International Journal of Computing Sciences Research (ISSN print: 2546-0552; ISSN online: 2546-115X)
Vol. 9, pp. 3659-3673
doi: 10.25147/ijcsr.2017.001.1.243
https://stepacademic.net

Short Paper

An Enhancement of K-Nearest Neighbor Algorithm's Data Pre-Processing for Dataset Classifications in Predicting Multiple Medical Diseases

Madeleine S. Tisang
Computer Science Department, Pamantasan ng Lungsod ng Maynila, Philippines
mstisang2020@plm.edu.ph
(corresponding author)

Jaira Venessa C. Obmina
Computer Science Department, Pamantasan ng Lungsod ng Maynila, Philippines
jvcobmina2020@plm.edu.ph

Francis Arlando L. Atienza
Computer Science Department, Pamantasan ng Lungsod ng Maynila, Philippines
fcatienza@yahoo.com

Jonathan C. Morano
Computer Science Department, Pamantasan ng Lungsod ng Maynila, Philippines
jcmorano@plm.edu.ph

Leisyl M. Mahusay
Computer Science Department, Pamantasan ng Lungsod ng Maynila, Philippines
lmocampo@plm.edu.ph

Jamillah S. Guialil
Computer Science Department, Pamantasan ng Lungsod ng Maynila, Philippines
jsguialil@plm.edu.ph

Date received: May 7, 2024
Date received in revised form: July 14, 2024
Date accepted: May 11, 2025

Recommended citation:

Tisang, M., Obmina, J. V., Atienza, F. A., Morano, J. C., Mahusay, L. M., & Guialil, J. S. (2024). An enhancement of k-nearest neighbor algorithm's data pre-processing for dataset classifications in predicting multiple medical diseases. *International Journal of Computing Sciences Research*, 9, 3659-3673. https://doi.org/10.25147/ijcsr.2017.001.1.243

Abstract

Purpose – This research intends to improve the K-Nearest Neighbor Algorithm's data preparation, emphasizing improving disease prediction across datasets of varied sizes by addressing imbalanced datasets and optimizing the selection of an effective k value.

Method – The researchers utilized SMOTE and GridSearch to address challenges in the K-Nearest Neighbor Algorithm. SMOTE balanced the datasets to prevent inaccurate representations, while GridSearch improved the k value accuracy, reducing challenges with constant fixed k values. These techniques contributed to the study's overall effectiveness in accurately predicting diseases.

Results – When compared to eight datasets, the improved K-Nearest Neighbor algorithm consistently surpasses the previous approach in terms of accuracy, precision, RMSE, MSE, and t-test evaluation. The findings suggest that the enhanced KNN algorithm outperformed the existing KNN method in terms of prediction. This resulted in improved performance in predicting a wide range of medical problems across eight datasets.

Conclusion – In conclusion, the study effectively aimed to boost the performance of the K-Nearest Neighbor (KNN) algorithm in categorizing medical conditions through enhanced data pre-processing techniques. Ultimately, the study's findings show that the enhanced KNN algorithm is effective in accurately predicting medical disease across a variety of datasets.

Recommendations – The researchers recommend employing high-dimensional datasets to address the 'Dimensionality Curse' and to further ascertain the significance of this study. The results of this study will help improve medical diagnostics by predicting diseases more accurately.

Research Implications – The outcomes of this study offer improved medical diagnostics through more precise disease prediction, hence improving the effectiveness of the K-Nearest Neighbor (KNN) algorithm in identifying various health conditions.

Practical Implications – Through these enhancements, healthcare practitioners will be able to take action quickly, providing early treatment interventions and individualized treatment approaches, as disease prediction becomes more accurate.

Keywords – KNN, machine learning, SMOTE, GridSearch

INTRODUCTION

K-Nearest Neighbor learns by retaining the entire training set and assigns a class to each query based on the majority label of its k-nearest neighbors. The classifier's performance hinges on the choice of K and the applied distance metric. The selection of K impacts the estimate's sensitivity, with smaller K values leading to poor local estimates due to data sparseness, noise, or mislabeled points. Increasing K smooths the estimate but risks over-smoothing and degraded performance with the introduction of outliers from other classes (Imandoust & Bolandraftar, 2013).

