Prediction model of financial suppliers in the vehicle manufacturing industry in Pereira.

Modelo de predicción de proveedores financieros en la industria manufacturera de vehículos en Pereira

D. C. Romero –Cárdenas ;A. O. Mejia ; L. A. Serna Cardona
DOI: https://doi.org/10.22517/23447214.25692
Scientific and technological research paper

***Abstract*— This research work presents the development of a model using data mining techniques to identify financial variables in a manufacturing company of automotive vehicle bodies in Pereira. The study is structured into four key phases. The first phase focuses on data preprocessing, including characterization, normalization, and dimensionality reduction using PCA, Relief, and Correlation. The second phase applies unsupervised learning with K-means and Gaussian Mixture Models (GMM) to cluster and validate data based on a defined target variable. In the third phase, supervised classifiers such as Bayesian Classifier, Artificial Neural Networks, Support Vector Machines, and KNN are employed to predict supplier efficiency, optimizing investment and costing processes. Finally, the fourth phase integrates preprocessing and prediction into a practical form, using libraries such as Plotly and Dash for detailed visualizations, and tools like GitHub and Heroku for application development. This study highlights the importance of artificial intelligence in business decision-making, demonstrating how data science techniques and visualization tools can facilitate the interpretation and utilization of data analysis results.**

***Index Terms* — Data Mining, Dash Plotly, Machine Learning, Neural Networks, Supplier Classification.**

***Resumen*— Este trabajo de investigación presenta el desarrollo de un modelo utilizando técnicas de minería de datos para identificar variables financieras en una empresa manufacturera de carrocerías para vehículos automotores en Pereira. El estudio se estructura en cuatro fases claves. La primera fase se centra en el preprocesamiento de datos, incluyendo la caracterización, normalización, y reducción de dimensionalidad mediante PCA, Relief y Correlación. La segunda fase aplica aprendizaje no supervisado con K-means y Mezclas Gaussianas (GMM) para agrupar y validar datos según una variable objetivo definida. En la tercera fase, se emplean clasificadores supervisados como el Clasificador Bayesiano, Redes Neuronales Artificiales, Máquinas de Soporte Vectorial y KNN para predecir la eficiencia de los proveedores, optimizando los procesos de inversión y costeo. Finalmente, la cuarta fase integra el preprocesamiento y la predicción en un formulario práctico, utilizando librerías como Plotly y Dash para visualizaciones detalladas, y herramientas como GitHub y Heroku para el desarrollo de la aplicación. Este estudio destaca la importancia de la inteligencia artificial en la toma de decisiones empresariales, demostrando cómo las técnicas de ciencia de datos y las herramientas de visualización pueden facilitar la interpretación y el aprovechamiento de los resultados del análisis de datos.**

***Palabras claves*— Clasificación de proveedores, Dash Plotly, Machine Learning, Minería de datos, Redes Neuronales.**

I. I. INTRODUCTION

In today's digital era, organizations face a constant challenge: it is no longer enough to store large volumes of data—they must also be able to manage it intelligently and convert it into actionable knowledge. This capability has become essential to maintaining competitiveness and improving operational efficiency in an increasingly demanding business environment [1]. In this context, data science has emerged as a key discipline, offering advanced tools and techniques to analyze large-scale information, identify patterns, predict behaviors, optimize processes, and discover new strategic opportunities [2].

This study stems from a specific need identified in a vehicle body manufacturing company located in Pereira, Colombia. Despite its growth in the national market and its International projection, the company lacks advanced mechanisms to objectively and automatically evaluate the efficiency of its financial suppliers.

This article was submitted for review on October 10, 2024, accepted on March 26, and published on March 31, 2025. This work was supported by CIAF University under the project entitled “Prediction Model of Financial Suppliers in the Vehicle Manufacturing Industry in Pereira.” The study was conducted by Diana Carolina Romero Cárdenas and Alejandro Ospina Mejía in collaboration with Master's student Luis Ariosto Serna Cardona. For correspondence: karol272@gmail.com, alejandro880411@gmail.com, moog22002@hotmail.com

Although it uses platforms such as Siesa, Power BI, and proprietary tools, its current analysis heavily depends on Excel macros and manual processes, limiting data integration, reducing analytical quality, and delaying strategic decision-making [3].

To address this issue, we propose the design of a predictive model based on data mining and machine learning techniques, capable of classifying suppliers into two categories: type 1 (efficient) and type 2 (less efficient). This model is built upon key information such as product identifiers, unit of measure, inventory type, batch type (domestic or imported), product line and subline, supplier and buyer codes, purchased quantities, unit values, local tax rates, and country of origin in the case of imports. Proper identification and analysis of these variables is essential to optimizing investments, reducing operational costs, and strengthening procurement management [4].

