Enhancing Employee Performance Management

A Data-Driven Decision Support Model using Machine Learning Algorithms

Zbakh Mourad[1], Aknin Noura[2], Chrayah Mohamed[3], Bouzidi Abdelhamid[4]
FS Tetuan, Abdelmalek Essaadi University, Tetuan, Morocco[1, 2, 4]
ENSA Tetuan, Abdelmalek Essaadi University, Tetuan, Morocco[3]

***Abstract*—Human resource management (HRM) plays a crucial role in the effective functioning of modern businesses. However, as the volume of data continues to increase, HR professionals are facing growing challenges in objectively gathering, measuring, and interpreting human resources data. The research problem addressed in this study is the need to improve methods for the objective classification of teams based on the most relevant performance factors considering the subjectivity of current tools. To tackle this issue, the research questions focus on the possibility of developing an efficient model for team classification using supervised machine learning algorithms. This study consists of developing and validating three team classification models using the support vector machine (SVM), the K-nearest neighbor (KNN) algorithm, and the multiple linear regression algorithm (MLR) after using PCA for data reduction. Following extensive validation, the module based on MLR was identified as the most effective, achieving an accuracy of 87.5% in Predicting employee performance, which makes it possible to anticipate and fill employee skills gaps and optimize recruiting efforts. This work provides human resources professionals with a data-driven decision support to enhance Human Resources Management using Machine Learning.**

Keywords*—HRM; HR analytics; Employee Performance Prediction; Support Vector Machine (SVM) Algorithm; K-Nearest Neighbor (KNN) Algorithm; Multiple Linear Regression (MLR) algorithm; Principal Component Analysis (PCA)

I. Introduction

Human resources management (HRM) is considered one of the most strategic functions of a company as it plays a vital role in boosting productivity and competitiveness. People are at the center of overall performance improvement, and companies must use the heterogeneity of their resources to create a competitive advantage. The development of skills such as Valuable, Rare, Inimitable and Non-substitutable (VRIN) is crucial for companies to align their resources with the overall business strategy [1].

With the era of digital transformation, managing team performance has become a complex and complicated task. Traditional IT tools are unable to collect and analyze the mass of data available through several new sources. However, companies that invest in big data software to understand and improve the performance of employees are likely to achieve organizational goals gain a competitive edge [2], including competitive advantages, identifying talents, and retaining high performers, better understanding of low performers, simulating the performance of candidates during recruitments, playing a strategic role in the structure of teams, and prioritizing HR investments to achieve greater work performance.

The deployment of HR analytics in HRM is no longer an option, and companies that ignore the revolution of digital technology risk being left behind. By deploying machine learning algorithms, companies can avoid significant costs if these mines of information are not exploited, namely, the costs associated with replacing employees especially when key skills are lost, the cost of hiring new staff, and the cost of training for replacements are important [3].

However, despite long-standing efforts to objectively assess performance to demonstrate factors that impact the performance of employees and quantify its impact on business outcomes, namely the effectiveness of training [4], but to this day, none of this work has been able to identify more than two factors and subsequently predict the performance results of a team using machine learning algorithms, identifying the specific factors that predict team performance remains a challenge.

To fill this gap, considering this context which makes increased competition, this study aims to respond to the need to improve performance management using machine learning algorithms, the results make possible the prediction of the performance of employees. the study consists of developing methods to objectively classify teams according to the most relevant performance factors, unlike the subjectivity of evaluation based primarily on interviews. three team classification models are developed and validated using support vector machine (SVM), K-nearest neighbor (KNN), and multiple linear regression (MLR) algorithm after using the principal component analysis (PCA) for the reduction of a dataset published by HRM professors at the New England College of Business, including 36 variables linked to 311 employees used to train and test the models. After evaluating the results, the MLR-based model appears to be the most effective, with an accuracy of 87.5% in predicting employee performance, making it possible to anticipate skills gaps and optimize recruitment efforts. This research provides a data-driven decision support tool to improve human resource management through machine learning.

In this paper, a structured approach to present our research is used. Firstly, in Section II, the relevant literature will be reviewed. Then, in Section III, the proposed approach and the dataset will be defined, and the steps involved in constructing the model will be outlined. The results obtained from the approach will also be presented. Next, in Section IV, the results obtained from the models will be evaluated. Finally, in Section V, the paper will be concluded.

II. Related Work

The review of the relevant literature is structured as follows: firstly, the evolution of human resource management and its technologies, with a particular focus on HR analysis, is examined. Next, the Performance appraisal process and its factors are delved into. Finally, the existing works that have employed machine learning algorithms in evaluating employee performances are explored.

A. Human Resource Management and HR Analytics

As per the literature, human resources management is a collection of practices that are employed to administer, mobilize, and develop human resources involved in the organization's activities to align them with the overall business strategy. In today's modern organizations, human resources have become the key to success, making the practices of human resources management crucial for the company's overall performance, especially in an era characterized by intense competition, globalization, and internationalization of markets.

The explosive growth of data in various fields of industry has made gathering, measuring, and interpreting HR data a complex and challenging task. As a result, new advanced practices and technologies have emerged, leading to the rise of human resources analytics (also known as people analytics) as a separate sub-field of business analytics [5].

According to the literature, HR analytics is defined as the collection and application of talent data to improve critical talent and business outcomes. HR analytics leaders enable HR leaders to develop data-driven insights that inform talent decisions, improve workforce processes, and promote a positive employee experience [6].

According to the information provided, Gardner's model, depicted in Fig. 1, highlights various aspects of HR Analytics, which include:

- Descriptive analytics: This dimension involves examining HR data to answer the question of "What happened?
- Diagnostic analytics: Diagnostics reveal the underlying causes of the events presented by descriptive data and answer the question of "Why did it happen?"
- Predictive analytics: The most important dimension of HR Analytics, which focuses on what might happen in the future based on the details of past events using statistical modeling (Machine learning).
- Prescriptive analytics: This dimension suggests data-driven options or actions to take based on the predictions. Unlike classic human decisions that are often subject to the process of gut feeling and illogical biases, it guides what to do in a particular situation based on given HR data.

As HR analytics emerged as a new trend, it has garnered significant attention and budget allocation. Numerous studies have been conducted to investigate its role, potential opportunities, and challenges associated with its implementation. [7], [8], [9], [10].