As medical diseases become more prevalent, the need for precise prediction models grows. This study aims to improve the performance of the K-Nearest Neighbor algorithm in medical risk prediction. To prevent biased predictions towards the majority sample, which may affect the accuracy of predicting medical conditions, it addresses the problem of imbalanced datasets. The research also concentrates on resolving the fixed k input to avoid underfitting and overfitting, which can affect the algorithm's performance. These enhancements aim to contribute to healthcare by paving the way for more reliable prediction systems, benefiting patient care and outcomes.

LITERATURE REVIEW

K-Nearest Neighbor in Disease Prediction

The k-nearest neighbor (KNN) algorithm is frequently employed for predicting diseases and is primarily used in classification tasks. This supervised algorithm predicts the classification of unlabeled data by considering the features and labels of the training data. Essentially, the KNN algorithm categorizes datasets by evaluating a training model against a testing query. It does this by assessing the k nearest training data points (referred to as neighbors) that closely match the query being examined. Ultimately, the algorithm employs a majority voting method to decide on the final classification. Renowned for its simplicity, the KNN algorithm finds extensive use in classification tasks owing to its straightforward and flexible structure (Uddin et al., 2022).

Imbalance Dataset

In the study of Thabtah et al. (2020), classifiers in machine learning seek to enhance predicted accuracy while decreasing misclassification errors. Real-world datasets, particularly in medical diagnostics, often show class imbalance, with one class having much fewer instances than the others. This imbalance is particularly common in

models that detect rare but severe disorders such as autism spectrum disorder.

In the study of Rattan et al. (2021), the goal of their experiment is to compare the prediction of 4 machine learning algorithms and to identify if SMOTE will enhance the four machine learning algorithms' predictive accuracy in classifying academic scores accurately. To balance the dataset during preprocessing, oversampling techniques such as SMOTE is employed. A dataset is considered unbalanced if the classification categories are not roughly represented in the same proportion. The experiment showed that with SMOTE the classifiers' performances improved by 10%. The 4 classifiers with SMOTE produced higher correctly classified instances, incorrectly classified instances, percentage of correctly classified instances, and a lower percentage of misclassified instances than the 4 classifiers without SMOTE.

Determining the k value

Authors Hassanat et al. (2014) note that, typically, the K parameter for the KNN classifier is selected empirically. Various nearest neighbor counts are tested for each issue, and the classifier's definition is based on the parameter that performs best (accuracy). Nevertheless, picking the best K for a range of issues is practically impossible because a KNN classifier's performance changes considerably when K and the distance measure it uses are modified. However, it has been demonstrated in the literature that it is challenging to predict the value of K in advance when the examples are not evenly distributed. In the study of Gupta and Goel (2020), the ideal K value has been studied to get the best performance out of the KNN classifier. Finding the ideal number of neighbors (K) where it performs best is a little difficult. It varies from dataset to dataset.

Dimas and Mukti (2021) made a comparative study of Grid Search and Random Search methods for hyperparameter tuning of Chronic Kidney Failure. In the authors' study, the Grid Search and Random Search methods are utilized in the Extreme Gradient Boosting Algorithm to accurately predict chronic kidney failure. In this process, hyperparameters were tuned to obtain the optimal parameter values. After training and evaluation, the results demonstrated that the Grid Search method successfully identified the best hyperparameters, achieving an accuracy of 99.28%.

METHODOLOGY

The study utilized eight distinct medical datasets, each distinguished by its own set of instances and features. The diversity of these datasets facilitated a comprehensive assessment of both the current and proposed KNN algorithms' performance in accurately predicting medical diseases. There are 768 instances and 8 matching features in the PIMA Diabetes dataset (UCI Machine Learning, 2016). An open-source dataset for a large diabetes dataset with 388,754 instances and 21 features was made available by the CDC in 2021. The Stroke dataset (Fedesoriano, 2021)

has 10 features and 4,909 instances. There are 583 instances in the UCI Machine Learning (2016) Liver dataset, and each instance has 10 corresponding features. This study also used 569 instances with 30 features collected from the UCI Machine Learning, 2016 Breast Cancer dataset. In the SEER Breast Cancer data (Mandala, 2023) are 11 features with 4,024 instances. A dataset with 8,763 cases and 17 features for heart attack risk prediction was made available by Banerjee (2024). The 195 cases in the Oxford Parkinson's Disease Dataset (Ahmad, 2023) correspond to 22 features. These datasets were acquired from various open-source websites. Below is the tabulated overview of the datasets, accompanied by their respective numbers of instances and features (Table 1):