This leads to the following research question: Is it possible to design a predictive model, based on data mining and machine learning techniques, that can classify the financial efficiency of suppliers in a manufacturing company? To address this question, a four-phase methodology is proposed: (i) data preprocessing and dimensionality reduction using PCA, correlation analysis, and Relief-F [2][3][4]; (ii) clustering using unsupervised algorithms such as K-Means [5] and Gaussian Mixture Models (GMM)[6][12]; (iii) classification using supervised algorithms such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Bayesian Classifiers[13][14][15]; and (iv) development of an interactive application, designed with Dash and Plotly, that enables real-time visualization of results to support decision-making [16].

The proposed approach is supported by existing scientific literature. Studies such as those by Ralambondrainy (1995) [5] and Müller & Guido (2016) [13] have demonstrated the effectiveness of hybrid models—combining clustering and classification techniques—for segmenting economic actors and predicting performance. More recently, Zhang et al. (2021) [17] applied hybrid models for supplier selection in the manufacturing industry, while Kim and Lee (2022)[18] used clustering algorithms to classify suppliers in complex industrial environments. However, these models have been developed and validated in contexts that differ significantly from the Colombian setting, presenting a valuable opportunity to adapt them to emerging economies such as those in Latin America.

This paper is structured as follows: Section II presents the methodology; Section III reports the experimental results; Section IV discusses the findings in light of previous research; and Section V concludes the study and outlines future research directions.

II. II. METHODOLOGY

1. Data Preprocessing

The first phase focused on how to preprocess 350,732 sample data points taken between 2018 and 2023, segmented into 23 columns, which has become a problem for the vehicle manufacturing company when optimizing the investment and costing process. Therefore, this section indicated the ideal way to classify the information, then preprocess it and normalize it to 20 characteristic variables.

This section focused on the recognition of categorical data, identifying variables using conventional algorithms from different methods [7].

a. Collection and Normalization

As part of the preprocessing, it was necessary to normalize the database before applying the prediction algorithms. This was done to organize the data into logical groups, placing all variables on the same scale, thus allowing fair comparisons and minimizing data dispersion [2].
For this research and based on the state of the art supporting the current document, the normalization method used was StandardScaler, which generates normalized data to be processed with unsupervised and supervised learning algorithms table I [8].

TABLE I.
SAMPLE OF DATA USED IN THE EXPERIMENT

Documento	*Ítem*	*Razón Social*	*Docto. orden*	*Docto. solicitud*	*Desc. ítem*
15921	9186	403	69	21763	1223
15922	9332	403	69	21763	4197
15926	5330	89	182	21763	2802

b. Dimensionality Reduction

Once the database was obtained, the first step was to perform categorical encoding of the information by classifying the data into nominal, ordinal, and numerical variables [9]. During data, preprocessing, variable selection was conducted with the goal of identifying or determining which methods or techniques to apply, as conventional machine learning algorithms do not handle categorical variables efficiently. Therefore, it is necessary to convert them to quantitative data [10].

2. Unsupervised Learning

A clustering approach was developed using unsupervised learning with K-Means and Gaussian Mixture Models, in which performance metrics were used to validate the algorithm's

accuracy level according to the target variable recommended by the expert [11].

Based on the state of the art reviewed in this research, the K-Means algorithm was used with ordinal variables because using nominal variables would considerably increase computational cost. Therefore, ordinal variables were used to divide the dataset into K distinct clusters, aiming to obtain two groups: efficient and less efficient suppliers [5].

A. K-Means Algorithm

The following details how the K-Means algorithm operates:

Initialization:
Select k initial points u1, u2, ..., uk called centroids.

Cluster Assignment:
For each data point xi, assign it to the cluster Cj whose centroid uj is closest (1).

$$C\_j = \{ x\_i : || x\_i - \mu\_j ||\text{^}2 \leq || x\_i - \mu\_l ||\text{^}2 , \forall l, , 1 \leq l \leq k \} \quad (1)$$

Centroid Update:
Recalculate the centroid of each cluster as the mean of the points assigned to the cluster (2).

$$\mu_j = \frac{1}{\{|C_j|\} \sum_{\{x_i \in C_j\}} x_i} \quad (2)$$

Convergence Criterion:

Continue iterating between steps 2 and 3 until the centroids do not change significantly, i.e., until (3)

$$\left|\left|\mu_j^{t+1} - \mu_j^t\right|\right|^2 < \varepsilon \quad (3)$$

Where t is the iteration number and ϵ\epsilonϵ is a small positive value defining the tolerance for convergence.

a. B. Gaussian Mixture Model (GMM)

After generating the labels with K-Means using the 20 features and in agreement with the expert, it was decided to use the labels generated by this model as the Gold standard to verify the performance of another unsupervised algorithm, the Gaussian Mixture Model (GMM). GMMs are a probabilistic approach to represent the presence of subpopulations within an unlabeled dataset. Unlike the K-Means algorithm, which assigns each data point to a single cluster, GMMs consider that data points can belong to multiple clusters with different probabilities [12].

