REVISITING DISTANCE METRICS IN K-NEAREST NEIGHBORS ALGORITHMS: IMPLICATIONS FOR SOVEREIGN COUNTRY CREDIT RATING ASSESSMENTS

Ali İhsan ÇETİN[1,3,*], Ali Hakan BÜYÜKLÜ[2]

[*1]Department of Statistics, Faculty of Science, Yildiz Technical University, İstanbul, Turkey

[2]Department of Statistics, Faculty of Science, Yildiz Technical University, İstanbul, Turkey

[3]Department of Finance and Banking, Ankara Yildirim Beyazit University, Ankara, Turkey

[*] Corresponding author; E-mail: aliihsancetin22@gmail.com

The k-Nearest Neighbors algorithm, a fundamental machine learning technique, typically employs the Euclidean distance metric for proximity-based data classification. This research focuses on the Feature Importance Infused k-Nearest Neighbors model, an advanced form of k-Nearest Neighbors. Diverging from traditional algorithm uniform weighted Euclidean distance, Feature Importance Infused k-Nearest Neighbors introduces a specialized distance weighting system. This system emphasizes critical features while reducing the impact of lesser ones, thereby enhancing classification accuracy. Empirical studies indicate a 1,7% average accuracy improvement with proposed model over conventional model, attributed to its effective handling of feature importance in distance calculations. Notably, a significant positive correlation was observed between the disparity in feature importance levels and the model's accuracy, highlighting proposed model's proficiency in handling variables with limited explanatory power. These findings suggest proposed model's potential and open avenues for future research, particularly in refining its feature importance weighting mechanism, broadening dataset applicability, and examining its compatibility with different distance metrics.

Key words*: kNN; Feature Importance; distance weighting; Credit Scoring; Accuracy*

1. Introduction

T In the rapidly advancing domain of machine learning, the k-Nearest Neighbors (kNN) algorithm is lauded for its simplicity and effectiveness. It finds extensive application in classification tasks across various sub-disciplines within finance. This algorithm is particularly prominent in segmentation and categorization processes, most notably within the realms of energy systems investments and renewable energy financing, where it stands out as a method of first consideration [1]. It is particularly renowned for its capabilities in pattern recognition, text classification, and various data mining applications, primarily utilizing the principle of proximity-based class labeling. The Euclidean distance metric is predominantly used to quantify this proximity. Despite its simplicity and apparent efficiency, the kNN algorithm encounters specific limitations that may impede its performance in certain scenarios. Critical factors such as the choice of 'k', the number of nearest

neighbors to consider, and the selection of an appropriate distance metric, are pivotal in influencing the predictive accuracy of the algorithm. Ongoing research endeavors [2,3] focus on the problem of optimal 'k' selection. Moreover, the imbalance in sample distribution can also adversely affect classification accuracy. Recent studies [4] propose novel methodologies for calculating 'k' and addressing issues like training sample imbalance, assessing their effectiveness using 15 UCI benchmark datasets.

The primary objective of this study is to augment the kNN algorithm's performance by optimizing the Euclidean distance metric, a fundamental component of the algorithm. We aim to enhance prediction accuracy by refining this essential element, introducing an adaptive method that customizes the distance metric to meet specific dataset requirements. The accurate determination of the distance metric is hypothesized to substantially elevate the predictive accuracy and, consequently, the overall efficiency of the kNN algorithm. In this research, we employ advanced optimization techniques to modify the traditional Euclidean distance metric in kNN, thereby rendering it more adaptable to the dataset in question. We conduct an exhaustive analysis to evaluate the impact of this optimization on the algorithm's predictive accuracy and computational efficiency. Building upon previous research acknowledging the necessity for advancements and adaptations in the algorithm, our study proposes an innovative and potentially impactful solution to some of the algorithm's fundamental challenges. We anticipate that our findings will contribute significantly to the fields of machine learning and data mining, especially in contexts where high predictive accuracy is critical. The structure of this paper is as follows: we begin with an extensive review of the kNN algorithm and its prevalent challenges. Subsequently, we introduce our novel method for optimizing the Euclidean distance metric within the kNN framework. We then present a thorough analysis of our experimental results, underscoring the effectiveness of our proposed approach. The paper concludes with a discussion of our findings, implications, and directions for future research.