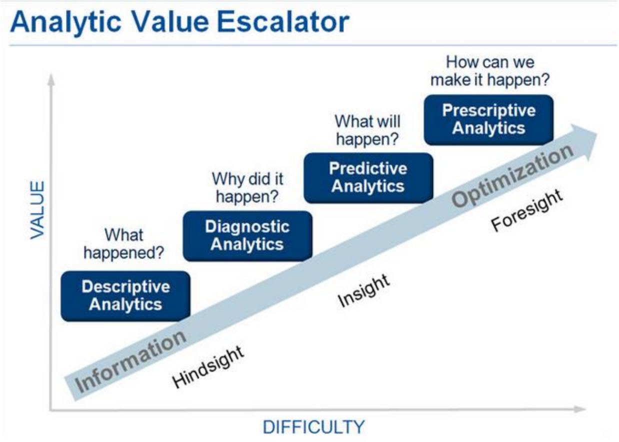


Fig. 1. Analytic value escalator (gardner's model).

Regarding potential opportunities, HR analytics can offer valuable insights into various issues such as attrition, strategic decisions related to performance management, and investments in training programs, as depicted in Fig. 2.

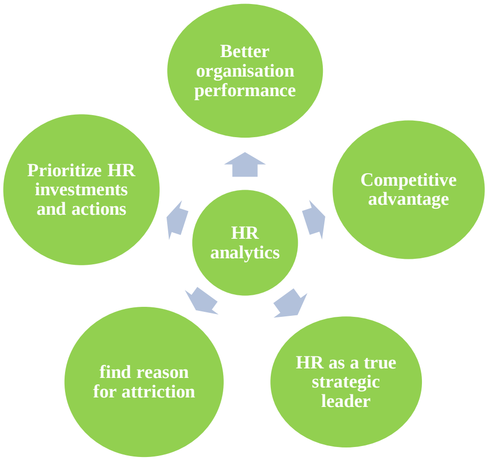

Better organisation performance
Competitive advantage
HR as a true strategic leader
find reason for attrition
Prioritize HR investments and actions
HR analytics

Fig. 2. Opportunities of HR analytics.

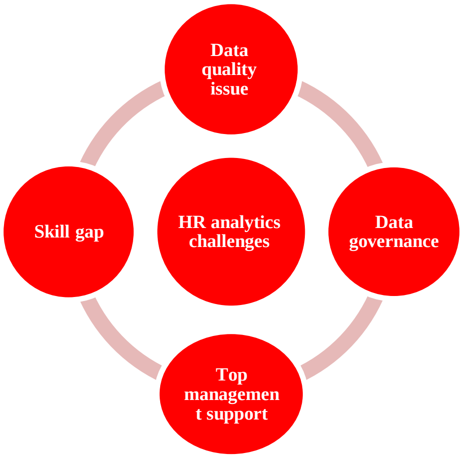


Fig. 3. Challenges of using HR analytics.

Despite the increasing attention that HR analytics has received, it is still in its early stages, and only a limited number of models have been developed so far, as noted in study [11]. Many studies have attempted to identify the challenges that organizations face when implementing HR analytics, as illustrated in Fig. 3.

B. Performance Management

After conducting a literature review, it was found that performance is a complex and multifaceted concept that is difficult to define, as noted in study [12]. However, in an industrial organizational context, performance is typically associated with excellence and is defined as an official report that records a result achieved at a specific moment in time, in a particular setting, based on objectives and expected outcomes measured using various indicators. This definition is closely tied to the company's vision, strategy, and objectives, as highlighted in study in [13] and [14]. It is therefore essential to develop an instrument for assessing job performance, as performance appraisal and management of employees and teams can be subject to perception and subjectivity if not based on data.

As such, performance management is a continuous process that aims to make informed decisions to achieve optimal outcomes by identifying and addressing problems and utilizing appropriate tools for measurement, as depicted in Fig. 4, and noted in study [15].

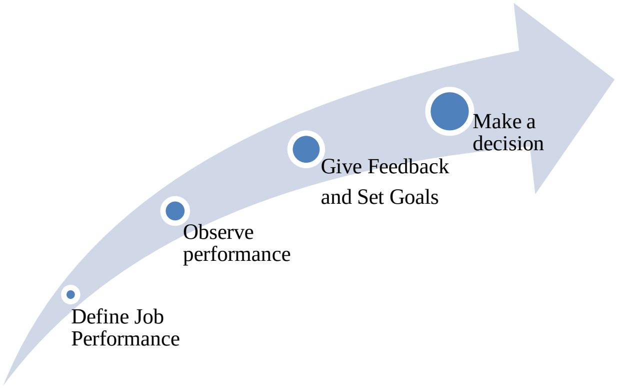


Fig. 4. Performance appraisal process.

C. Performance Assessment

Numerous research studies have investigated the factors that influence job performance. Training is one such factor, and the study conducted by Joshua S. Bendickson and Timothy D. Chandler in 2019 [16] served as the inspiration for our own research. Additionally, other studies have explored the impact of factors such as workforce diversity [17], leadership style [18], determinants of employee engagement [19], the role of employee satisfaction as a mediator of compensation and career development [20], and the effect of organizational communication and culture [21] on job performance.

Despite extensive research, the exact factors that influence job performance have not been pinpointed. In 2013, a study called "The Analytics Era" examined more than 200 indicators and concluded that the most appropriate indicators vary depending on the company and activity. Therefore, HR must prioritize aligning HR analytics with the business priorities of the company to select the appropriate indicators.

Several studies have proposed models based on machine learning algorithms that predict employee performance based on HR data. For example, Iwamoto et al. [22] proposed a model based on multiple regression that evaluates individual performance based on the financial outcomes of employees that influence the performance of the organization. This represents a significant step towards an objective assessment of individual performance. Later, Abdullah et al. [23] established a model that assesses individual performance against knowledge and skills through a case study in Malaysia, wherein the analytical hierarchy process is used to integrate the multifaceted preferences of the five criteria of human capital to determine the importance of the four identified indicators.

Furthermore, Chen and Chen [24] and QA Al-Radaideh, E Al Nagi [25], applied a data mining algorithm based on decision trees and association rules to employee characteristics and performance. In a recent study by JM Kirimi and CA Moturi [26], data mining classification was used to predict employee performance. They compared the results of three different machine learning algorithms, namely ID3, C4.5, and Naïve Bayes. This study found that the C4.5 algorithm had the highest accuracy due to several factors that had a significant impact on employee performance. For instance, the experience attribute had the maximum gain ratio. This study serves as a starting point for the research.

III. Methodology

A. Research Design

Comparing several processes for executing machine learning projects, such as KDD, Scrum, Kanban, SEMMA, or TDSP, the Cross Industry Standard Process for Data Mining (CRISP-DM) model has been chosen. This model is the most widely used industry-independent form of data mining since 2017, owing to its various advantages that have resolved existing problems. The CRISP-DM methodology provides a uniform framework for guidelines, planning, and managing a project [27].