Table 1. Dataset Table

Dataset	Instances and Features
PIMA Diabetes	(768,8)
Large Diabetes Dataset	(388754,21)
Stroke	(4909,10)
Liver	(583,10)
Breast Cancer	(569,30)
Large Breast Cancer Dataset	(4024, 11)
Heart Attack	(8763,17)
Parkinson's Disease	(195, 22)

Enhanced K-Nearest Neighbor Conceptual Framework

To further enhance the K-Nearest Neighbor (KNN) framework, we concentrate on tuning parameters and resolving class imbalances. The aim is to increase the algorithm's precision and performance, in a diverse dataset, especially when handling large datasets (Figure 1).

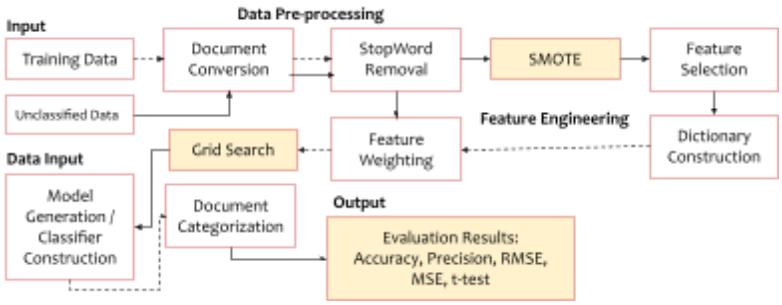

Figure 1. Enhanced K-Nearest Neighbor Conceptual Framework

Data pre-processing involves a series of steps to ensure the dataset's cleanliness, consistency, and compatibility with the algorithm's requirements, contributing to the robustness and reliability of the KNN classification model. It begins with the normalization or standardization of features to bring them to a consistent scale, the categorical features encoded to a numerical format to facilitate distance calculations, then the data will undergo the SMOTE technique to correct any imbalanced data in the dataset.

To address the imbalanced datasets, the researchers employed the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is often utilized in situations involving class imbalance as it provides synthetic samples for the minority class to balance the dataset. In this scenario, because the dataset is uneven, SMOTE is used to ensure that the model is trained on a more representative collection of data, enabling it to learn patterns from the minority class and improve the overall performance of the model. The researchers used the SMOTE technique to balance the dataset, which involved identifying the X and Y variables in the dataset, initiating the SMOTE technique for balancing, creating Z at a random point on the line segment using the formula $Z = X0 + w (X-X0)$ and storing the balanced dataset in a designated variable.

Furthermore, to determine the optimal value of k in the K-Nearest Neighbor algorithm, the researchers employed the GridSearch technique for hyperparameter tuning. GridSearch is a widely utilized algorithm in machine learning for identifying the best combination of hyperparameters for a given model. In this study, it was utilized to systematically evaluate different k values and enhance the algorithm's performance in medical risk prediction tasks through cross-validation. The researchers conducted hyperparameter tuning, initializing parameters, setting the number of cross-folds to 5, and defining a grid from 1 to 50. A model object was then created with these parameters and fitted to the training data. The researchers aim to enhance the algorithm's performance in medical risk prediction tasks by systematically evaluating different k values through cross-validation.

The data will be classified using the KNN technique after pre-processing and enhancements. This algorithm uses the updated features to find the data's k-nearest neighbors. Following the KNN algorithm's classification of the data, the model's accuracy is assessed. It is then evaluated in terms of performance metrics, including receiver operating curve graph, accuracy, precision, and t-test.

Evaluation and Testing Metrics

To evaluate the performance of the modified KNN algorithm, the following are employed to assess the prediction accuracy of the KNN model:

Accuracy: The sum of two accurate projections is used to calculate accuracy. (TP + TN)/total number of data sets (P + N) (Equation 1). The best possible accuracy is 1.0, while the worst possible accuracy is 0.00 (Vujović, 2021):

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} = \frac{TP + TN}{P + N} \qquad Equation\ 1$$

Precision: It computes the ratio of the number of precise positive projections (TP) to the overall number of positive predictions (TP + FP). 1.0 is the highest accuracy while
0.0 is the lowest (Equation 2). True positives and false positives are denoted by the

Equation 2

letters TP and FP, respectively (Vujović, 2021):