2. Related Works

A multitude of studies have been conducted on the modification of distance metrics in the k-Nearest Neighbors (kNN) algorithm. The study presented in [5] critically analyzes the kNN algorithm, offering a novel weighting function that integrates neighbor features alongside distances. This approach is rigorously tested on various real-world and synthetic datasets, allowing for a comprehensive assessment of its strengths and limitations, thereby charting potential areas for future improvements. In [6], paper introduces an ensemble method called EkNN-RF, which enhances the kNN algorithm. This method optimizes the nearest neighbor number and distance function based on a validation set. The feature set and training set for each base classifier are obtained through bootstrap sampling, giving higher weight to more important features. Each base classifier contributes to the final classification through voting. Experimental results suggest that EkNN-RF outperforms Adaboost, Naive Bayes, Random Forest, and other kNN variants in terms of classification accuracy on certain datasets. In [7], the DCTkNN, a variant of the k-Nearest Neighbor algorithm, is proposed. This algorithm refines the process of selecting k-nearest neighbors and the classification approach by factoring in class contribution and feature weighting. It calculates weighted distances by contrasting accuracies from a complete dataset against those from data missing each dimension feature. The algorithm assigns final sample labels based on a blend of the count of k-nearest neighbors and their average distance, showing enhanced classification accuracy on UCI datasets [8]. In reference [9], the

Bayesian-kNN (BCkNN) is introduced as an enhancement of the Citation-kNN (CkNN) for multi-instance classification. This improved algorithm utilizes a Bayesian framework for selecting k references and employs a distance-weighted majority vote for q citers, as opposed to the conventional majority voting system. Empirical evidence suggests that BCkNN surpasses both its predecessors, Bayesian-KNN (BkNN) and CkNN, in performance while maintaining computational efficiency comparable to CkNN. The study [10] introduces the Parameter Independent Fuzzy class-specific Feature Weighted k-Nearest Neighbor (PIFW-kNN) classifier, enhancing traditional FkNN by optimizing 'k' and feature weights using the Success-History based Adaptive Differential Evolution (SHADE) algorithm, yielding improved accuracy. Further, [11] discusses an improved Weighted k-Nearest Neighbor (WkNN) algorithm tailored for indoor localization. This method leverages both spatial and physical distances of Received Signal Strength (RSS) for reference point (RP) selection and employs a fusion weighted algorithm for accurate position calculation. Experimental validations demonstrate that this novel approach markedly surpasses conventional methodologies like kNN, E-WkNN, and P-WkNN, delivering superior positioning accuracy and outperforming recent advancements in WkNN algorithms. Lastly, [12] introduces an innovative distance metric tailored for specific datasets, employing a metaheuristic optimization technique known as differential evolution (DE). The effectiveness of this metric, when integrated into the k-Nearest Neighbor algorithm and evaluated against 30 benchmark datasets, was substantiated, demonstrating its successful application and adaptability.

3. Methodology

3.1. A Sovereign Credit Ratings and Its Applications on Machine Learning

Sovereign credit ratings, crucial assessments of a country's creditworthiness conducted by prominent international agencies such as Standard & Poor's, Moody's, and Fitch Ratings, offer a comprehensive appraisal of the risk associated with lending to nations [13,14]. These evaluations incorporate a variety of factors, including but not limited to economic indicators (e.g., GDP growth, inflation rates, fiscal balance), political stability, debt burdens, and historical instances of default. The implications of these ratings are significant, influencing borrowing costs for sovereign entities and playing a pivotal role in the global financial landscape. Recent advancements in the field of sovereign credit rating prediction have increasingly favored the use of Machine Learning (ML) methodologies, demonstrating potential for superior accuracy over conventional models [15,16]. This research paper delves into the efficacy of an innovative application of the k-Nearest Neighbors (kNN) algorithm in predicting sovereign credit ratings. Employing an extensive dataset that spans a range of economic, political, and social indicators, the study compares the performance of this novel kNN approach with traditional econometric models and existing methods used by leading credit rating agencies. Notably, this new kNN variant exhibits a marked proficiency in leveraging spatial relationships within the multidimensional feature space, thereby enhancing predictive accuracy. The findings of this study underscore the substantial promise of this new weighted kNN methodology in refining the precision of sovereign credit ratings.

3.2. Dataset

The efficacy and validity of the proposed Feature Importance Infused k-Nearest Neighbors (FIIkNN) approach were assessed using an original dataset of sovereign country credit ratings. This distinctive dataset, comprising 20 features and a binary target variable, was sourced from the Fitch Ratings agency's website, supplemented with open-source data from the World Bank and the United Nations [17,18]. A comprehensive analysis of the data modeling results, employing this unique dataset, is elaborated in Section 4.1 of the study.