The CRISP-DM model is a six-phase process model that encompasses the entire data mining project, from business understanding to deployment, as depicted in Fig. 5. The six phases are as follows:

- Business Understanding: This phase involves identifying the problem, defining the project objectives, and determining the data mining goals. It also includes assessing the resources required for the project.
- Data Understanding: This phase involves collecting and exploring the data to understand its characteristics, quality, and relationships. This provides a foundation for the subsequent phases.
- Data Preparation: This phase involves cleaning, transforming, and integrating the data to prepare it for

modeling. It also includes selecting the appropriate data to use for modeling.

- Modeling: This phase involves selecting and applying appropriate modeling techniques to the prepared data. It includes creating and evaluating multiple models to determine the optimal model for the project.
- Evaluation: This phase involves assessing the performance of the model and determining its effectiveness. It also includes determining if the model meets the project objectives.
- Deployment: This phase involves deploying the model in the production environment and monitoring its performance. It also includes preparing documentation and training materials for stakeholders.

By following the CRISP-DM model, it can be ensured that the machine learning [28] project is well-structured, well-documented, and well-executed, resulting in high-quality and actionable insights.

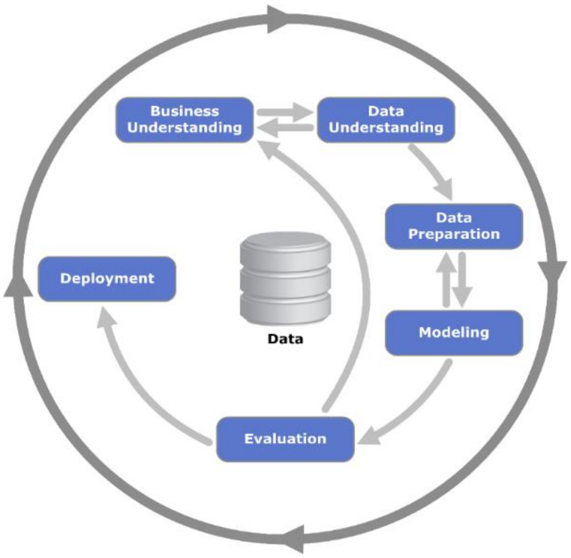


Fig. 5. CRISP-DM process model.

B. *Dataset Description*

Considering the factors that affect job performance, as identified in the literature review, a database published by HRM professors at New England College of Business has been selected. This database comprises 311 employee profiles and includes 36 variables that offer insights into employee performance. Utilizing this database, a comprehensive analysis of the factors that impact employee performance can be conducted, enabling informed decisions based on the results. This will facilitate gaining a deeper understanding of the factors influencing job performance and developing effective strategies to enhance employee performance.

The variables in the database cover various aspects such as income, engagement, satisfaction, projects, number of days an employee was late in the last 30 days, absences, and other HR-related data. The complete list of variables is summarized in Table I.

TABLE I. DESCRIPTION OF THE VARIABLES

Dataset	**Variables**	
	Description	***Data type***
Employee Name	Employee's full name	Text
EmpID	Employee ID is unique to each employee	Text
MarriedID	Is the person married (1 or 0 for yes or no)	Binary
MaritalStatusID	Marital status code that matches the text field MaritalDesc	Integer
EmpStatusID	Employment status code that matches text field EmploymentStatus	Integer
DeptID	Department ID code that matches the department the employee works in	Integer
PerfScoreID	Performance Score code that matches the employee's most recent performance score	Integer
FromDiversityJobFairID	Was the employee sourced from the Diversity job fair? 1 or 0 for yes or no	Binary
PayRate	The person's hourly pay rate. All salaries are converted to hourly pay rate	Float
Termd	Has this employee been terminated - 1 or 0	Binary
PositionID	An integer indicating the person's position	Integer
Position	The text name/title of the position the person has	Text
State	The state that the person lives in	Text
Zip	The zip code for the employee	Text
DOB	Date of Birth for the employee	Date
Sex	Sex - M or F	Text
MaritalDesc	The marital status of the person (divorced, single, widowed, separated, etc)	Text
CitizenDesc	Label for whether the person is a Citizen or Eligible NonCitizen	Text
HispanicLatino	Yes or No field for whether the employee is Hispanic/Latino	Text
RaceDesc	Description/text of the race the person identifies with	Text
DateofHire	Date the person was hired	Date
DateofTermination	Date the person was terminated, only populated if, in fact, Termd = 1	Date
TermReason	A text reason / description for why the person was terminated	Text
EmploymentStatus	A description/category of the person's employment status. Anyone currently working full time = Active	Text
Department	Name of the department that the person works in	Text
ManagerName	The name of the person's immediate manager	Text
ManagerID	A unique identifier for each manager.	Integer
RecruitmentSource	The name of the recruitment source where the employee was recruited from	Text
PerformanceScore	Performance Score text/category (Fully Meets, Partially Meets, PIP, Exceeds)	Text
EngagementSurvey	Results from the last engagement survey, managed by our external partner	Float
EmpSatisfaction	A basic satisfaction score between 1 and 5, as reported on a recent employee satisfaction survey	Integer
SpecialProjectsCount	The number of special projects that the employee worked on during the last 6 months	Integer
LastPerformanceReviewDate	The most recent date of the person's last performance review.	Date
DaysLateLast30	The number of times that the employee was late to work during the last 30 days	Integer

According to the literature review, nine significant variables were identified for building the model: "MarriedID," "GenderID," "Salary," "EngagementSurvey," "EmpSatisfaction," "SpecialProjectsCount," "DaysLateLast30," and "Absences" as dependent variables, and "PerfScoreID" as the independent variable. The data was then split into two datasets, with 248 records used for training the model and 63 records for testing and validating it. By focusing on the most significant variables, a model that accurately predicts employee performance can be developed. Data processing and model building were conducted using the R software, with the HR dataset transformed into a data frame consisting of 311 rows and 36 columns. Data cleaning and checking for missing values were performed to ensure accuracy and completeness, improving the reliability of the model's predictions. With the cleaned data, the model was built using R software, leveraging its powerful data analysis, and modeling capabilities to develop an accurate prediction model for employee performance.