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

Root Mean Squared Error (RMSE) (Equation 3): The root mean square error is expressed in terms of what would occur if a basic predictor were applied. Below is the formula for the root mean square error (RMSE) E_i of an individual model (Vujović, 2021):

$$E_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(P_{(ij)} - T_j)^2} \qquad Equation\ 3$$

The target value for record j is T_j, and the value predicted by individual model *i* for record *j* (of n records) is $P_{(ij)}$. For a perfect prediction, $P_{(ij)} = T_j$ and $E_i = 0$. As a result, the ideal value of the index E_i is 0, and its range is 0 to infinity

Mean Absolute Error (MSE) (Equation 4): the mean of all the test set's instances' absolute values for each prediction mistake. The difference for each instance between the actual and expected values is known as the prediction error (Vujović, 2021):

$$E_i = \frac{1}{n}\sum_{j=1}^{n} | P_{(ij)} - \sum_{j=1}^{n}(P_{(ij)} - T_j | | \qquad Equation\ 4$$

where T_j is the goal value for record *j* and $P_{(ij)}$ is the value predicted by individual model *i* for record *j* (of n records). A perfect prediction has $E_i = 0$ and $P_{(ij)} = T_j$. As a result, the index E_i is a number between 0 and infinity, where 0 represents the idea.

The paired t-test is essentially a variation of the one-sample t-test (Equation 5),

focusing on the differences within pairs of data points. When there is no difference between the paired data points (null hypothesis), the test statistic T2 follows a t-distribution with degrees of freedom equal to the sample size minus one (df = n-1). Below is the t-test equation (Xu et al., 2017):

$$T_2 = \frac{X_d}{\sqrt{\frac{1}{n\,(n-1)}\sum_{j=1}^{n}(X_j - \overline{X_d})}} \qquad \textit{Equation 5}$$

RESULTS

Evaluation Results of the Existing and Enhanced KNN Algorithm across Eight Medical Datasets

The researchers employed the k value derived from the GridSearch and applied it to both the existing and enhanced KNN to compare their results (Table 2). Significant differences between various datasets are observed when comparing the accuracy results of the k value from the enhanced KNN with the k value obtained from improvements made to an existing algorithm. In the PIMA Diabetes dataset, the enhanced KNN achieves an accuracy of 76.67%, surpassing the existing KNN's accuracy of 71.35% by 5.32%. Similarly, in the larger Diabetes dataset with 388,754 instances and 21 features, the enhanced KNN achieves a significantly higher accuracy of 72.93% compared to the existing KNN's accuracy of 61.56%, representing an 11.37% difference.

For the Stroke dataset, the enhanced KNN yields an accuracy of 88.23%, outperforming the existing KNN's accuracy of 77.14% by 11.09%. In the evaluation of the Breast Cancer dataset, the enhanced KNN achieves an accuracy of 81.82% with a precision of 90.9%, while the existing KNN with the same k value of 48 achieves an accuracy of 67.13%. Likewise, in the large Breast Cancer dataset, the enhanced KNN attains an accuracy of 76%, showing a 3.83% difference compared to the existing KNN's accuracy of 72.18%. In contrast, the existing KNN performs better in the Liver disease dataset with an accuracy of 69.86% compared to the enhanced KNN's accuracy of 68.18%, demonstrating only a 1.68% difference. Overall, the enhanced KNN algorithm showcases superior performance across the remaining seven medical datasets, reflecting the effectiveness of the enhancements in enhancing prediction accuracy.

Notably, the enhanced KNN demonstrates improvements in Precision, RMSE, and MSE compared to the existing KNN in nearly all datasets. Furthermore, improvements are seen in the RMSE and MSE values, with the enhanced KNN consistently producing lower values than the existing KNN. Lower RMSE and MSE values suggest that the enhanced KNN algorithm makes predictions that are more accurate and reliable.