3.3. k-Nearest Neighbors (kNN) algorithm

The k-Nearest Neighbors (kNN) algorithm, essential in supervised machine learning, classifies an instance by referencing the labels of the 'k' nearest instances in its feature space. The choice of distance metric, such as Euclidean or Manhattan, is pivotal, as it defines 'proximity' and influences algorithm performance. The number of neighbors, 'k', is also crucial; an ideal 'k' balances sensitivity to noise and relevance of the considered neighbors. While kNN benefits from simplicity and minimal data distribution assumptions, optimal performance requires careful selection of both distance metric and 'k', often using techniques like cross-validation.

The algorithm follows two primary steps:

The algorithm computes the Euclidean distance between two data points or tuples $X_1(x_{11}, x_{12}, ..., x_{1n})$ and $X_2(x_{21}, x_{22}, ..., x_{2n})$ using the following formula:

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \quad (1)$$

To handle the potential issues associated with wide-ranging feature values, features are often normalized:

$$\upsilon' = \frac{\upsilon - \min(A)}{\max(A) - \min(A)} \quad (2)$$

The algorithm assigns the class of a query point based on the majority class among its kNN:

$$y = \arg\max_{C_j} \sum_{l \in X_k} I(y_l = j) \quad (3)$$

In this equation, X_k represents the k-nearest neighbors, y_l is the class label of each neighbor, and $I()$ is an indicator function that returns 1 if its argument is true and 0 otherwise.

The efficacy of the kNN algorithm is intrinsically linked to the choice of distance metric. Traditional metrics such as Euclidean, Manhattan, and weighted Euclidean are commonly utilized, yet their applicability may vary across different data types, necessitating the exploration of adaptive distance measures [19]. To enhance the suitability of these measures for diverse datasets, optimization methods like Differential Evolution (DE) are frequently applied, refining their parameters for increased adaptability [20]. However, this parameter expansion can result in heightened computational demands. In pursuit of improved kNN accuracy, various methodologies have been proposed. These include density-based sample reduction strategies and inverse proportion weight kNN techniques, designed to mitigate class imbalance issues. Also, hybrid approaches have been explored, combining kNN with other analytical methods scuh as SVMs [21].

3.4. Euclidean Distance

The Euclidean distance (L2 norm) is mathematically defined as the square root of the aggregate of squared discrepancies between corresponding components of two vectors. This measure fundamentally represents the linear distance traversed along a straight path connecting two points in a Euclidean space [22].

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4}$$

3.5. Random Forest Feature Importance

The application of decision trees in the realm of machine learning is prevalent [23]. The concept of feature importance in Random Forest models is a critical technique employed to ascertain the extent of influence or contribution of various features towards the target variable. Random Forests determine this feature importance by computing the Gini impurity, which essentially quantifies the error reduction or 'pollution' introduced by each feature during the data splitting process. High-level description of the process, which can be represented in:

For each tree in the forest:
a. For each feature:
i. Calculate the decrease in impurity caused by splits on the feature.
ii. Accumulate the value for the feature.

The significance of a feature within a Random Forest model is directly proportional to its ability to decrease impurity. The average reduction in impurity, aggregated for each feature across all trees, is calculated as follows:

The importance $I(f)$ of feature f is:

$$I(f) = \frac{1}{|T|} \sum_{t \in T} \Delta I(t, f) \tag{5}$$

T : Set of all trees, F : Set of all features, $\Delta I(t, f)$: Decrease in impurity caused by feature f in tree t.

3.6. Euclidean Distance Form with Infused Feature Importance Weighting

In classical k-Nearest Neighbors (kNN) algorithms, the calculation of distances between samples typically employs the Euclidean distance metric, which utilizes a uniform weighting approach. This method does not consider the varying impact of individual features on the target variable, potentially leading to an inaccurate representation of each feature's unique importance. To address this, our study introduces an enhancement to the kNN algorithm by incorporating feature importance, as determined by the Random Forest algorithm, into the distance calculation. This modification aims to provide a more accurate and refined measure of sample distances. The Random Forest algorithm determines the importance of a feature by assessing the increase in prediction error when the feature's values are altered. The importance scores derived from this process are then normalized, ensuring that their sum equals 1. These normalized scores are denoted as $FII_{(i)}$ for each feature with index *i*. In our modified approach, these importance scores are utilized in the Euclidean distance formula. Traditionally, in the Euclidean formula, each term is equally weighted, represented

as $(1/i)$, where i is the number of features. In our weighted Euclidean kNN algorithm, we replace the equal weight of $(1/i)$ with the feature importance scores, maintaining the sum of weights at 1. The formula for calculating distance in this modified algorithm, within a d-dimensional feature space between two instances x and y, is then adapted to incorporate these weights, thus enhancing the distance metric's relevance to the specific features of the dataset. The formula is shown as follows:

$$D(x,y)=\sqrt{\sum_{i=1}^{d} i\times\frac{1}{i}\times\left(x_{[i]}-y_{[i]}\right)^2} \tag{6}$$

where $FII_{(i)}$ instead of $\frac{1}{i}$

$$D(x,y)=\sqrt{\sum_{i=1}^{d} i\times FII_{(i)}\times\left(x_{[i]}-y_{[i]}\right)^2} \tag{7}$$

