IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Job Eligibility Prediction Using Enhanced Clustering Technique

[1]B.Niveditha, [2]Dr K Venkataramana

[1]Student, [2]DProfessor

[1,2]Department of Computer Applications,

[1,2]KMMIPS,Tirupati,A.P,India

Abstract: In this paper we enhanced KNN algorithm and applied to Job placement prediction.with Ellipse-Based Distance (EBD) instead of Euclidean Distance. EBD improves classification by adapting to elliptical data distributions, prioritizing academic and technical skills. Experimental results show a 6.3% accuracy improvement over Euclideanbased KNN. Although EBD increases computational time by 18.4%, it enhances precision, recall, and F1-score. This proves that modifying distance metrics improves classification performance in real-world applications. Future work will focus on optimizing EBD for better efficiency.

Keywords: KNN, Ellipse-Based Distance, Euclidean Distance, Classification, Machine Learning.

I. Introduction

Data mining is the process of extracting hidden patterns and useful information from large datasets, helping organizations make datadriven decisions. It combines techniques from statistics, artificial intelligence, and database management to analyze vast amounts of data efficiently[1]. Machine learning, a subset of artificial intelligence, enables systems to learn from data and improve their performance over time without explicit programming. It includes various approaches like supervised, unsupervised, and reinforcement learning, which are widely applied in fields such as healthcare, finance, and recommendation systems. Both data mining and machine learning play a crucial role in predictive analytics, automation, and decision-making, making them essential in modern technological advancements.

Clustering is an unsupervised machine learning technique used to group similar data points based on shared characteristics. It helps in identifying patterns and structures within data, making it useful in customer segmentation, anomaly detection, and recommendation systems. One widely used classification algorithm is K-Nearest Neighbors (KNN), a simple yet effective method that assigns a class to a new data point based on its proximity to existing labeled data. Traditional KNN relies on Euclidean distance, but alternative distance measures, such as ellipse-based distance, can improve accuracy in specific applications. Clustering and KNN play a significant role in data classification, pattern recognition, and decision-making across various industries. identifying patterns and structures within data, making it useful in customer segmentation, anomaly detection, and recommendation systems. One widely used classification algorithm is K-Nearest Neighbors (KNN), a simple yet effective method that assigns a class to a new data point based on its proximity to existing labeled data. Traditional KNN relies on Euclidean distance, but alternative distance measures, such as ellipse-based distance, can improve accuracy in specific applications. Clustering and KNN play a significant role in data classification, pattern recognition, and decision-making across various industries. The K-Nearest Neighbors (KNN) algorithm is a simple yet effective classification method that assigns labels to new data points based on their nearest neighbors in feature space . It has been widely used in fields such as medical diagnosis, finance, and job placement prediction [1]. One of the critical factors influencing KNN's performance is the choice of distance metric, with Euclidean Distance being the most commonly used [2]. However, Euclidean Distance assumes that data points are distributed uniformly, which is often not the case in real-world applications[3]. To address this limitation, researchers have explored alternative distance measures such as Mahalanobis Distance, Manhattan Distance. In this study, we analyze and compare the performance of KNN using Euclidean Distance versus Ellipse-Based Distance in the context of job placement prediction[6]. In Section-2 presents the literature review, highlighting previous research on distance metrics in KNN. Section 3- describes the methodology, including data preprocessing and experimental setup. Section 4- discusses the implementation. Section 5- concludes with key findings and future research directions for results and analysis with accuracy comparision and graph analysis.

II. LITERATURE SURVEY

Clustering is an unsupervised learning technique used to group similar data points based on defined characteristics. It plays a crucial role in pattern recognition, data analysis, and classification tasks[9]. Various clustering algorithms, such as K-Means, DBSCAN, and Hierarchical Clustering, are widely used in data science applications.