C. Model Building

1) Steps of model construction: To predict employee performance based on their profile, the steps outlined in Fig. 6 were followed. Firstly, the most relevant factors impacting employee performance were identified based on the literature review. Next, the database was prepared by cleaning the data and checking for missing values. To improve result visualization, the data was compressed using PCA, which can synthesize a large dataset compared to other compression techniques, aiding in better data visualization and model accuracy. Three models using the SVM, MLR, and KNN algorithms were established to predict employee performance, each trained using the compressed training set. Finally, the results of each model on the test set were evaluated to determine the most efficient model. By following these steps, an accurate and reliable model for predicting employee performance based on their profile can be developed, facilitating informed decisions and effective strategies to enhance employee performance and achieve organizational goals.

2) Data compression using Principal component analysis: PCA is a dimensionality reduction algorithm that involves transforming interrelated variables, also known as "correlated" variables in statistics, into new variables that are decorrelated from each other. These new variables are known as "principal components" or "principal axes" [29].

By utilizing PCA, the number of variables in the dataset can be reduced and the visualization of the data improved. This is particularly useful when dealing with large datasets that are difficult to visualize or analyze. The principal components generated by PCA can be used to represent the data in a lower-dimensional space, making it easier to analyze and interpret.

Overall, PCA is a powerful technique that can be used to improve the accuracy and efficiency of machine learning models by reducing the number of variables and improving the visualization of the data. By incorporating PCA into the model, A more accurate and reliable model can be developed to effectively predict employee performance based on their profile.

To apply the Principal Component Analysis (PCA) algorithm to the training set, the "FactoMineR" package in R software was utilized. By analyzing the distribution of variables in each factor generated by PCA, factor 1 was interpreted as the "behavior factor," which accounts for 41.42% of the variance, and factor 2 was interpreted as the "achievement factor," which accounts for 35.97% of the variance, as summarized in Table II and Fig. 7.

The behavior factor, factor 1, is the most significant in terms of variance and includes variables that assess days of late and engagement. The achievement factor, factor 2, comprises variables that assess salary and special projects done for the society. Factors with an absolute value greater than 0.40 were considered significant and retained, while those with an absolute value less than 0.40 were deleted for the clarity of the table.

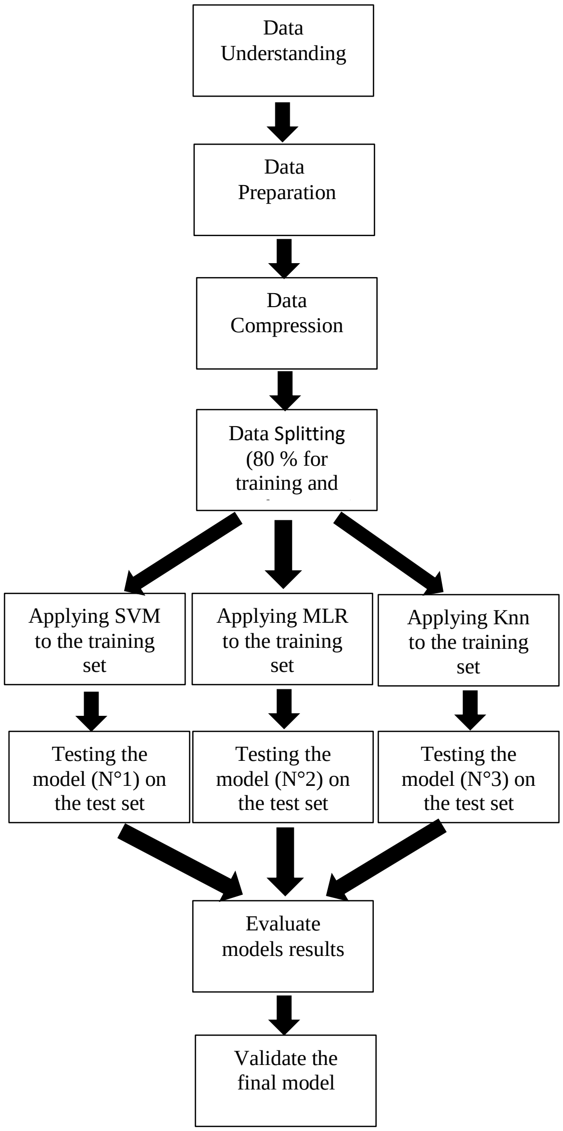

Data Understanding
Data Preparation
Data Compression
Data Splitting (80 % for training and
Applying SVM to the training set
Applying MLR to the training set
Applying Knn to the training set
Testing the model (N°1) on the test set
Testing the model (N°2) on the test set
Testing the model (N°3) on the test set
Evaluate models results
Validate the final model

Fig. 6. Steps of model construction.

TABLE II. COORDINATES FOR THE VARIABLES

Variables	Factors			
	Factor 1	***Factor 2***	***Factor 3***	***Factor 4***
Salary	0.57	0.65	0.47	
EngagementSurvey	0.69	-0.55		0.44
SpecialProjectsCount	0.57	0.66	-0.46	
DaysLateLast30	-0.72	0.52		0.43

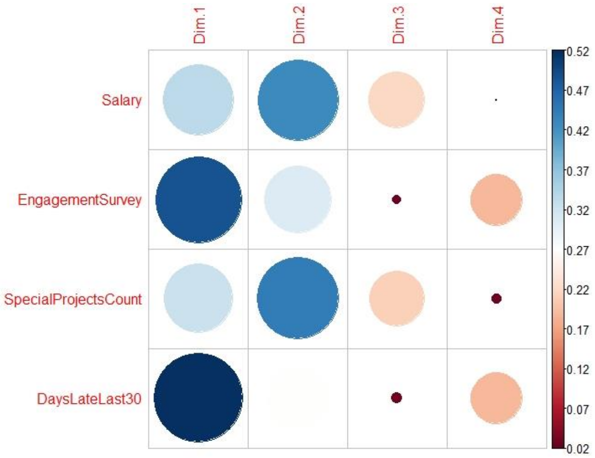


Fig. 7. The variables's contributions.

By utilizing PCA and analyzing the factors generated, better understanding of the underlying variables that impact employee performance can be achieved. This will facilitate the development of more effective strategies to enhance employee performance and achieve organizational goals.

The figure labeled as Fig. 8 provides a visual representation of the distribution of individuals based on the first two factors of the Principal Component Analysis of the chosen data set. The analysis has been performed to identify and understand the variables that have the most significant impact on the data set. The figure helps to illustrate how different employee are distributed based on the two factors and provides insights into how is performing in relation to performance score ID of each one.