In this methodology, i signifies an index ranging from 1 to d, where $x_{[i]}$ and $y_{[i]}$ represent the values of the i^{th} feature in the x and y samples, respectively, and $FII_{(i)}$ denotes the significance of the ith feature as determined by the Random Forest algorithm. This framework assigns greater influence in the Euclidean distance calculation to features deemed more important, thus proportionally contributing as per the feature count. Consequently, this tailored approach to measuring sample proximity accounts for the individual importance of each feature, thereby facilitating more precise neighbor identification and potentially enhancing the algorithm's classification accuracy. Further, this novel method capitalizes on the robustness and bias-variance tradeoff strengths inherent in the Random Forest algorithm, thereby augmenting its resilience to common kNN challenges such as overfitting and sensitivity to irrelevant features.

Future empirical studies could involve the experimental validation of this enhanced kNN algorithm, utilizing the weighted Euclidean distance across diverse datasets and varying parameter settings. This would enable a thorough assessment of the method's effectiveness in boosting prediction performance and classification accuracy. The principal pseudo-code of the Feature Importance Infused k-Nearest Neighbors (FIIkNN) algorithm is outlined as follows:

1. **Input:** Dataset *D*, Number of neighbors *k*, Random Forest parameters *RF_params*
2. **Data Preprocessing:** Conduct necessary preprocessing on Dataset *D* if required
3. **Random Forest Training:** Implement Random Forest on Dataset *D* utilizing *RF_params* to determine feature importance
4. **Feature Importance Calculation:** For each feature *i* in Dataset *D:*
   4.1 FIIkNN(i) = Calculate the importance of feature *i* using Random Forest
5. Definition of Weighted Euclidean Distance Function:
   5.1 Initialize variable *Sum* as *0*
   5.2 Iterate *i* from 1 to *d*(dimension):
      5.2.1 Increment *Sum* by (i) * FIIkNN(i) * (x[i] - y[i])^2
   5.3 Return the square root of *Sum* $\left(\sqrt{Sum}\right)$
6. **Classification Procedure:** For each test instance *x* in Dataset *D:*
   6.1 Compute the Weighted Euclidean Distance between *x* and all distances in *D*
   6.2 Order the computed distances in ascending sequence
   6.3 Identify the top *k* instances with the smallest distances
   6.4 Assign to *x* the majority class label from these *k* instances
7. **Output:** Predicted class labels for all test instances in Dataset *D*

4. Results

4.1. Efficacy of the Proposed FIIkN

In the experimental design of our study, we meticulously established the parameterization for the Feature Importance Infused k-Nearest Neighbors (FIIkNN) methodology. Initially, the selection of 'k' values, comprising 3, 5, and 7, was aligned with those commonly used in financial research studies. And, bootstrap samples of 300, 600, and 1000 were generated for each chosen 'k' value, aiming to optimize the balance between accuracy and computational efficiency. The mean accuracy for these varying configurations was then calculated. To ensure robustness and impartiality in our analysis, the size of the feature sets was deliberately limited to 2, 4, 6, 10, and 20. Comprehensive evaluations were subsequently conducted for each of these feature sets, considering the different combinations of bootstrap sample sizes and 'k' values.

4.2. Assessment of the Performance of the FIIkNN

This study endeavors to examine the efficacy metrics of an innovatively parameterized algorithm, developed by uniquely adapting the traditional kNN model through the incorporation of a Euclidean distance weighting mechanism. A wide array of configurations was employed in our evaluation, entailing various permutations stemming from distinct bootstrap sample sizes, dimensions of feature sets, and selected 'k' values. The results, highlight the intricate variations and interplays among these factors.