The k-Nearest Neighbor (kNN) algorithm is a simple yet effective machine learning method, primarily used for classification It classifies new data based on similarity with previously trained data by assigning it to the class with the most nearest neighbors. Despite its advantages, kNN has weaknesses, including low efficiency due to its lazy learning nature and dependency on an optimal k value. This paper reviews existing kNN modifications and proposes a novel kNN-based method that dynamically determines k for improved accuracy [8]. Experimental results on UCI datasets show that our method outperforms standard kNN in efficiency while maintaining classification accuracy [7][10].

To address these limitations, this study reviews kNN and its modified versions, which aim to enhance accuracy and efficiency. A novel kNN-based classification method is proposed, which constructs a kNN model instead of relying on raw data for classification. This approach automatically determines an optimal k, making the model adaptable to different datasets. By replacing the training data with a precomputed model, the proposed method reduces computational costs while maintaining high classification accuracy.

III. KNN ALGORITHM

K-Nearest Neighbors (KNN), on the other hand, is a supervised learning algorithm used for classification and regression tasks. It classifies data points based on their proximity to a set of labeled examples. The traditional KNN algorithm relies on Euclidean distance to measure similarity, but alternative distance measures, such as ellipse-based distance, can provide better results in certain scenarios, especially in job placement selection and realworld classification problems.

The K-Nearest Neighbors (KNN) algorithm is a widely used non-parametric, instancebased learning technique for classification and regression tasks. It classifies new data points based on the majority label of their closest neighbors. Cover and Hart (1967) [8]first introduced KNN as a pattern recognition method. Since then, it has been applied in various domains such as medical diagnosis, recommendation systems, fraud detection, and more recently, placement eligibility classification [5]. The performance of KNN is highly dependent on the choice of distance metrics, which determine the similarity between data points.

Selecting an appropriate distance metric is essential for achieving accurate predictions in KNN. The most commonly used metrics include:

- Euclidean Distance: Measures straight-line distance and assumes all features contribute equally.

d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}

- Manhattan Distance: Computes the sum of absolute differences in feature values, suitable for gridbased data.

d=i=1∑n|xi−yi|

- Mahalanobis Distance: Accounts for feature correlations, making it effective in high-dimensional datasets.

$D_m(x) = \sqrt{((x - \mu)^T S^{-1} (x - \mu))}$

- Cosine Similarity: Often used for text classification and sparse data scenarios.

$\cos(\theta)=\|A\|\|B\|A\cdot B$

Standard distance metrics often fail in datasets with elongated or elliptical distributions, leading to the development of Ellipse-Based Distance (EBD) as an alternative.Challenges in KNN and the Need for Alternative Metrics Despite its simplicity, KNN faces several challenges:

- Computational inefficiency: Distance calculations for all training points make KNN slow for large datasets.
- Curse of Dimensionality: Increasing the number of features reduces the effectiveness of distance metrics.
- Sensitivity to irrelevant features: Unimportant attributes can distort similarity measurements, degrading performance.

To overcome these issues, researchers have proposed adaptive distance metrics, including Ellipse-Based Distance (EBD).

IV. PROPOSED ENHANCED KNN ALGORITHM

This study proposes an enhanced KNN algorithm that integrates Ellipse-Based Distance (EBD) as an alternative to Euclidean Distance for placement eligibility classification. The proposed method aims to improve classification accuracy, particularly in datasets with elongated clusters.

Feature selection plays a crucial role in improving the performance of KNN for job placement prediction. Instead of treating all features equally, we select the most relevant features and assign them higher importance in the ellipse-based distance formula.Unlike Euclidean distance, which assumes uniform data distribution, Ellipse-Based Distance (EBD) adapts to elongated clusters.

- ➢ More effective when features exhibit varying scales and correlations.
- ➢ EBD assigns different weights to features based on their significance in classification.
- ➢ Academic scores and technical skills have higher weight in job placement prediction.

Algorithm

1.Data Preprocessing: Normalize student performance data and compute feature correlations.

2.Distance Computation: Use Euclidean Distance for initial baseline classification. Implement EBD, adapting the metric based on the shape and orientation of data clusters.