The following details how the Gaussian Mixture Model operates(4):

A Gaussian mixture is defined by:

1. K: The number of Gaussian components.

2. π_k**:** The weight of the k-th Gaussian component where

$$\sum_{k=1}^{K} \pi_k = 1 \quad (4)$$

3. u_k The mean vector of the k-th component.
4. $\sum k$: The covariance matrix of the k-th component.
The Gaussian mixture model is defined as:

$$p(x) = \sum_{k=1}^{K} \pi_k \, N\,(x \,|\mu_k, \sum\nolimits_k) \quad (5)$$

Where $N(x \mid u_k, \sum_k)$ is the probability density function of the multivariate normal distribution(5).

3. Supervised Learning

In this phase, supervised learning algorithms were applied through classification tasks to validate their performance. This research utilized the Bayesian Classifier, the K-Nearest Neighbors (KNN) algorithm, Artificial Neural Networks (ANN), and Support Vector Machines (SVM). It should be noted that for supervised algorithms, classification is a very important approach for recognizing categorical data. However, there are methods for this purpose that focus on Kernels, which have accuracy problems and high computational costs [13]. For this reason, an approach to identifying ordinal categorical variables was proposed using conventional classifiers such as Bayesian Classification, ANN, KNN, and SVM, which not only improved accuracy levels but also offered lower computational costs.
Subsequently, the performance of the different proposed methods was evaluated, identifying that the three standard classifiers significantly improved the accuracy of the database without any procedure, making this applicability an ideal way to group and classify categorical data [14].

The following describes the supervised KNN algorithm, the most accurate one applied in this research.

The K-Nearest Neighbors algorithm is a supervised classification method that assigns a label to a new instance based on the labels of its k nearest neighbors in the feature space [15]. The learning process of the KNN method is based on calculating the distance of the new element to each existing one and sorting these distances from smallest to largest to select the group to which it belongs. This group is the one with the highest frequency with the smallest distances [16].

The method operates as follows:

1. The training data X = x1, x2, … xN with labels y = y1; y2, … yN (N being the number of data samples) are stored in memory.
2. For a new sample xi e RD, where D is the number of attributes, the k nearest neighbors are found using a distance d in the entire training set (k can be 1, 3, 5, 7,…).
3. A procedure is carried out to select the class of the new sample xi.The common distances d are (6):

4. The common distances d are:

- **Mahalanobis Distance:**

$$D_{M(x,y)} = \sqrt{\{(x - y)^T \Sigma^{\{-1\}(x-y)}\}}$$

$$donde \sum^{-1} es\ la\ matriz\ de\ covarianza\ entre\ x\ y\ y.$$

- **Euclidean Distance:**

$$||x - y||_2 = \sqrt{\{(x - y)^T (x - y)\}}$$

- **Manhattan Distance:**

$$(Mathb(x,y) = \left| (x - y)^{\top(x-y)} \right| \quad (6)$$

III. III. Dataset and Experimental Setup

A. Dataset

The dataset used in this study was obtained from a vehicle manufacturing company in the city of Pereira. Table II presents the relevant information. It is important to clarify that the information was anonymized to maintain the confidentiality of the processed data. The data used in this project consists of 350,732 sample records collected between the years 2018 and 2023.

TABLE II.
SAMPLE OF THE ORIGINAL DATA USED FOR THIS EXPERIMENT

Tipo docto	Documento	Docto. referencia	Item	Razón Social	Sucursal
EA	EA-00187457	FACT-120731	9186	ALEJANDRO S.A.S	1
EA	EA-00000949	FACT-21	71630	PRIETO LUIS	1
EA	EA-00187463	FACT-41941	5330	CAROLINA COLOMBIA	1

A comma-separated flat file containing this information was used for subsequent processing in Python using the Spyder development environment. The data underwent preprocessing, dimensionality reduction, and feature extraction relevant to the experiment. Following this, unsupervised and supervised algorithms were applied, measuring their levels of accuracy and precision to determine the most accurate algorithm. This allows the company to predict the types of suppliers it has, classified as Rating 1 (Efficient Suppliers) and Rating 2 (Less Efficient Suppliers).

A. B. Specifications and Training

The identification and selection of the most relevant data sources on supplier purchases in the financial area of a manufacturing company in Pereira were carried out using the ERP system Siesa. The information was exported to a CSV file for analysis. The Extraction, Transformation, and Loading (ETL) process ensured the proper formatting of the data, including data cleaning.

Handling of missing and duplicate values and correction of descriptive errors.
A supplier classification module was developed, validating state-of-the-art methods such as k-means and Gaussian mixtures to generate grouping labels, which were evaluated using performance metrics like accuracy, precision, recall, and F1-Score. Additionally, a user interface with a form was designed to enter relevant data, demonstrating the applicability of the results through the use of the supervised KNN algorithm and Dash Plotly, with **70%** of the data for training and **30%** for validation. The combination of these methods and tools facilitated effective classification and clear visualization of the data, providing valuable information for financial decision-making.