Table 2. Evaluation of Enhanced vs. Existing KNN Algorithm in PIMA Diabetes Dataset

KNN Algorithm	Dataset	K value	Accuracy (%)	Precision (%)	RMSE	MSE
Existing	PIMA Diabetes (768,8)	31	71.35	65	0.5352	0.2864
Enhanced	**PIMA Diabetes (768,8)**	**31**	**76.67**	**68.57**	**0.483**	**0.2334**
Existing	Large Diabetes Dataset (388754,21)	48	61.56	69.94	0.5306	0.2815
Enhanced	**Large Diabetes Dataset (388754,21)**	**48**	**72.93**	**72.39**	**0.5202**	**0.2706**
Existing	Stroke (4909,10)	1	77.14	77.34	0.478	0.229
Enhanced	**Stroke (4909,10)**	**1**	**88.23**	**86.5**	**0.343**	**0.118**
Existing	Liver (583,10)	26	68.78	73.08	0.549	0.3014
Enhanced	**Liver (583,10)**	**26**	**71.27**	**75**	**0.564**	**0.3181**
Existing	Breast Cancer (569,30)	48	60	68.07	0.5732	0.3286
Enhanced	**Breast Cancer (569,30)**	**48**	**92.1**	**90.9**	**0.4264**	**0.1818**
Existing	Large Breast Cancer Dataset (4024, 11)	1	66.81	75.31	0.5274	0.2782
Enhanced	**Large Breast Cancer Dataset (4024, 11)**	**1**	**78.75**	**78.14**	**0.4899**	**0.24**
Existing	Heart Attack (8763,17)	1	55.77	37.73	0.665	0.4422
Enhanced	**Heart Attack (8763,17)**	**1**	**58.89**	**58.88**	**0.6411**	**0.411**
Existing	Parkinson's Disease (195, 22)	1	84.75	88.89	0.39057	0.1525
Enhanced	**Parkinson's Disease (195, 22)**	**1**	**92.5**	**100**	**0.2236**	**0.05**

T-test Results of the Existing and Enhanced KNN Algorithm across Eight Medical Datasets

A paired t-test was conducted to assess the before and after enhancement of the same dataset to analyze the significance done by the enhancement. In contrast, the significance should be less than the standard significance level of 0.05 alpha for it to be deemed statistically significant. Below are the following results of the t-test (Table 3).

The evaluation of the t-test on the enhanced KNN algorithm above demonstrates the significance of the enhancements in each dataset. Seven out of eight datasets' p-values obtained from the t-test are less than 0.05. indicating statistical significance. In particular, for datasets related to PIMA Diabetes, Stroke, Liver, Breast Cancer, Parkinson's Disease, Diabetes (Large dataset), and Breast Cancer (Large Dataset), the p-values are all less than 0.05, suggesting that the enhancements have led to statistically significant changes in these datasets. However, for the Heart Attack dataset, the obtained p-value is greater than 0.05, indicating that the enhancements did not result in a statistically significant change in this dataset. Overall, this analysis demonstrates that enhancements to the KNN algorithm have led to significant improvements in the majority of the datasets evaluated.

Table 3. T-test of the Enhanced KNN Algorithm

Accuracy		
t-Test: Paired Two Samples for Means		
	P(T<=t) two-tail	Result
PIMA Diabetes	0.031208	< 0.05 therefore significant
Stroke	0.006948	< 0.05 therefore significant
Liver	0.022435	< 0.05 therefore significant
Breast Cancer	0.003883	< 0.05 therefore significant
Heart Attack	0.370749	>0.05 therefore NOT significant
Parkinson's Disease	0.016227	< 0.05 therefore significant
Diabetes (Large dataset)	0.000787	< 0.05 therefore significant
Breast Cancer (Large Dataset)	0.045612	< 0.05 therefore significant

DISCUSSION

This study aims to enhance the data pre-processing of the K-Nearest Neighbor

(KNN) algorithm to improve classifications across various types of datasets for different medical diseases. The study underwent two phases to evaluate the performances of the existing and enhanced KNN in predicting medical diseases accurately. The first phase of the study is the evaluation of the enhanced KNN. The enhanced KNN is then run to train and evaluate eight medical datasets. The second phase of the study is the evaluation of existing KNN when introduced to the derived k value from the enhanced KNN. After completing the two phases, the results from the enhanced KNN, and the existing KNN with the introduction of the GridSearch k value are compared. This comparison aims to differentiate significant differences in the overall performance of the algorithm, ultimately providing an accurate representation of the data.

