4i1.4668

Hota, L. & Dash, K. (2021). Comparative Analysis of Stock Price Prediction by ANN and RF Model. *Computational Intelligence and Machine Learning*.

Liu, Z., Zhu, G., Jiang, X., Zhao, Y., Zeng, H., Jing, J. & Ma, X. (2020). Survival Prediction in Gallbladder Cancer Using CT Based Machine Learning. *Frontiers in Oncology*, *10*. https://doi.org/10.3389/fonc.2020.604288

Paramita, A. S., Maryati, I., & Tjahjono, L. M. (2022). Implementation of the K-Nearest Neighbor Algorithm for the Classification of Student Thesis Subjects. *Journal of Applied Data Sciences*, *3*(3), 128–136.

Ren, Z., Tang, Y., & Zhang, W. (2021). Quality-related fault diagnosis based on k-nearest neighbor rule for non-linear industrial processes. *International Journal of Distributed Sensor Networks*, *17*(11). https://doi.org/10.1177/15501477211055931

Rosdiana, R., Novalia, V., Aidilof, H. A. K., Danil, M., & Fikri, M. I. (2021). APPLICATION AND ATTRIBUTE ANALYSIS IN THE MODEL OF CLASSIFYING HEART DISEASE. *MULTICA SCIENCE AND TECHNOLOGY (MST) JOURNAL*, *1*(2), 72–75. https://doi.org/10.47002/mst.v1i2.280

Sandeep, V. & Devi. T. (2022). *Journal of Pharmaceutical Negative Results*, *13*(SO4). https://doi.org/10.47750/pnr.2022.13.s04.210

Shin, K., Song, J. J., Bang, W., & Lee, G. W. (2021). Quantitative precipitation estimates using machine learning approaches with operational dual-polarization radar data. *Remote Sensing*, *13*(4), 1–23. https://doi.org/10.3390/rs13040694

Tong, X. & Duan, J. (2022). Research on the Prediction of Nonbreakeven Financial Products' Yield of Commercial Banks Based on Machine Learning. *Mobile Information Systems*, *2022*. https://doi.org/10.1155/2022/8731261

Wayahdi, M. R. & Ruziq, F. (2022). KNN and XGBoost Algorithms for Lung Cancer Prediction. *Journal of Science Technology (JoSTec)*, *4*(1), 179–186. https://doi.org/10.55299/jostec.v4i1.251

Wayahdi, M. R., Ruziq, F., & Ginting, S. H. N. (2024). AI APPROACH TO PREDICT STUDENT PERFORMANCE (CASE STUDY: BATTUTA UNIVERSITY). In *Journal of Science and Social Research* (Issue 4). http://jurnal.goretanpena.com/index.php/JSSRhttp://jurnal.goretanpena.com/index.php/

Wayahdi, M. R., Syahputra, D., & Ginting, S. H. N. (2020). *EVALUATION OF THE K-NEAREST NEIGHBOR MODEL WITH K-FOLD CROSS VALIDATION ON IMAGE CLASSIFICATION*. http://infor.seaninstitute.org/index.php/infokum/index