Future work will focus on advanced predictive analysis and machine learning techniques to optimize the identification of critical financial variables and improve operational efficiency. Deep neural networks, real-time data processing, advanced interactive dashboards, and the expansion of the model to other business areas will be explored. A feedback loop system is also proposed for continuous improvement, explainable artificial intelligence techniques to ensure transparency, and fostering interdisciplinary collaborations, thus enhancing financial

management and establishing a solid foundation for the application of AI and data analysis in the organization.

IV. RESULTS AND DISCUSSIONS

A. Dimensionality Reduction Algorithms

a. PCA Algorithm

The covariance matrix was used, utilizing the Eigenvalues and Eigenvectors of the function, reducing the dimensionality of the variables by activating the PCA algorithm, as shown in Fig. 1. Once the process was completed, the variance method was used, which allowed determining which variables provided the most information to the algorithm.

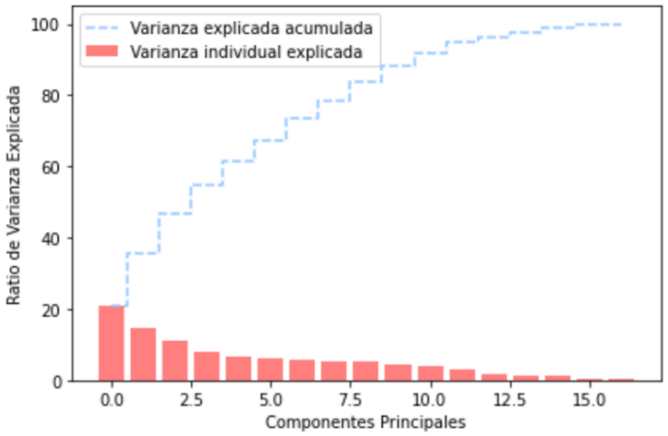

Fig. 1. Explained and Cumulative Variance

b. Correlation

As part of the dimensionality reduction process and with the objective of identifying the linear relationship and proportionality between statistical variables, correlation analysis was performed on the information obtained after applying principal component analysis in the previous step [3]. In Fig. 2, the 18 features are observed with respect to each other, where, for example, the total correlation between column 15 and column 1 can be seen, as well as between column 13 and column 3. The aim is to reduce dimensionality by removing the correlation that exists between the features.

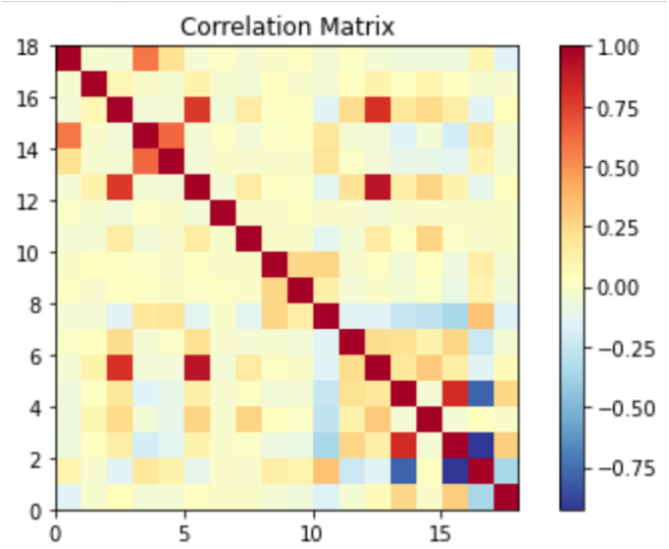

Fig. 2 Matrix with Full Correlation

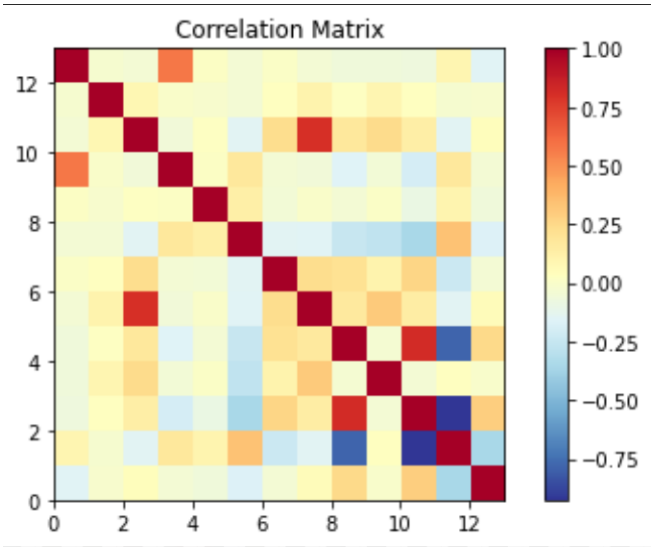

Fig. 3 Reduced Correlation Matrix with PCA

In this matrix, by eliminating 5 features, it becomes evident that the tone is lighter, indicating less correlation between the columns Fig. 3.