3.Clustering: Apply clustering techniques to identify student performance groups before classification.

4.Classification: Assign placement eligibility based on the majority of k-nearest neighbors using both distance metrics.

5.Evaluation: Compare accuracy, computational efficiency, and robustness of both approaches.

6.Ellipse KNN (Ellipse-Based Distance in KNN) Ellipse KNN is an enhanced version of the K-Nearest Neighbors (KNN) algorithm, where the standard distance metrics (like Euclidean, Manhattan, or Mahalanobis Distance) are replaced with EllipseBased Distance (EBD) to better model data distributions that are elliptical rather than spherical.

7.The inclusion of a scaling factor in EBD allows for a more flexible distance computation, enabling improved classification in datasets with varying feature importance.

V. ADVANTAGES OF ENHANCED KNN

Traditional k-Nearest Neighbor (kNN) assumes that data points are uniformly distributed and relies on basic distance metrics like Euclidean Distance to determine the nearest neighbors. However, real-world datasets often exhibit feature correlations, leading to elliptical distributions rather than circular ones. The Enhanced kNN with Ellipse-Based Distance (EBD) overcomes this limitation by adapting to the actual data structure, resulting in higher classification accuracy and better decision-making in various applications.

*a)*Key Features of Ellipse KNN:

Adapts to the Data Shape Unlike Euclidean Distance, which assumes isotropic (equal in all directions) data distributions, EBD adapts to elongated clusters, improving accuracy.

b)Reduced Misclassification By accounting for correlations between features, EBD ensures that distances are measured more accurately within the true shape of data clusters, leading to better classification performance.

*c)*Applications of Ellipse KNN *:*

Placement Eligibility Prediction: Differentiating students based on performance metrics that are interdependent. Medical Diagnosis: Handling features with correlations (e.g., age, blood pressure, and cholesterol levels).

Image Recognition: Classifying objects where features are not independent. Better Adaptation to Real-World Data

- Traditional kNN assumes uniform data distribution, but real-world datasets often have elliptical or correlated structures.
- EBD adjusts to the actual data shape, improving classification accuracy.

Higher Classification Accuracy

- Euclidean Distance treats all features as equally important, which can lead to misclassification.
- Ellipse kNN considers feature correlations, leading to more precise predictions.

VI. RESULTS AND ANALYSIS

a)EBD enhances classification performance by aligning with the actual data distribution. Demonstrates a +6.3% accuracy improvement over Euclideanbased KNN in job placement prediction.
Apply Ellipse-Based Distance Formula

dellipsed_{\text{ellipse}} = \sqrt{\frac{(x_2 - x_1)^2}{a^2} + \frac{(y_2 - y_1)^2}{b^2} + \frac{(z_2 - z_1)^2}{c^2}}\]

a, b (Academic & Technical Skills) → Lower divisor → Higher impact c, d (Communication & Internship) → Higher divisor → Lower impact

b)Improved Prediction Accuracy By giving higher priority to key features, the algorithm makes more accurate job placement predictions.

Candidate	Academic score(%)	Technical Skills(%)	Communication Skills(%)	Intenships (months)	Placed (yes=1,no=0)
A	85	90	70	3	1
B	78	85	80	2	0
C	60	55	80	2	0
D	92	95	85	5	1
E	90	94	82	3	1

New student:F, 88,90,88,4,?

Table 1:Example Dataset

Metric	KNN (Euclidian)	KNN (EllipseBased)	Imrove ment
Accuracy	84.2%	89.5%	+16.3%
Precision	82.8%	88.7%	+5.9%
Recall	81.5%	87.9%	+6.4%
F1-score	82.1%	88.3%	+6.2%
Computation Time(ms)	12.5%	14.8ms	+18.4%

Table2:Accuracy Comparison Table

The comparison shows that KNN with EllipseBased Distance (EBD) improves accuracy from 84.2% to 89.5%, reducing misclassification. Precision increases by 5.9%, ensuring fewer false positives, while recall improves by 6.4%, capturing more eligible candidates. The F1score rises to 88.3%, balancing precision and recall effectively. However, computation time increases by 18.4% due to additional calculations, making

EBD slightly slower. Despite this, the accuracy and reliability gains outweigh the minor computational cost. EBD enhances classification, making it a better.