From the provided data, it's evident that the enhanced KNN algorithm consistently outperforms the existing KNN algorithm across various datasets and evaluation metrics. The findings of the assessment of the enhanced KNN algorithm provide substantial evidence that the objectives specified in the research have been successfully met. Notably, as compared to the existing KNN method, the modifications have significantly improved prediction performance across a variety of medical datasets, as shown by higher Accuracy, Precision, and lower RMSE and MSE values. This is consistent with the study's objective of enhancing the algorithm's performance in medical risk prediction. Furthermore, the modifications have effectively addressed the issue of imbalanced datasets by consistently outperforming the existing KNN algorithm on datasets with variable class distributions. This indicates the modifications' effectiveness in reducing biases associated with imbalanced data sets while also enhancing prediction accuracy regardless of class imbalances. Additionally, the flexibility given by the modifications allows the algorithm to adapt to different datasets and avoid the challenges of fixed k input, resulting in improved predicted accuracy without being restricted by specific k numbers. Overall, the results show that the improvements made to the KNN algorithm help to design more reliable prediction systems in healthcare, which benefits patient care and outcomes.

Aside from the obtained Accuracy, Precision, RMSE, and MSE, the enhanced KNN also showed significance across the seven medical datasets in the t-test. By establishing statistical significance across different datasets, the improved KNN algorithm demonstrates its effectiveness in dealing with the challenges given by medical data. This significance indicates that the enhancements have resulted in real changes that can influence clinical decision-making and patient outcomes, rather than just numerical increases in prediction accuracy.

CONCLUSIONS AND RECOMMENDATIONS

In conclusion, the study successfully aimed to enhance the performance of the K-Nearest Neighbor (KNN) algorithm in classifying medical diseases by improving its data pre-processing. The experiment's utilization of the techniques - SMOTE and

Grid search - has been successful for the study. The evaluation and results of the enhanced KNN show that it consistently outperformed the existing KNN, achieving higher accuracy, precision, RMSE, and MSE in disease prediction while avoiding the challenges of overfitting and underfitting. The overall findings of the study underscore the effectiveness of the enhanced KNN algorithm in accurately predicting medical diseases across diverse datasets. Furthermore, the significance of the enhanced KNN algorithm across multiple medical datasets signify a promising advancement in medical data analysis. By demonstrating consistent improvements in predictive performance and statistical significance, the enhanced KNN algorithm holds great potential for enhancing clinical decision support systems, improving patient outcomes, and advancing medical research.

The proposed enhancements in this study for KNN exhibited superior performance compared to the existing KNN. However, there are still unexplored improvements that can be made. Given the diversity of datasets in the medical field, further exploration is recommended using different datasets. Additionally, for future studies, testing these enhancements in high-dimensional datasets will provide insights into their performance in dealing with the "Dimensionality Curse." To further ascertain the significance of this study, the researchers also proposed the exploration of its applicability beyond medical datasets.

IMPLICATIONS

The findings of this study hold promising improvements in medical diagnostics by predicting diseases more accurately and contributing to a more robust framework for diagnosing diverse health issues by improving the performance of the K-Nearest Neighbor (KNN) algorithm in the classification of medical diseases. Healthcare practitioners will be able to take action earlier, resulting in early treatment interventions and individualized treatment approaches, as disease prediction becomes more accurate.

ACKNOWLEDGEMENT

We would like to extend our heartfelt gratitude to everyone who helped make this study possible. We would like to express our deepest gratitude to our thesis adviser whose expertise and constant assistance in ensuring the successful completion of this study. We would also especially want to express our gratitude to our consultant for her encouragement, thoughtful suggestions, and assistance during the whole process.

FUNDING

The study did not receive funding from any institution.

DECLARATIONS

Conflict of Interest

The researchers declare no conflict of interest in this study.

Informed Consent

Informed consent is not necessary for this study because it involves the analysis of openly accessible data that has been published and does not contain identifiable personal information.

Ethics Approval

Given that the study involves the analysis of publicly available data, ethics approval is not applicable. Any data used will be handled following relevant regulations and best practices in machine learning research.

REFERENCES

Ahmad. (2023, October 16). *Oxford Parkinson's Disease Detection Dataset*. Kaggle. https://www.kaggle.com/datasets/pypiahmad/oxford-parkinsons-disease-detection-dataset

Banerjee, S. (2024, May 11). *Heart attack risk prediction Dataset*. Kaggle. https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset?fbclid=IwAR2dSxWoGWlZ2u_4iXNBPlZ55I2GdKo8JTGE9VdtDovSeggn_TpibbO1MNs

CDC. (2021, November 8). *Diabetes Health Indicators Dataset*. Kaggle. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_health_indicators_BRFSS2015.csv