c. Relief F Algorithm

For the development of the Relief algorithm, the most recommended approach in the state of the art was used. In this section, the Relief algorithm estimated the relevance of the features based on how neighboring instances of different classes differ, setting the weight of each feature to 0, defining the number of nearest neighbors K and the number of iterations m. After executing all iterations, the features were selected based on their highest weights as the most relevant features [17].

In Fig. 4, the Eigenvalues are shown to determine which values are below the required threshold according to state-of-the-art recommendations. In this context, it is determined that the values below the threshold or negative values in the Eigenvalues are features that do not contribute value to the function.

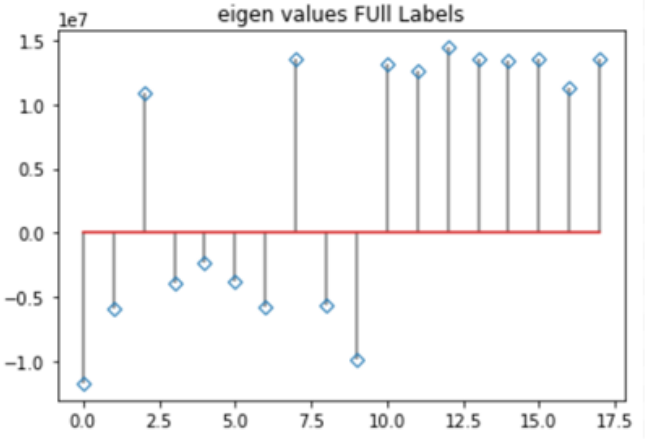

Fig. 4. Labels with Eigenvalues

After preprocessing the threshold and removing the less important features, Fig. 5 visualizes the most influential features in the database.

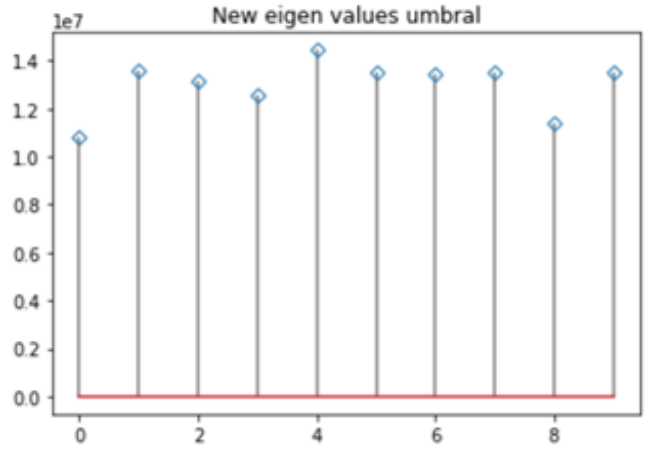

Fig. 5. Threshold with New Eigenvalues

For data preprocessing and dimensionality reduction, three techniques were used: PCA, Correlation, and Relief F, as recommended by the company expert. The results showed a reduction to 12 features with PCA, 13 features with Correlation, and 9 features with Relief F. Consequently, it was decided not to use any of the three techniques for dimensionality reduction and to use the complete database with all 20 features, as recommended by the expert.

B. Unsupervised Algorithms

a. K-Means Algorithm vs. Gaussian Mixture Model (GMM)

Based on the state of the art reviewed in this research, the K-Means algorithm was used with ordinal variables because using nominal variables would considerably increase computational costs. Therefore, ordinal variables were used to divide the dataset into K distinct clusters, aiming to obtain two groups: efficient suppliers and less efficient suppliers [5].

The following presents the prediction data obtained by applying GMM to the test dataset with 20 features using the labels generated by K-Means. It is important to highlight that the labels generated by the GMM algorithm were duly reviewed by the company table III.

TABLE III.
PERFORMANCE OF THE GMM ALGORITHM WITH 20 FEATURES

Métricas	Valores
Matriz de Confusión	
TP (Prov. Clase 0)	43.110
FN (Prov. Clase 0)	0
FP (Prov. Clase 1)	10.582
TN (Prov. Clase 1)	297.041
Exactitud	97%
Clase 0	
Precisión	0.80
Recall	1.00
F1-Score	0.89
Clase 1	
Precisión	1.00
Recall	0.97
F1-Score	0.98

Based on this, an accuracy of **97%** was identified, indicating that it is a high-performing model. Subsequently, performance metrics were analyzed by class, revealing that class 0 had a **precision of 80%** and a **recall of 100%,** while class 1 achieved a precision of 100% and a recall of 97%. This led to the conclusion that both classes demonstrated satisfactory performance, with an **F1-score of 89%** for class 0 and **98%** for class 1.