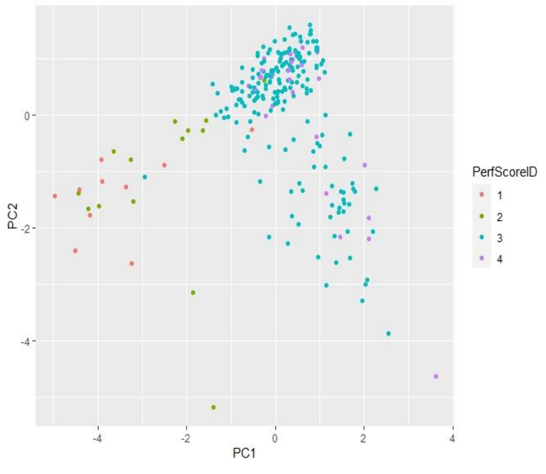


Fig. 8. Individual's graph (PCA).

3) SVM : As the first model, the support vector machine (SVM) algorithm is performed on the compressed database using Principal Component Analysis (PCA). SVMs are a powerful machine learning algorithm applicable for classification, regression, and outlier detection purposes. The basic model of the SVM classifier is a linear classifier processing a set of linearly separable data points with two class labels. It devises an optimal separating surface based on support vectors that maximizes the distance to the nearest training-data point of any class, also known as the functional margin. The optimal separating surface is referred to as a hyperplane or a set of hyperplanes in a higher-dimensional space [30].

Let, $s_a = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ be a training set for two classes, where $x_i \in \mathbb{R}^n$ denotes the input vectors,

$y_i \in \{-1,1\}$ stands for their class label, and N is the number of the observations (samples).

In sum, SVM algorithm help to find computationally the "maximum-margin hyperplane" that divides the group of points x_i for which $y_i = 1$ from the group of points for which $y_i = -1$.

The following diagram in Fig. 9, illustrates these concepts visually:

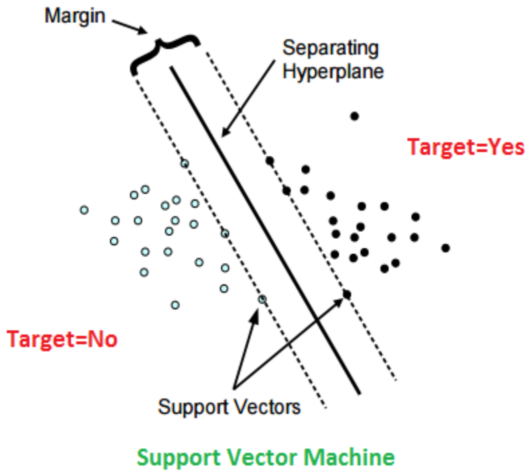


Fig. 9. Linear SVM concept.

To solve nonlinear problems, SVMs can perform a non-linear classification using kernel functions such as polynomial, sigmoid, Gaussian radial basis function (also called RBF or Gaussian kernel) or sigmoid kernel, that converts non-linear separable problems to linear separable problems by adding more dimensions to it, that implicitly mapping their inputs into high-dimensional feature spaces where the problems may be solved linearly. The discriminant function with kernel $K(x, x_i)$ is defined in Eq. (1).

$$f(x) = sgn\{\sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b\} \qquad (1)$$

where, sgn(u) is the sign function where:

if $u > 0$ then $sgn\ (u) = 1$, $if\ u < 0\ then\ sgn\ (u) = -1$;

x is the sample to be recognized; b is called the bias or threshold and α_i is the Lagrange multiplier.

The "e1071" package on the R software was utilized to apply the SVM algorithm to the training set. Upon examining the graph of individuals, it was observed that the data was not linearly separable. Therefore, the decision was made to apply Kernel SVM. For the rest of the study, the Gaussian radial basis function (GRBF) was chosen as the kernel function in the model. This function is defined in Eq. (2) and has shown excellent performance [31]:

$$K(x, x_i) = exp(-\gamma\|x - x_i\|^2) \qquad (2)$$

The figure labeled as Fig. 10 illustrates the results obtained from the model using the GRBF kernel to classify the training dataset based on the first two factors of the Principal Component Analysis. The graph shows how well the model has been able to classify the training data based on the chosen factors and provides valuable insights into its performance.

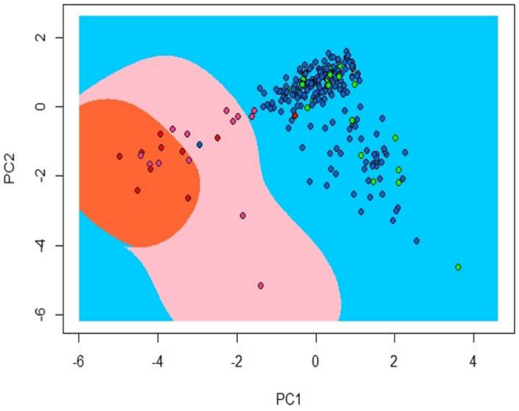


Fig. 10. GRBF SVM results on the training set.

To evaluate the effectiveness of the model, it was applied to the test database and achieved an accuracy rate of 87%. A graphical representation of the results obtained from the model using the Gaussian radial basis function (GRBF) kernel to classify the test dataset based on the first two factors of the Principal Component Analysis has been provided in the figure labeled as Fig. 11. The graph illustrates how well the model has been able to classify the test data based on the chosen factors and provides valuable insights into its performance.

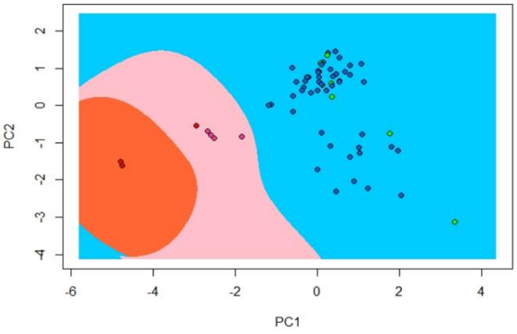


Fig. 11. GRBF SVM results on the test set.

Although the model achieved a high level of precision, it was observed that it was not able to accurately predict the most efficient employees (class 4). However, it was able to accurately predict the other classes. This indicates that the model may not be suitable for identifying the most efficient employees, and there may be other factors that are contributing to their performance that are not accounted for in the model. Further analysis and investigation may be required to identify these factors and improve the accuracy of the model.

4) Multinomial logistic regression: In the second employee performance classification model, multinomial logistic regression (MLR) was employed on the compressed database using Principal Component Analysis (PCA). MLR is a statistical method utilized to model and analyze categorical outcomes with more than two categories. It extends binary logistic regression, enabling the prediction and estimation of probabilities for multiple categories simultaneously.