Dimas, A., & Mukti, S. S. (2021). Performance comparison of grid search and random search methods for hyperparameter tuning in extreme gradient boosting algorithm to predict chronic kidney failure. *International Journal of Intelligent Engineering and Systems*, 14(6), 198–207. https://doi.org/10.22266/ijies2021.1231.19

Fedesoriano. (2021, January 26). Stroke Prediction Dataset. Kaggle. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Gupta, S. C., & Goel, N. (2020). Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. https://doi.org/10.1109/icssit48917.2020.9214129

Hassanat, A. B. A., Abbadi, M. A., Altarawneh, G. A., & Alhasanat, A. A. (2014). Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach. *arXiv (Cornell University)*. https://arxiv.org/pdf/1409.0919

Imandoust, S. B., & Bolandraftar, M. (2013). Application of K-nearest neighbor (KNN) approach for predicting economic events theoretical background. *ResearchGate*. https://www.researchgate.net/publication/304826093_Application_of_K-nearest_neighbor_KNN_approach_for_predicting_economic_events_theoretical_background on machine learning classifiers. *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*. https://doi.org/10.1109/esci50559.2021.9396962

Mandala, S. (2023, May 19). *SEER Breast Cancer data*. Kaggle. https://www.kaggle.com/datasets/sujithmandala/seer-breast-cancer-data?select=SEER+Breast+Cancer+Dataset+.csv

Rattan, V., Mittal, R., Singh, J., & Malik, V. (2021). Analyzing the application of SMOTE on machine learning classifiers. 2021 International Conference on Emerging Smart Computing and Informatics (ESCI). https://doi.org/10.1109/esci50559.2021.9396962

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. H. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences (Print)*, *513*, 429–441. https://doi.org/10.1016/j.ins.2019.11.004

UCI Machine Learning. (2016, September 25). Breast Cancer Wisconsin (Diagnostic) data set. Kaggle. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

UCI Machine Learning. (2016b, October 6). PIMA Indians Diabetes Database. Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

UCI Machine Learning. (2017, September 20). *Indian liver patient records*. Kaggle. https://www.kaggle.com/datasets/uciml/indian-liver-patient-records

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbor (KNN) algorithm and its different variants for disease prediction. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-10358-x

Vujović, Ž. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications,* *12*(6). https://doi.org/10.14569/ijacsa.2021.0120670

Xu, M., Fralick, D., Zheng, J. Z., Wang, B., Tu, X. M., & Feng, C. (2017). The Differences and Similarities Between Two-Sample *T*-Test and Paired *T*-Test. *Shanghai archives of psychiatry*, *29*(3), 184–188. https://doi.org/10.11919/j.issn.1002-0829.217070

Author's Biography

Madeleine Tisang is an aspiring programmer who is proactive in learning and utilizing programming languages tailored to specific projects. She also practices designing with the aim of professional growth. She demonstrated the ability to document progress during system analysis. Through their collective efforts and experiences, both authors are committed to continuously seeking knowledge in expanding their expertise along with emerging technologies.

Jaira Venessa Obmina is a dedicated leader and creative designer committed to personal and professional growth. Together with her partner, she poured their hearts into completing this research study, demonstrating resilience and determination. Driven by a passion for learning, Jaira continually seeks new experiences to enhance her skills and contributions to her field. Her innovative approach reflects her commitment to excellence and adaptability.

Francis Arlando L. Atienza is a professor at Pamantasan ng Lungsod ng Maynila. He is also the adviser who guided his advisee and provided helpful insights and studies that contributed a lot to the study.

Jonathan C. Morano is a Lecturer I at Pamantasan ng Lungsod ng Maynila. He has more than 20 years of experience working and teaching in the field of Information Technology. In this position, Jonathan oversees the thesis process, guaranteeing that students obtain the necessary guidance and resources to complete their research.

Leisyl M. Mahusay serves as the Thesis Coordinator at the College of Information Systems and Technology Management, Pamantasan ng Lungsod ng Maynila, Philippines. She supervises the thesis process, guaranteeing that all elements, from proposal to defense, are executed seamlessly and effectively.

Jamillah S. Guialil is a Bachelor of Science in Computer Studies, majoring in Computer Science, graduate in 2018. She is currently pursuing a Master of Information Technology at Pamantasan ng Lungsod ng Maynila. She also teaches as a part-time faculty member at the College of Information Systems and Technology Management at the same university. As an expert in information systems, is instrumental in assessing and offering feedback on student theses to ensure compliance with academic and industry standards.