In test 2, the prediction data obtained by applying the GMM algorithm to the test dataset with the 13 features reduced by PCA using the labels generated by K-Means are presented. It is important to highlight that the labels generated by the GMM algorithm were duly reviewed by the company table IV.

TABLE IV.
PERFORMANCE OF THE GMM ALGORITHM WITH 13 FEATURES USING PCA

Métricas	Valores
Matriz de Confusión	
TP (Prov. Clase 0)	43.125
FN (Prov. Clase 0)	0
FP (Prov. Clase 1)	14.303
TN (Prov. Clase 1)	293.305
Exactitud	96%
Clase 0	
Precisión	0.75
Recall	1.00
F1-Score	0.86
Clase 1	
Precisión	1.00
Recall	0.95
F1-Score	0.98

Based on this, an accuracy of **96%** was identified, indicating that it is a high-performing model. Subsequently, performance metrics were analyzed by class, revealing that class 0 had a **precision of 75%** and a recall of 100%, while class 1 achieved a **precision of 100%** and a **recall of 95%.** This led to the conclusion that both classes demonstrated satisfactory performance, with an **F1-score of 86%** for class 0 and **98% for class 1.**

In test 3, the prediction data obtained by applying the GMM algorithm to the test dataset with the 9 features reduced by Relief F using the labels generated by K-Means are presented. It is important to highlight that the labels generated by the GMM algorithm were duly reviewed by the company.

The following presents the prediction data obtained by applying the GMM algorithm to the test dataset with the 9 features reduced by Relief F using the labels generated by K-Means. It is important to highlight that the labels generated by the GMM algorithm were duly reviewed by the company table V.

TABLE V.
PERFORMANCE OF THE GMM ALGORITHM WITH 9 FEATURES USING RELIEF F

Métricas	Valores
Matriz de Confusión	
TP (Prov. Clase 0)	307.181
FN (Prov. Clase 0)	441
FP (Prov. Clase 1)	33
TN (Prov. Clase 1)	43.078
Exactitud	100%
Clase 0	
Precisión	1.00
Recall	1.00
F1-Score	1.00
Clase 1	
Precisión	0.99
Recall	1.00
F1-Score	0.99

Within this framework, it was observed that the experiment of K-Means with Gaussian Mixtures, using Relief F, also achieved an accuracy of 100%. However, considering all the features, the algorithm achieved a high accuracy rate of 97%. In light of these results, the expert ultimately decided to retain all the labels generated by K-Means, recognizing their relevance in the context of the intended financial analysis.

C. Supervised Algorithms

Subsequently, the performance of SVM, ANN, Bayesian Classifier, and KNN was evaluated, where it was identified that all four standard classifiers significantly improved the accuracy of the database without any preprocessing, making this approach ideal for grouping and classifying categorical data [14].

a. SVM Algorithm

Below are the prediction results obtained by applying each supervised algorithm.

TABLE VI.
PERFORMANCE OF THE SVM CLASSIFICATION MODEL

Métricas	Valores
Matriz de Confusión	
TP (Prov. Clase 0)	8.834
FN (Prov. Clase 0)	5.301
FP (Prov. Clase 1)	26.699
TN (Prov. Clase 1)	29.313
Exactitud	54%
Clase 0	
Precisión	0.25
Recall	0.62
F1-Score	0.36
Clase 1	
Precisión	0.85
Recall	0.52
F1-Score	0.65

The confusion matrix in Table VI shows that out of 14.135 instances of class 0, 8.834 instances were correctly predicted and 5.301 were misclassified. For class 1, out of 56.012 instances, 29.313 were correctly predicted and 26.699 were misclassified, with an **accuracy of 54%.**

b. Bayesian Classification Algorithm

TABLE VII.
PERFORMANCE OF THE BAYESIAN CLASSIFICATION MODEL

Métricas	Valores
Matriz de Confusión	
TP (Prov. Clase 0)	9.270
FN (Prov. Clase 0)	4.865
FP (Prov. Clase 1)	27.503
TN (Prov. Clase 1)	28.509
Exactitud	82%
Clase 0	
Precisión	1.00
Recall	1.00
F1-Score	1.00
Clase 1	
Precisión	0.03
Recall	1.00
F1-Score	0.06

The confusion matrix in Table VII shows that out of 14.135 instances of class 0, 9.270 were correctly predicted and 4.865 were misclassified. For class 1, out of 56.012 instances, 28.509 were correctly predicted and 27.503 were misclassified, with an **accuracy** of **82%.**

c. Algoritmo ANN

TABLE VIII.
ANN MODEL PERFORMANCE

Métricas	Valores
Matriz de Confusión	
TP (Prov. Clase 0)	3.848
FN (Prov. Clase 0)	10.287
FP (Prov. Clase 1)	1.422
TN (Prov. Clase 1)	54.590
Exactitud	83%
Clase 0	
Precisión	0.73
Recall	0.27
F1-Score	0.40
Clase 1	
Precisión	0.84
Recall	0.97
F1-Score	0.90