Table 1. Average Accuracy Attained Using FIIkNN with Two Features in Sovereign Country Credit Rating Dataset

		KNN			FIIKNN		
		k=3	**k=5**	**k=7**	**k=3**	**k=5**	**k=7**
Feature=2	**Boostrap=300**	0,87987	0,88577	0,87234	**0,90501**	**0,88712**	**0,89562**
	Boostrap=600	0,88015	0,87609	**0,88768**	**0,90516**	**0,89935**	0,88077
	Boostrap=1000	0,87980	0,87595	**0,89246**	**0,90501**	**0,89921**	0,88151

Table 1 offers a comprehensive comparative analysis of the classification results, which are based on datasets sampled in increments of 300, 600, and 1000, using an 80-20% random selection protocol. This comparison sheds light on the efficacy of the standard kNN algorithm's Euclidean distance metric against the proposed novel metric in this study. Distance calculations were conducted using 'k' values of 3, 5, and 7, aligning with commonly referenced standards in existing literature. Each experimental iteration was associated with a specific number of bootstrap samples, directly correlating with the mean accuracy obtained. The analysis primarily focused on the two most critical independent variables influencing credit decisions, as determined by the Random Forest Feature Importance technique. Following these assessments, an average of these outcomes was computed to derive the final results. As detailed in Table 1, it is evident that the classification accuracy achieved by the Feature Importance Infused k-Nearest Neighbors (FIIkNN) approach significantly surpasses that of the classic kNN methodology.

Table 2. Average Accuracy Attained Using FIIkNN with Four Features in Sovereign Country Credit Rating Dataset

		KNN			FIIKNN		
		k=3	**k=5**	**k=7**	**k=3**	**k=5**	**k=7**
Feature=4	**Boostrap=300**	**0,93981**	0,90514	0,90327	0,92712	**0,91958**	**0,91413**
	Boostrap=600	0,91041	**0,91885**	0,90011	**0,92492**	0,90919	**0,91374**
	Boostrap=1000	**0,92973**	0,90487	0,90501	0,92101	**0,91907**	**0,91358**

Table 2 demonstrates that, in scenarios involving a model with four features, the classification accuracy of the FIIkNN approach high mostly exceeds that of the traditional kNN method.

Table 3. Average Accuracy Attained Using FIIkNN with Six Features in Sovereign Country Credit Rating Dataset

		KNN			FIIKNN		
		k=3	**k=5**	**k=7**	**k=3**	**k=5**	**k=7**
Feature=6	**Boostrap=300**	0,97256	0,96844	0,96376	**0,98642**	**0,98365**	**0,97754**
	Boostrap=600	0,97275	**0,97869**	0,96403	**0,98763**	0,97789	**0,97875**
	Boostrap=1000	0,97454	0,96914	0,96311	**0,98744**	**0,98164**	**0,97852**

Upon analyzing Table 3 in conjunction with Tables 1 and 2, a discernible trend emerges, indicating a positive correlation between the number of features utilized and the Feature Importance Infused k-Nearest Neighbors (FIIkNN) model relative to the conventional k-Nearest Neighbors (kNN) algorithm.

Table 4. Average Accuracy Attained Using FIIkNN with Ten Features in Sovereign Country Credit Rating Dataset

		KNN			FIIKNN		
		k=3	**k=5**	**k=7**	**k=3**	**k=5**	**k=7**
Feature=10	**Boostrap=300**	0,97852	0,96511	0,96012	**0,98943**	**0,98422**	**0,97910**
	Boostrap=600	0,96980	0,96513	0,95996	**0,98960**	**0,98409**	**0,97883**
	Boostrap=1000	0,96973	0,96503	0,96002	**0,98942**	**0,98400**	**0,97881**

Upon examining Table 4, which presents analyses for feature numbers of 10, it is evident that the performance of the proposed approach consistently exhibits 100% superiority over the kNN's performance.

Table 5. Average Accuracy Attained Using FIIkNN with Twenty Features in Sovereign Country Credit Rating Dataset

		KNN			FIIKNN		
		k=3	**k=5**	**k=7**	**k=3**	**k=5**	**k=7**
Feature=20	**Boostrap=300**	0,98102	0,97836	0,97617	**0,99698**	**0,99343**	**0,98941**
	Boostrap=600	0,98111	0,97844	0,97422	**0,99700**	**0,99357**	**0,98962**
	Boostrap=1000	0,98110	0,97838	0,97611	**0,99690**	**0,99346**	**0,98944**

In our comparative analysis, the Feature Importance Infused k-Nearest Neighbors (FIIkNN) model, which integrates an expanded 20-feature set, demonstrated a 100% performance enhancement over traditional methods, as shown in Table 5. This aligns with findings in Table 4, indicating superior efficacy of FIIkNN in handling complex, high-dimensional data. The FIIkNN's robustness and adaptability are further affirmed by its significant accuracy improvement, ranging from 1% to 1.7%, over the conventional kNN algorithm employing weighted Euclidean distance. This improvement, verified through repeated cross-validation across various datasets, is notable in binary target variable contexts with baseline accuracies above 90%. The FIIkNN's success is attributed to its effective integration of feature importance in distance computation, optimizing the impact of key features and leading to more precise classifications. The study's results emphasize the scalability and effectiveness of the FIIkNN model in advanced machine learning tasks.