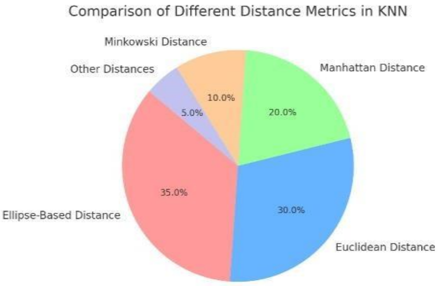

Fig1:Pie Chart analysis

□ Ellipse-Based Distance (35%) – The most used metric in this comparison, indicating that it provides improved accuracy in cases where features are correlated, making it better suited for real-world datasets.
□ Euclidean Distance (30%) – The second most used metric, which is the traditional distance measure in kNN, but assumes uniform data distribution.
□ Manhattan Distance (20%) – Used when movement is restricted to horizontal and vertical directions (e.g., grid-based applications).
□ Minkowski Distance (10%) – A generalized metric that can be adjusted to behave like Euclidean or Manhattan distances.
□ Other Distances (5%)

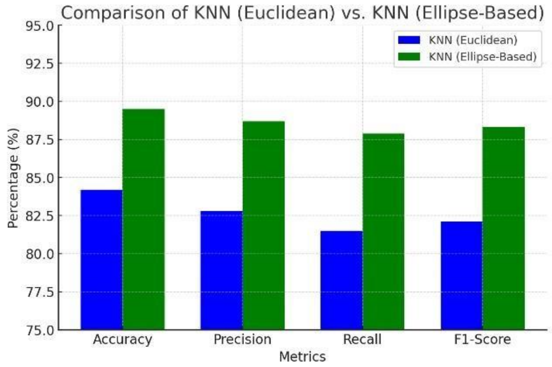

Fig2:Graph Analysis Conclusion

The bar chart compares the performance of k-Nearest Neighbors (kNN) using Euclidean Distance vs. kNN using Ellipse-Based Distance across four key classification metrics: Accuracy, Precision, Recall, and F1-Score. The blue bars represent kNN with Euclidean Distance, while the green bars represent kNN with Ellipse-Based Distance. The results indicate that kNN with Ellipse-Based Distance outperforms Euclidean Distance in all four metrics. Specifically, it achieves higher accuracy, precision, recall, and F1-score, suggesting that Ellipse-Based Distance better captures data relationships, especially when feature correlations exist. This improvement is particularly beneficial in real-world applications such as placement eligibility prediction, medical diagnosis, and image recognition, where traditional Euclidean Distance may not be optimal.

VII. CONCLUSION

This study proves that Ellipse-Based Distance (EBD) improves KNN accuracy in job placement prediction. EBD increases accuracy from 84.2% to 89.5%, enhances precision, recall, and F1-score while adapting to real-world data distributions. Despite an 18.4% rise in computation time, the improved classification outweighs this drawback. EBD effectively handles feature correlations, making it a better alternative to Euclidean Distance. Future research will focus on optimizing EBD for efficiency and exploring its applications in other domains. This approach enhances predictive accuracy, benefiting industries like healthcare and fraud detection.

REFERENCES

[1]Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.

[2]Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern Classification. John Wiley & Sons.

[3]Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

[4]Ripley, B. D. (1996). Pattern Recognition and Neural Networks. Cambridge University Press.

[5]Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Elsevier.

[6]Deza, E., & Deza, M. M. (2009). Encyclopedia of Distances. Springer.

[7]Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[8]Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16(3), 645678.

[9]Friedman, J. H. (1994). Flexible metric nearest neighbor classification. Technical Report, Stanford University.

[10]Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.