While linear regression is suited for continuous outcome variables, aiming to find a linear relationship between predictors and the outcome, multinomial logistic regression is designed for categorical outcomes with multiple categories. It estimates probabilities for each category based on predictor variables. Linear regression assumes a linear relationship and normality of residuals, while multinomial logistic regression assumes independence of observations and the absence of multicollinearity. Understanding the nature of the outcome variable and the research question is crucial in choosing between these two methods.

Logistic regression is the most common form of classification algorithm employed, especially in industry. Its range is bounded between 0 and 1, and when the target is categorical, the outcome represents the probability that the output is true [32]. Intuitively, it is a process of modeling the probability of an outcome given an input variable that can be extended into multiple classes.

Multinomial logistic regression involves estimating the coefficients for each independent variable in the model. The estimation is typically performed using maximum likelihood estimation, which aims to find the values of the parameter from the training set that maximize the likelihood of observing the given set of outcomes. Once the coefficients are estimated, they can be used to predict the probabilities of each outcome category for new observations.

The mathematics of the classifier relies on the outcome to distribution P(Y|X) where Y is a dependent variable and X={x_1,…,x_n } is independent variable. Due to applying a nonlinear log transformation using the logistic function called also sigmoid function [33], the parametric model of Logistic Regression can be written as in Eq. (3):

$$P(Y = 1|X, W) = \frac{1}{1+e^{(w_0+\sum_{i=1}^{n} w_i x_i)}} \qquad (3)$$

When there are more than two categories, as is the case in the study, multinomial logistic regression is utilized, which is a powerful statistical classification algorithm that generalizes logistic regression to multiclass problems.

Intuitively, since y = {0,1...n}, the problem is divided into n+1 binary classification problems, where in each one, the probability that y is a member of one of the classes is predicted. This process is repeated, applying logistic regression to each case, and then using the hypothesis that returned the highest value as the prediction.

The output obtained is a probability vector Y, containing probabilities {y,…,y_k } for the k target classes, since the total probability of all the possible events in a system is always 1. Finally, the outcome with the highest probability will be the predicted outcome for the given feature set [34], as shown in Eq. (4) and Eq. (5).

$$\mathrm{y_i} = \mathrm{P(y = i|x, w)} \tag{4}$$

$$Y = max(y_i) \tag{5}$$

The figure labeled as Fig. 12 illustrates the results obtained from the model using MLR based on the first two factors (PC1 and PC2) of the training set. The graph depicts how well the model has been able to classify the training data based on the chosen factors and provides valuable insights into its performance.

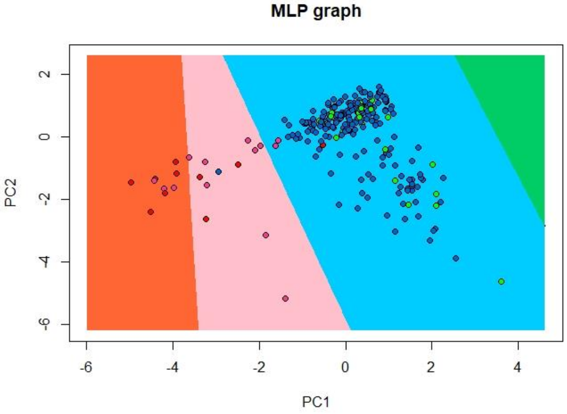


Fig. 12. Multinomial logistic regression on the compressed training set.

Based on the outcomes, the model can differentiate between low performers (indicated by the red region), employees with potential to improve (represented by the pink region), individuals who perform well (denoted by the blue region), but it cannot predict high performers (depicted by the green region). Finally, assessing the model's performance on the test set, the graph depicted in Fig. 13 illustrates the results of the model using MLR in relation to PC1 and PC2 of the test data set.

The classifier has achieved an accuracy of 87.3%, which indicates that it is capable of accurately identifying the performance level of employees.

However, It has been observed that the model is unable to predict high-performing employees (class 4), despite its satisfactory accuracy in predicting the other classes. It is worth mentioning that the time taken to find the nearest neighbors is also relatively short, indicating good efficiency.

5) K-nearest neighbors algorithm: Utilizing the K-nearest neighbors (KNN) algorithm as the third model for employee performance classification, the compressed database with PCA was employed. KNN is an intuitive and widely used machine learning algorithm for classification and regression tasks [35]. It is particularly useful when the decision boundaries are non-linear, as is the case with the data. The KNN algorithm is non-parametric and relies on the concept of similarity to make predictions. One of its advantages is its simplicity and ease of implementation. It does not make any assumptions about the underlying data distribution, making it applicable to a wide range of problems. KNN can also handle both numerical and categorical data.

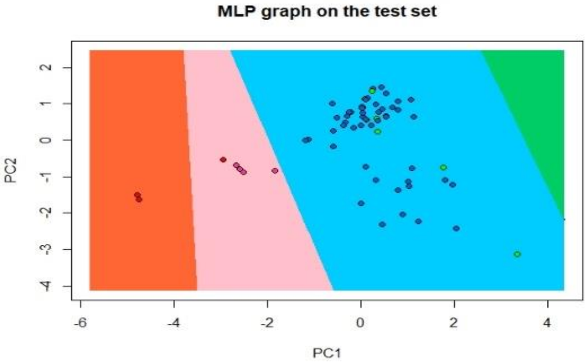


Fig. 13. Multinomial logistic regression results on the test set.

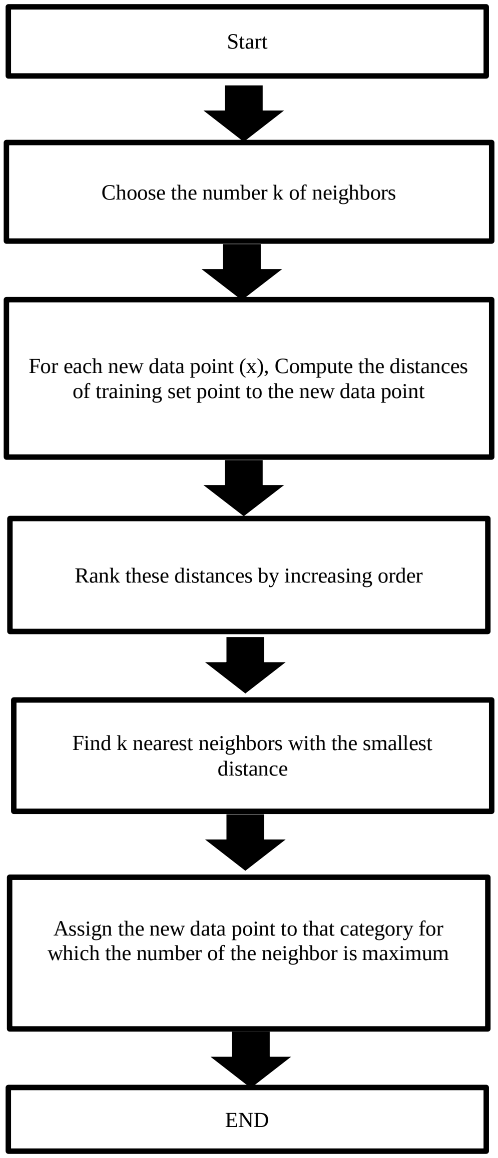
In KNN, the training data consists of labeled instances with known classes or values. When a new instance is to be classified or predicted, the algorithm looks for the k closest instances in the training set based on a distance metric, typically Euclidean distance. The predicted class or value of the new instance is determined by majority voting (for classification) or averaging (for regression) the labels or values of its k nearest neighbors. This means that the category Y of the new data is assigned by calculating the distance to each point in the training set and assigning it to the majority class of the k nearest neighboring data. The only parameter to be fixed is k, the number of neighbors to consider.