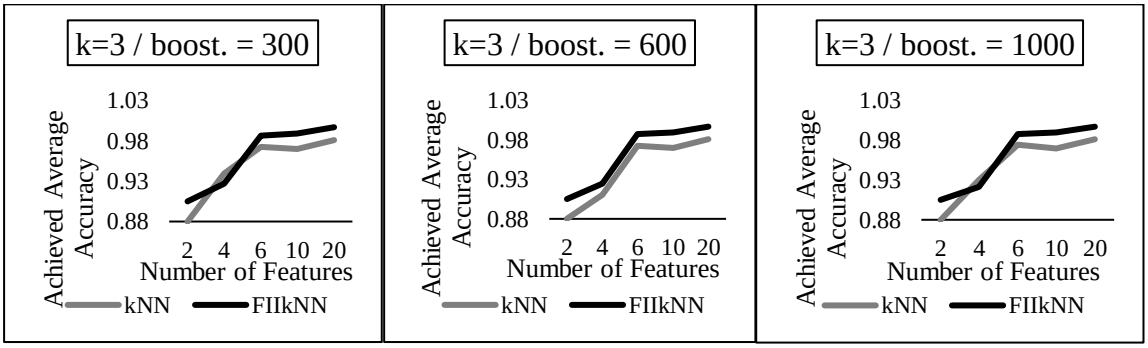

Figure 1. Achieved average accuracies based on increasing number of features for k=3

In Figure 1, the graphs depict an enhancement in the performance as the number of features escalates. Notably, the accuracy between FIIkNN and kNN peaks in the model incorporating 2, 10, and 20 features.

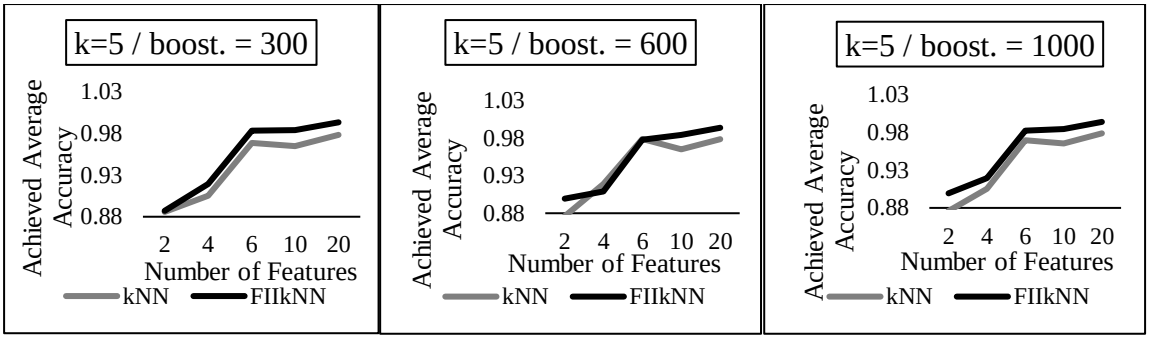

Figure 2. Achieved average accuracies based on increasing number of features for k=5

Figure 2 also displays the increasing success of accuracy while increasing in number of features using in modelling.

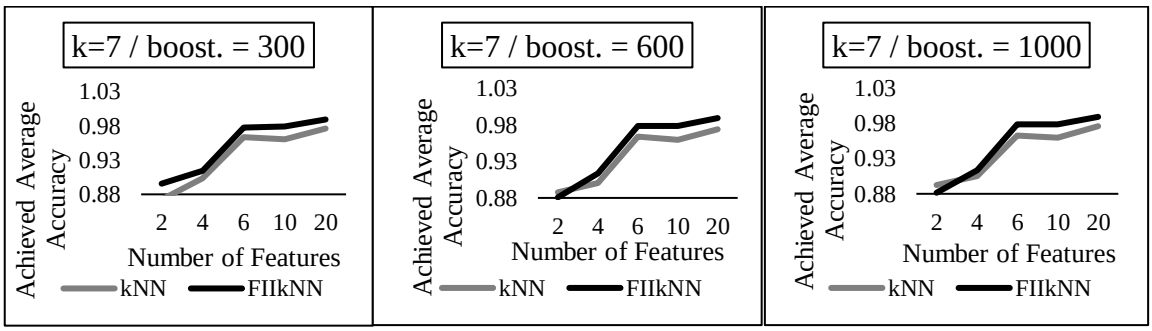

Figure 3. Achieved average accuracies based on increasing number of features for k=7