The confusion matrix in Table VIII shows that out of 14.135 instances of class 0, 3.848 were correctly predicted and 1.422 were misclassified. For class 1, out of 56.012 instances, 54.590 were correctly predicted and 1.422 were misclassified, with an **accuracy** of **83%.**

d. KNN Algorithm

TABLE IX.
KNN MODEL PERFORMANCE

Métricas	Valores
Matriz de Confusión	
TP (Prov. Clase 0)	8.197
FN (Prov. Clase 0)	5.938
FP (Prov. Clase 1)	3.301
TN (Prov. Clase 1)	52.711
Exactitud	87%
Clase 0	
Precisión	0.71
Recall	0.58
F1-Score	0.64
Clase 1	
Precisión	0.90
Recall	0.94
F1-Score	0.92

The confusion matrix in Table IX shows that out of 14.135 instances of class 0, 8.197 were correctly predicted, and 5.938 were misclassified. For class 1, out of 56.102 instances, 52.711 were correctly predicted and 3.301 were misclassified, achieving **87% accuracy.**

Based on this, a **87% precision** was achieved, indicating that it is a high-performance model. Subsequently, performance metrics by class were analyzed, finding that both **precision** and **recall** were **71%** and **90%** for both classes, and an **F1-score** of **64% and 92%** was observed for both class 0 and class 1. This concluded that the supervised KNN algorithm is the most accurate among those studied in this research.

The results obtained in this research demonstrate the high effectiveness of the KNN model, achieving an accuracy of 87% and an F1-Score of 92% for efficient suppliers. This validates the model's ability to accurately classify categorical data in a business context.

These findings are consistent with studies such as that of Zhang et al. [17], who applied hybrid models for supplier selection in the manufacturing industry, reporting high levels of precision. Similarly, Kim and Lee [18] showed that clustering algorithms applied prior to supervised learning significantly improved segmentation performance in complex industrial environments.

Unlike those studies, which were developed in advanced organizational contexts in Asia, this research adapts the models to a Colombian setting using data extracted from a local ERP system and tailored to the conditions of a real manufacturing company in Pereira. This contextual difference explains why the KNN model proved more effective than SVM or ANN in this case, particularly due to its lower computational cost and its compatibility with visual tools such as Dash and Plotly [16].

In conclusion, this study offers a methodological contribution by demonstrating that classification models like KNN, when supported by robust feature reduction and interactive visualization processes, can yield results comparable to international research and remain applicable to emerging business environments.

V. CONCLUSIONS AND FUTURE WORK

This work presented an approach for characterizing and grouping categorical data using ordinal variables. The variable "Línea," considered key by the company expert, was used for dimensionality reduction and supplier classification. Preprocessing techniques such as Z-score scaling were applied, improving experimental results and providing structured data for supervised and unsupervised algorithms.

The clustering approach included K-Means to generate labels, showing great similarity with the target variable. The labels were compared with the Gaussian Mixture Models (GMM) algorithm, achieving **97% accuracy** with 20 features, **96%** with 13 features (PCA), and **100%** with 9 features. However, the expert preferred using the 20 features from K-Means due to their business relevance. Finally, an optimized KNN model with various kernels and features was constructed, successfully identifying and classifying suppliers with high accuracy. This model is useful for analysts and investors in financial decision-making for the vehicle manufacturing company in Pereira, demonstrating the effectiveness of the K-Means and GMM algorithms in improving data separability and reducing computational times.

Future work will focus on advanced predictive analysis and Machine Learning techniques to optimize the identification of critical financial variables and improve operational efficiency. Deep neural networks, real-time data processing, advanced interactive dashboards, and the expansion of the model to other business areas will be explored. A feedback loop system for continuous improvement, explainable artificial intelligence techniques to ensure transparency, and interdisciplinary collaborations are also proposed to enhance financial management and establish a solid foundation for the application of AI and data analysis in the organization.

The performance of both supervised and unsupervised algorithms in this study is inherently conditioned by the characteristics of the dataset, which originates from a single manufacturing company in Pereira. As such, the predictive models may not generalize directly to other organizations without conducting similar analyses to determine the most suitable algorithms for each specific context. While the current approach effectively classifies supplier types within the studied company, its applicability to other environments requires contextual adaptation and possible retraining of the models.

IV. VI. ACKNOWLEDGMENTS

V.

Thanks to the Master's program in Systems and Computing Engineering at the Universidad Tecnológica de Pereira, to the Director of the Institutional Corporation for Administration and Finance (CIAF), Luis Ariosto Serna, and to Ph.D. Julián David Echeverry.