However, the KNN algorithm has some limitations. It can be computationally expensive, especially for large datasets, as it requires calculating distances between the new instance and all training instances. Furthermore, KNN is sensitive to the choice of k and the distance metric. Selecting an appropriate value for k and determining the most suitable distance metric can significantly impact the algorithm's performance. The steps of the algorithm are illustrated in Fig. 14.

The "class" library in R was used to generate a graph depicting the performance of the KNN model, based on the first two factors (PC1 and PC2) of the training dataset. The graph is shown in Fig. 15.

In order to test and validate the model, it was applied to the test database, resulting in an accuracy of 85.71%. The results are illustrated in Fig. 16.

Despite the high precision achieved, it is noted that the model is unable to predict the most efficient employees (class 4), although it performs well in predicting the other classes.



Start

Choose the number k of neighbors

For each new data point (x), Compute the distances of training set point to the new data point

Rank these distances by increasing order

Find k nearest neighbors with the smallest distance

Assign the new data point to that category for which the number of the neighbor is maximum

END

Fig. 14. KNN steps.

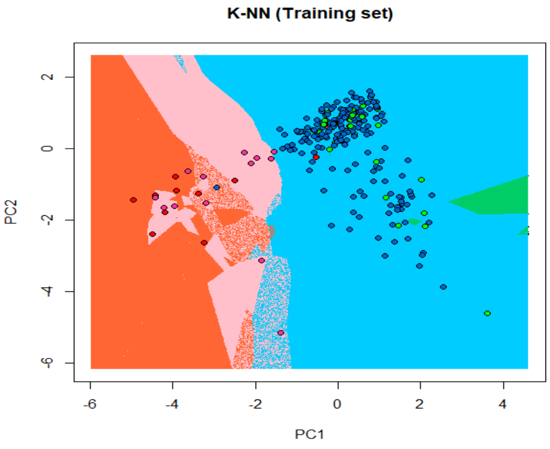


Fig. 15. KNN results on the training set.

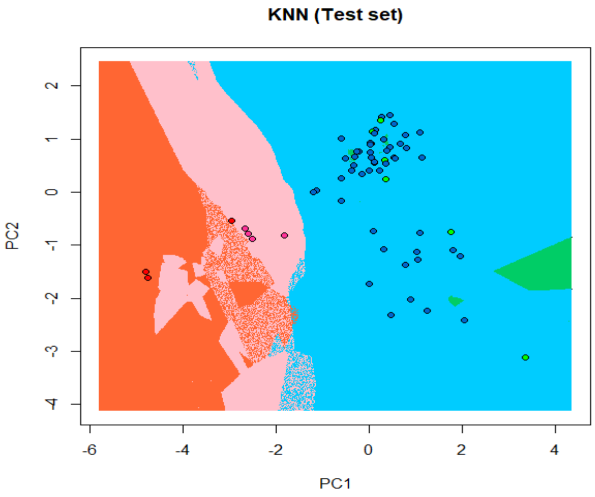


Fig. 16. KNN results on the test set.

Additionally, it is worth mentioning that the KNN algorithm has some limitations. Firstly, it can be computationally expensive, particularly for larger datasets, as it requires the calculation of distances between the new instance and all training instances. Secondly, the performance of KNN is highly dependent on the choice of k and the distance metric used. Selecting an appropriate value for k and determining the most suitable distance metric can significantly impact the algorithm's performance.

IV. Results and Discussion

Upon testing the three constructed models, it was observed that each model achieved varying levels of accuracy on the test set. The SVM, MLR, and KNN models achieved accuracy scores of 87%, 87.5%, and 85.7% respectively. The results indicated that the MLR model performed the best, with an accuracy score of 87.5%.

In the study, despite the challenges of acquiring HR data, a model was established to address this challenge. Firstly, the opportunities and limitations of applying machine learning techniques in the field of human resource management, particularly performance management, were identified. This allowed the research efforts to focus on leveraging the opportunities and mitigating the associated risks.

Secondly, a literature review was conducted to identify the most relevant factors and classify them into fields of performance evaluation based on related works.

Thirdly, after identifying the most relevant performance factors and selecting the most appropriate database, the PCA algorithm was used to compress the selected factors into two factors - the "behavior factor" with 41.42% of the variance and the "achievement factor" with 35.97%.

Finally, three classification algorithms (SVM, MLR, and KNN) were applied separately to the compressed database, and their accuracy was tested. Based on the results, it was determined that the MLR model achieved the highest accuracy score of 87.5%.

Therefore, the conclusion is drawn that the second model, which uses MLR as a classifier, is the most efficient model to adopt for the objective classification of employees based on the most relevant performance factors.

V. Conclusions

In today's world, it is becoming increasingly challenging for human resources managers to objectively and quantifiably classify employees based on relevant performance factors, especially in a highly competitive environment where data continues to grow exponentially, making the task more difficult. However, it is possible to classify the performance of human resources based on data, including salary, commitment, productivity (reflected by the number of projects carried out), absence, and the number of days in arrears.

The model enables the prediction of the performance class, allowing for the identification of low performers (red region), employees with potential to improve (pink region), individuals who perform well (blue region), and high performers who cannot be predicted by any of the three models (green region).

The quantified results obtained through the model provide human resources managers with a powerful classifier, enabling them to:

1) Identify talented employees with a good level of performance (blue region) and retain them.

2) Boost the performance of employees with an average score (pink region).

3) Plan for better organizational performance by understanding low performers (red region).

4) Achieve other economic benefits, such as optimizing time and recruitment results, which are no longer subject to subjective decisions. Additionally, the training strategy can be adapted to the specific needs of each performance region, thereby reducing the cost of training.