In Figure 3, consistent with the trends observed in previous graphical representations, demonstrates that the Feature Importance Infused k-Nearest Neighbors (FIIkNN) model consistently outperforms the traditional k-Nearest Neighbors (kNN) algorithm, particularly in terms of prediction results.

5. Conclusions and Discussions

This study's extensive analysis has elucidated several key insights regarding the performance and characteristics of the Feature Importance Infused k-Nearest Neighbors (FIIkNN) model. the study revealed a consistent trend wherein the accuracy of the FIIkNN algorithm surpassed that of the traditional kNN algorithm as the number of features increased. This trend can be attributed to the FIIkNN model's enhanced ability to integrate and leverage information from a larger set of features, thus offering a more refined and accurate classification. This observation underscores the model's

potential in handling complex, multi-variable datasets, which is invaluable in advanced predictive modeling tasks. In instances where the success rate surpassed 90% with only four categories, a significant element of randomness was evident. Despite high recommendations for the model's application under these conditions, the differential impact remained marginal. Intriguingly, this approach demonstrated a propensity for superior performance with variables that had lower explanatory power for the target, as opposed to those with a higher explanatory influence, compared to k classical NN. This suggests the model's unique ability to extract and utilize subtle patterns within the data. Additionally, an inverse relationship was observed between the number of neighbors (k) and the success rate, indicating a decrease in accuracy as k increased. This highlights the importance of optimal k-value selection for maximizing the model's efficacy. The role of feature importance was also found to be crucial in determining the accuracy outcomes. Situations where feature importance values were closely similar resulted in minimal accuracy disparities. Conversely, significant differences in feature importance values led to notable variations in accuracy, underscoring the impact of feature prioritization in the model.

Furthermore, the analysis of the relationship between the increasing number of features and resultant accuracy is paramount in understanding the scalability and adaptability of the FIIkNN model. The patterns observed in accuracy relative to feature count, as detailed in this study, provide essential insights into the model’s performance across various dimensionalities. These findings not only attest to the FIIkNN model’s utility in managing complex, feature-dense datasets but also highlight its prospective application in sophisticated machine learning scenarios, marking a significant contribution to the field.

Moreover, this study's insights into the Feature Importance Infused k-Nearest Neighbors (FIIkNN) model offer profound implications for the sectors of energy systems and renewable energy financing. Specifically, the model's enhanced classification accuracy and adeptness in handling multi-variable datasets underscore its potential to revolutionize decision-making processes in these fields. By enabling more precise segmentation and prediction, the FIIkNN model could significantly contribute to optimizing investment strategies and risk assessments, thereby facilitating more efficient allocation of resources towards sustainable energy projects.

6. Future Works and Recommendations

This research lays the foundation for numerous prospective investigations and advancements in the realm of machine learning algorithms. While the current study primarily utilizes the Euclidean distance metric, future research could benefit from examining the applicability and performance of the proposed Feature Importance Infused k-Nearest Neighbors (FIIkNN) model using alternative distance metrics, such as the Manhattan and Minkowski metrics. Such exploration would significantly broaden our understanding of the model’s versatility and effectiveness when applied in varied metric contexts. Another promising direction for future research involves developing and incorporating a mechanism to ascertain the optimal number of neighbors (k) for the model. This addition could potentially enhance the model's predictive accuracy and efficiency, tailoring it more closely to specific dataset characteristics. Furthermore, the prospect of creating a hybrid model that amalgamates the principles of feature importance with the methodologies of differential evolution presents an intriguing possibility. This hybridization could yield a more refined and powerful version of the model, one that leverages the strengths of both approaches to optimize performance and increase adaptability. Given

the complexity and potential impact of such integrations, a dedicated research effort, possibly culminating in a separate scholarly article, would be beneficial. This focused study would allow for a deeper investigation into the nuances and synergistic effects of combining the FIIkNN approach with other established machine learning techniques. The exploration of these potential advancements not only promises to contribute significantly to the field of machine learning but also opens the door to more sophisticated and accurate predictive models, catering to the ever-evolving demands of data-driven decision-making.

