Assessing Musculoskeletal Disorder Susceptibility in Professional Drivers Using K-Means Algorithms

Imane Benallou*, Abdellah Azmani, Monir Azmani
Intelligent Automation and BioMedGenomics Laboratory (IABL), Abdelmalek Essaadi University, FST of Tangier, Km 10 Ziaten B.P: 416 Tangier 90000, Morocco
E-mail: ibenallou@uae.ac.ma, abdellah.azmani@gmail.com, monir.azmani@gmail.com
*Corresponding author

Keywords: driver segmentation, musculoskeletal disorders, machine learning, k-means, clustering

Received: November 10, 2024

The work of professional drivers is crucial in many economic sectors. Truck and bus drivers, taxi drivers, and delivery vehicle drivers are at the heart of the action, transporting goods and people to keep businesses running and ensure they reach their daily destinations. Behind this essential activity, significant challenges arise from working conditions and their impact on health. Because of that, drivers are exposed to different risk factors, possibly contributing to the onset of musculoskeletal disorders (MSDs), such as low back pain and other symptoms. Various factors were identified, including exposure to car vibrations, long hours of sitting while driving, repetitive manual activities, psychosocial factors, and individual characteristics, which contribute to the development of these problems in these professionals. This paper proposes a driver profiling model using the K-means clustering algorithm to establish risk profiles associated with the occurrence of MSDs. The model involves integrating personal and professional variables to identify the most vulnerable. The model estimates suggest that only 21% of drivers are at low risk of developing MSDs, highlighting the high prevalence of these disorders within this occupation.

Povzetek: Ocenjevanje dovzetnosti poklicnih voznikov za mišično-skeletne motnje z uporabo nenadzorovanega učenja. Na podlagi ergonomskih, demografskih in delovnih dejavnikov je s pomočjo algoritma K-means izvedena razvrstitev voznikov v skupine tveganja, kar omogoča zgodnje prepoznavanje ogroženih posameznikov in podporo preventivnim ukrepom na delovnem mestu.

1 Introduction

Professional driving involves operating a vehicle for business and personal purposes over an extended period [1]. Professional drivers experience harsh conditions that could make them vulnerable to MSDs [2]. Bus drivers are the most affected by low back pain, with a rate of 59%, compared to car drivers (26%) and truck drivers (16%) [3]. These disorders result from an overload of the musculoskeletal system, often caused by repetitive movements, awkward postures, and excessive and prolonged use of force in the workplace. Prolonged sitting, vehicle ergonomics, and vibrations are major risk factors [4].

In addition, the risk of MSDs increases with age and career years [5]. Long-term exposure to vibration throughout the body [6], especially on uneven roads, amplifies seat wobbles when accelerating [7]. Another important risk factor is the length of the journey and the lack of regular stops, which increases muscle tension [8]. Additionally, a high body mass index (BMI) and a lack of regular exercise contribute more to the development of musculoskeletal disorders [9].

The consequences of MSDs have a significant impact on both an individual and social level, resulting in a variety of costs. In European companies, more than half of workers affected by MSDs report absence from work, and these employees are generally absent for more extended periods than those with other health problems. In addition, MSDs are the leading cause of permanent disability in 60% of reported cases [10]. Therefore, preventing MSDs remains a fundamental concern for all actors: organizations, researchers, and practitioners. This article presents an innovative model for segmenting drivers based on their risk profile for developing musculoskeletal disorders (MSDs). This model is based on integrating key variables such as age, work experience, body mass index (BMI), weekly hours worked, physical exercise, and other factors related to working conditions. Grouping drivers according to their level of risk for MSDs will provide a structured and practical approach to identifying the most vulnerable groups.

The body of this article is structured as follows: section 2 presents the related work. Section 3 explains the methodology used to build the proposed model; the results obtained are also presented in this section. Section 4 interprets the results obtained and discusses their role in improving the well-being of professional drivers. Finally, the conclusion is given in section 5.

2 Literature review

Much research has been explored on the study of MSDs, and this review touches on three primary levels: prevention, diagnosis, and rehabilitation of MSDs.

Regarding the prevention of MSDs, several researchers have proposed combining wearable sensors with machine learning algorithms to mitigate ergonomic risks associated with work-related musculoskeletal disorders. Matos et al. [11] proposed a system that monitors workers with textile machinery to detect movements with a high risk of MSDs. This system includes three modules: a Motion Capture System, a Time Series Forecasting Integrating Machine Learning algorithms (SVM, XG, MLP, and deep LSTM), and a risk detection module based on rules for work-related musculoskeletal disorders. Su et al. [12] investigated the application of decision trees for assessing ergonomic risks associated with musculoskeletal disorders among sewing machine operators. The developed model highlighted the existing relationship between body segments and the possible risk patterns. Zhao et al. [13] have developed a portable inertial measurement unit detection system to identify the risks of musculoskeletal disorders (MSDs) in construction workers. This system is based on convolutional LSTM to recognize the awkward postures of workers in daily tasks.

In diagnosing MSDs, significant research has integrated artificial intelligence algorithms, particularly deep learning, into the detection of various musculoskeletal pathologies. Cohen et al. [14] investigated the use of deep neural networks to detect wrist fractures on X-rays. The results showed that the performance of the proposed model in diagnosing wrist fractures in radiology is significantly higher than that of non-radiologists. Hess et al. [15] proposed using a deep learning network in 3D diagnostics to detect rotator cuff tears. This model enables automatic slice-by-slice segmentation of the humerus, scapula, and rotator cuff muscles. This approach could replace manual segmentation, which is often cumbersome and tedious. Georgeanu et al. [16] have developed a model that utilizes convolutional neural networks to detect bone tumor malignancy from MRI scans. This tool operates without the need for manual segmentation by a specialist, thereby increasing the reliability of the diagnosis provided by the orthopedist.