As a perspective of the research, the model needs to be tested on a larger database in a real industrial context, using questionnaires targeted around the chosen performance factors.

In summary, the results confirm that performance management can shift from being curative to predictive, and the model combining PCA and GRBF is a promising tool for predicting performance based on HR data and making data-based decisions for the three performance classes.

Acknowledgment

I would like to express my sincere gratitude to all those who have given me the opportunity to complete this paper. I would like to extend a special thanks to Pr. AKNIN Noura and Pr. CHRAYAH Mohamed, whose valuable contributions through stimulating suggestions and encouragement were instrumental in helping me finalize this work.

References

[1] J.Barney, Firm resources and sustained competitive advantage". Journal of Management, 17(1), 99–120, 1991.

[2] Elinor Friedman, Andrew Harley and Klayton Southwood. Insurance big data insurance big data can improve business, Towers Watson and Willis, 2006.

[3] Lee, T. W., Hom, P., Eberly, M., Li, J. J. (2018). Managing employee retention and turnover with 21st century ideas. Organizational Dynamics, 47, 88-98, 2018.

[4] Arthur, Bennett, Edens and Bell, Effectiveness of training in organizations: a meta-analysis of design and evaluation features, J Appl Psychol, 88(2):234-45, 2003.

[5] Marler and Boudreau, An evidence-based review of HR Analytics. The International Journal of Human Resource Management,· November 2016.

[6] Gartner glossary: https://www.gartner.com/en/human-resources/glossary/ hr-analytics.

[7] HR analytics in Business: Role, Opportunities, and Challenges of Using, http://dx.doi.org/10.37896/JXAT12.07/2441.

[8] Nocker, M.; Sena, V. Big Data and Human Resources Management: The Rise of Talent Analytics. Soc. Sci. 2019, 8, 273, 2019.

[9] Guenole, Nigel, Jonathan Ferrar, and Sheri Feinzig. 2017. The Power of People: Learn How Successful Organizations Use Workforce Analytics to Improve Business Performance. New York: Pearson Education. Available online: https://www.thepowerofpeople.org (accessed on 21 April 2018.

[10] OrgVue. 2019. Making People Count: 2019 Report on Workforce Analytics. London: OrgVue. Pease, Gene, Boyce Byerly, and Jac Fitz-enz. 2012. Human Capital Analytics: How to Harness the Potential of Your Organization's Greatest Asset. New York: Wiley.

[11] Nocker, M.; Sena, V. Big Data and Human Resources Management: The Rise of Talent Analytics. Soc. Sci. 8, 273, 2019.

[12] Pestieau Pierre, Gathon Henry-Jean. La performance des entreprises publiques. Une question de propriété ou de concurrence ? Revue économique. Volume 47, n°6, pp. 1225-1238, 1996.

[13] Zineb Issor : La performance de l'entreprise : un concept complexe aux multiples dimensions », Projectics / Proyéctica / Projectique 17(2):93, January 2017.

[14] Notat NN., "Une question centrale", Acteurs de l'Économie, dossier spécial performance, p. 72, octobre 2007.

[15] Boxall, P., Purcell, J., & Wright, P. 2007. Human Resource Management: Scope, analysis and significance. In P. Boxall, J. Purcell, & P. Wright (Eds.), Oxford Handbook of Human Resource Management: 364-381. Oxford: Oxford University Press.

[16] Joshua S. Bendickson et Timothy D. Operational performance: The mediator between human capital developmental programs and financial performance. Journal of Business Research. Volume 94, 2019.

[17] Deepu Kumar, B. H. Suresh. Workforce Diversity and its Impact on Employee Performance, International Journal of Management Studies V(4(1)):48, October 2018.

[18] Joyce Chua, Abdul Basit and Zubair Hassan. Leadership Style and Its Impact on Employee Performance, April 2018.

[19] Anitha Jagannathan. Determinants of employee engagement and their impact on employee performance, International Journal of Productivity and Performance Management 63(3):308-323, April 2014.

[20] Kholilah Kholilah, Yukke Sartika Sari. The impact of employee satisfaction as a mediator of compensation and career development on employee performance, May 2021.

[21] Idris Gautama So, Noerlina, A.A Djunggara and Athapol Ruangkanjanases. Effect of organisational communication and culture on employee motivation and its impact on employee performance, June 2018.

[22] H. Iwamoto, M. Takahashi, A quantitative approach to human capital management, Proc.-Soc. Behav. Sci. 172 112–119, 2015.

[23] L. Abdullah, S. Jaafar, I. Taib, Ranking of human capital indicators using analytic hierarchy process, Proc.-Soc. Behav. Sci. 107 22–28 (2013).

[24] C.F. Chen, L.F. Chen, Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry, Expert Syst. Appl. 34 (1) 280–290,2008.

[25] QA Al-Radaideh, E Al Nagi, Using data mining techniques to build a classification model for predicting employees performance, International

Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 144-151,2012.

[26] J.M. Kirimi, C.A. Moturi, Application of data mining classification in employee performance prediction, Int. J. Comput. Appl. 146 (2016).

[27] Christoph Schröer, Felix Kruse, Jorge Marx Gómez. A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science, Volume 181, 2021, Pages 526-534 ,2021.

[28] M.Hiri, M.Chrayah,N. Ourdani and N. Aknin, "Machine Learning Techniques for Diabetes Classification: A Comparative Study", International Journal of Advanced Computer Science and Applications(IJACSA), Volume 14 Issue 9, 2023.

[29] Philippeau, G. Comment Interpréter les Résultants d'une Analyse en Composantes Principales. Cited 61 times. Paris: Institut Techniques des Céréales et Fourrages , 1986.

[30] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, 2000.

[31] H.T. Lin, C.J. Lin, A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods, Technical report, Department of Computer Science, National Taiwan University, 2003.

[32] Tolles & Meurer,"Logistic Regression: Relating Patient Characteristics to Outcomes", JAMA The Journal of the American Medical Association, August 2016.

[33] Böhning, D, "Multinomial logistic regression algorithm", Annals of the Institute of Statistical Mathematics, pp. 197–200, 1992.

[34] Krishnapuram, B.; Carin, L, Figueiredo, M.A.T; Hartemink, A.J, "Sparse multinomial logistic regression: fast algorithms and generalization bounds, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 27, Issue: 6, June 2005.

[35] Youqiang Zhang, Guo Cao, Bisheng Wang, Xuesong Li,A novel ensemble method for k-nearest neighbor,Pattern Recognition,Volume 85,Pages 13-25, 2019.