References

[1] Kalaiarasi, K., et al., Optimization of the average monthly cost of an EOQ inventory model for deteriorating items in machine learning using PYTHON, *Thermal Science*, 25 (2022), Spec. issue 2, pp. 347-358

[2] Cheng, Debo., et al., k NN algorithm with data-driven k value, *Proceedings*, 10th, Advanced Data Mining and Applications: 10th International Conference, Guilin, China, 2014, pp. 499-512

[3] Zhang, S., Challenges in KNN Classification, *IEEE Transactions on Knowledge and Data Engineering*, 34 (2022), 10, pp. 4663-4675

[4] Dastile, X., *et al*., Statistical and machine learning models in credit scoring: A systematic literature survey, *Applied Soft Computing*, 91 (2020), pp. 106263

[5] Mladenova, T., A Feature-Weighted Rule for the K-Nearest Neighbor, *Proceedings*, 5th, International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Bolu, Turkey, 2021, pp. 493-497

[6] Liang, J., An ensemble method, *Proceedings*, 4th, International Conference on Communication and Information Processing, New York, USA, 2018, pp. 186–190

[7] Huang, J., *et al.*, An Improved kNN Based on Class Contribution and Feature Weighting, *Proceedings*, 10th, International Conference on Measuring Technology and Mechatronics Automation, Changsha, China, 2018, pp. 313-316

[8] ***, School of Computing and Information Sciences, http://archive.ics.uci.edu/ml

[9] Liangxiao, J., *et al*., Bayesian citation-KNN with distance weighting, *International Journal of Machine Learning and Cybernetics*, 5 (2014), 2, pp. 193-199

[10] Biswas, N., *et al.*, A parameter independent fuzzy weighted k-nearest neighbor classifier, *Pattern Recognition Letters*, 101 (2018), pp. 80-87

[11] Peng, X., *et al*., An improved weighted K-nearest neighbor algorithm for indoor localization, *Electronics*, 9 (2020), 12, pp. 2117

[12] Ertuğrul, Ö. F., A novel distance metric based on differential evolution, *Arabian Journal for Science and Engineering*, 44 (2019), pp. 9641-9651

[13] Alsakka, R., Gwilym, O., Leads and lags in sovereign credit ratings, *Journal of Banking & Finance*, 34 (2010), 11, 2614-2626

[14] Ahmed, S. E., Çetin, A. İ., Determinants of Credit Ratings and Comparison of the Rating Prediction Performances of Machine Learning Algorithms, *Proceedings*, 17th, In *E3S Web of Conferences* 2023, Cape Town, South Africa, Vol. 409, p. 05013

[15] Ekmekcioglu, M., *et al*., Predicting Sovereign Credit Ratings Using Machine Learning Algorithms, *Proceedings*, 1st, Industrial Engineering in the Covid-19 Era: Selected Papers from the Hybrid Global Joint Conference on Industrial Engineering and Its Application Areas, GJCIE 2022, Switzerland, 2023, pp. 52-61

[16] Takawira, O., Mwamba, J. W. M., Sovereign Credit Ratings Analysis Using the Logistic Regression Model, *Risks*, 10 (2022), 4, pp. 70-93

[17] ***, Worldbank Databank, https://databank.worldbank.org/home.aspx

[18] ***, Human Development Reports, https://hdr.undp.org/data-center

[19] Ali, N., *et al.,* Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets, *SN Applied Sciences*, 1 (2019), pp.1-15

[20] Obiedat, R., *et al.,* An Intelligent Hybrid Sentiment Analyzer for Personal Protective Medical Equipments Based on Word Embedding Technique: The COVID-19 Era, *Symmetry*, 13 (2021), 12, pp. 2287

[21] Gothai, E., *et al.,* Map-Reduce based Distance Weighted k-Nearest Neighbor Machine Learning Algorithm for Big Data Applications. Scalable Computing, *Practice and Experience*, 23 (2022), 4, pp. 129-145

[22] Bajpai, A., *et al.,* Performance enhancement of automatic speech recognition system using euclidean distance comparison and artificial neural network, *Proceedings*, *3th*, International Conference On Internet of Things: Smart Innovation and Usages (IoT-SIU), IEEE 2018, pp. 1–5

[23] Abdulrahim, H., *et al.,* Machine learning models to prediction OPIC crude oil production, *Thermal Science*, 26 (2022), Spec. issue 1, pp. 437-443

Submitted: 11.11.2023.
Revised: 30.01.2024.
Accepted: 01.02.2024.