VI. VII. REFERENCES

[1] Gonzalez Disla, R. R. (2013). Big data: El cambio en los paradigmas de la información. ResearchGate. https://www.researchgate.net/profile/Renato-GonzalezDisla/publication/311950584_BIG_DATA_El_Cambio_en_el_Paradigma_de_la_Informacion/links/58645e6208ae329d6203a9d5/BIG-DATA-El-Cambio-en-el-Paradigma-de-laInformacion.pdf

[2] Sablón, B., et al. (2019). Gestión de la información y toma de decisiones en organizaciones educativas. Revista de Ciencias Sociales, XXV(2), 120–130. https://doi.org/10.31876/rcs.v25i2.27341

[3] Amo Cubillo, A. (n.d.). Research data management. Universidad de Valladolid. Recuperado el 25 de agosto de 2024, de http://uvadoc.uva.es/handle/10324/31269

[4] Chien, C.-F., Chang, Y.-J., & Wang, W.-C. (2018). AI and big data analytics for wafer fab energy saving and chiller optimization to empower intelligent manufacturing. IEEE Xplore. https://ieeexplore.ieee.org/document/8374411

[5] Ralambondrainy, H. (1995). A conceptual version of the K-means algorithm. Pattern Recognition Letters, 16(11), 1147–1157. https://doi.org/10.1016/0167-8655(95)00075-R
[6] Liang, Y., Quan, D., Wang, F., Jia, X., Li, M., & Li, T. (2020). Financial big data analysis and early warning platform: A case study. IEEE Access, 8, 36515–36526. https://doi.org/10.1109/ACCESS.2020.2969039
[7] Romero, A. C., Sanabria, J. S. G., & Cuervo, M. C. (n.d.). Utilidad y funcionamiento de las bases de datos NoSQL. Facultad de Ingeniería, 21(33), 21–32. Recuperado el 25 de agosto de 2024, de https://www.redalyc.org/articulo.oa?id=413940772003
[8] Iturria Aguinaga, A. (n.d.). Reduction of false positives in online outlier detection over time series using ensemble learning. Recuperado el 25 de agosto de 2024, de https://hdl.handle.net/10481/82540
[9] Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining, 2(1), 1–10. https://doi.org/10.30880/jscdm.2021.02.01.003
[10] Macqueen, J. (n.d.). Some methods for classification and analysis of multivariate observations. Recuperado el 25 de agosto de 2024, de https://doi.org/
[11] Karatzoglou, A., Meyer, D., Wien, W., & Hornik, K. (n.d.). Journal of Statistical Software: Support vector machines in R. Recuperado el 25 de agosto de 2024, de http://www.jstatsoft.org/
[12] Rasmussen, C. E. (2004). Gaussian processes in machine learning. Lecture Notes in Computer Science, 3176, 63–71. https://doi.org/10.1007/978-3-540-28650-9_4
[13] Castrillón, O. D., Sarache, W., & Ruiz-Herrera, S. (2020). Prediction of academic performance using artificial intelligence techniques. Formación Universitaria, 13(1), 93– 102. https://doi.org/10.4067/S0718-50062020000100093
[14] Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In Proceedings of the ninth international workshop on Machine learning (pp. 249-256). Morgan Kaufmann.
[15] Ahmad, A., & Dey, L. (2007). A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. Pattern Recognition Letters, 28(1), 110–118. https://doi.org/10.1016/J.PATREC.2006.06.006
[16] Moya, R. (n.d.). Selección del número óptimo de clusters. Recuperado el 25 de agosto de 2024, de https://jarroba.com/seleccion-del-numero-optimo-clusters/
[17] Fin de Máster, T., Tutor, Q., Pérez, J., Cotutor, Á. A., & Bernabeu, P. (n.d.). Métodos de clasificación con Python: Aplicaciones empresariales. Universitat Politècnica de València, Escuela Politécnica Superior de Alcoy. Recuperado el 25 de agosto de 2024, de https://riunet.upv.es/handle/10251/195236

Diana Carolina Romero Cárdenas: graduated in Systems Engineering in 2006 at the National Open and Distance University. His research interests include Data Science, Data Analysis, Machine Learning, and Big Data.
ORCID: https://orcid.org/0009-0006-0568-874X

Alejandro Ospina Mejía: received his degree in Systems Engineering in 2011. His research interests include Data Science, Data Analysis, Machine Learning, and Big Data. ORCID: https://orcid.org/0009-0009-6902-8910.

Luis Ariosto Serna Cardona received his undergraduate degree in physical engineering (2017) and his M.Sc. degree in engineering (2021) and PsD Student. He is director of research at CIAF educacation superior and researcher in the department of engineering. Research interests: machine learning and deep learning. ORCID: https://orcid.org/0000-0003-3985-4014