In musculoskeletal rehabilitation, much research has explored the use of machine learning models to predict the outcomes of rehabilitation programs. Zmudzki and Smeets [17] explored the use of machine learning (ML) in enhancing the personalization and effectiveness of interdisciplinary and multimodal treatments for patients with chronic musculoskeletal pain. The indicators studied encompass classic dimensions of rehabilitation, covering biomedical, psychosocial, and functional dimensions. Thirteen machine learning algorithms were trained and combined to develop a reliable patient stratification model. Obukhov et al. [18] proposed a model for monitoring musculoskeletal rehabilitation exercises. This model analyzes and classifies user movements to enhance the accuracy of human movement recognition during musculoskeletal rehabilitation exercises. The multilayered neural network algorithms, KNN and Random Forest, have given good results, reinforcing the potential of this tracking system.

Regarding the application of machine learning algorithms in studying musculoskeletal disorders in drivers. Balakrishnan et al.[19] proposed a machine learning classifier model for monitoring driving positions that are considered the leading cause of musculoskeletal disorders in sedentary workers. Aliabadi et al. [20] employed two machine learning algorithms: linear regression, and Random Forest, to investigate musculoskeletal discomforts in mining truck drivers. The results demonstrated increased accuracy for the Random Forest method compared to linear regression. These results also highlighted the role of uncomfortable body posture, vibrations, and age in the onset of musculoskeletal discomfort.

Hanumegowda and Gnanasekaran [26] developed machine learning models to predict work-related musculoskeletal disorders in bus drivers. Three algorithms: decision tree, random forest, and naïve Bayes were used in the study and trained on data extracted from a structured questionnaire based on the Modified Nordic Musculoskeletal Questionnaire (MNMQ), supplemented by direct observations. These variables encompass three broad domains: sociodemographic, occupational, and behavioral/health, to predict the frequency of MSD-related pain over the past 12 months.

Raza et al. [27] investigated the prevalence of musculoskeletal disorders (MSDs) in two occupational groups, truck drivers and office workers, using three machine learning algorithms: decision tree, random forest, and naïve Bayes. The results indicated an increased vulnerability of truck drivers to MSDs compared to office workers. The study emphasized the importance of establishing suitable ergonomic conditions in the workplace to minimize the risk of musculoskeletal disorders.

According to the extensive literature review of previous studies, research gaps were found. One of the most critical limitations noted was the absence of objective clinical validation of the reported musculoskeletal disorders, which restricts the validity of the collected data and, ultimately, the external validity of the results. Furthermore, in most studies, some potentially decisive variables, such as the ergonomics of the seating position, road surface quality, or effort intensity, were not taken into account. The exclusion of these variables will limit the explanatory value of the developed models and may lead to an underestimation or overestimation of the risks involved.

To this end, our research will partly fill these gaps by taking a clustering algorithms-based approach to identify risk driver profiles in terms of several key variables. This methodological option enables an unconstrained segmentation of the driver population as well as an exploration of the underlying vulnerability patterns to MSDs. Applying these variables, the purpose of our work will be to present an enhanced understanding of the risk factors by combining individual, behavioral, and contextual variables. This option may contribute to the

development of specific preventive measures and a more efficient ergonomic design policy in the transport sector.

3 Implementation

In this study, the k-means algorithm was employed, a clustering algorithm that falls under the category of unsupervised machine learning, which operates on unlabeled data [21]. The primary disadvantage of this method is the need to accurately predict the number of clusters to perform the grouping operations effectively. Several statistical techniques, including the Silhouette and Elbow methods, are employed [22]. Figure 1 illustrates the different stages of its implementation. To assess the robustness of the clustering approach, three alternative algorithms were also applied for comparison purposes: hierarchical clustering, DBSCAN, and Gaussian Mixture Models (GMM). The results of this comparison are discussed in the results analysis section.

3.1 Methodology

To predict the onset of musculoskeletal disorders in professional drivers, the procedure illustrated in Figure 2 was followed.

The proposed methodology is divided into four essential steps. The first step involved collecting data through interviews with 277 professional drivers of various vehicle types, including taxis, trucks, buses, and service vehicles. The data gathered covered a wide range of demographic, physiological, and occupational variables.

Drivers were categorized by age into four groups: under 35, 35 to 44, 45 to 55, and over 55 years. Estimates of height and weight were used to calculate the Body Mass Index (BMI), which was classified as underweight or normal (BMI < 25), overweight (25 ≤ BMI < 30), or obese (BMI ≥ 30). Driving experience was recorded and grouped into three categories: less than 5 years, between 5 and 20 years, and over 20 years. Weekly driving time was also collected and categorized into three ranges: less than 40 hours, between 41 and 60 hours, and more than 60 hours. Additionally, the frequency of heavy lifting was assessed on a daily scale, ranging from 0 to 8, reflecting the frequency with which drivers were required to handle physically demanding loads. Physical activity was evaluated by asking participants to report the number of hours per week they engage in physical effort, also on a scale from 0 to 8. The study also collected information on vehicle ergonomics, specifically the presence of adjustable seats and steering wheels, as well as the availability of armrests. Finally, the condition of the roads and the nature of the journeys regularly undertaken by the drivers were recorded to provide insight into their work environment.

Table 1 provides a detailed description of the columns in the dataset. Then, the data was cleaned to detect and correct any inaccurate, incomplete, or erroneous entries. To enable the application of clustering algorithms, corresponding numerical values were generated using Python code for the variables age, BMI, weekly driving time, and driving experience, based on the collected data. For other non-numeric values, the representation shown in Table 2 was adopted. An extract of the resulting dataset is represented in Table 3.

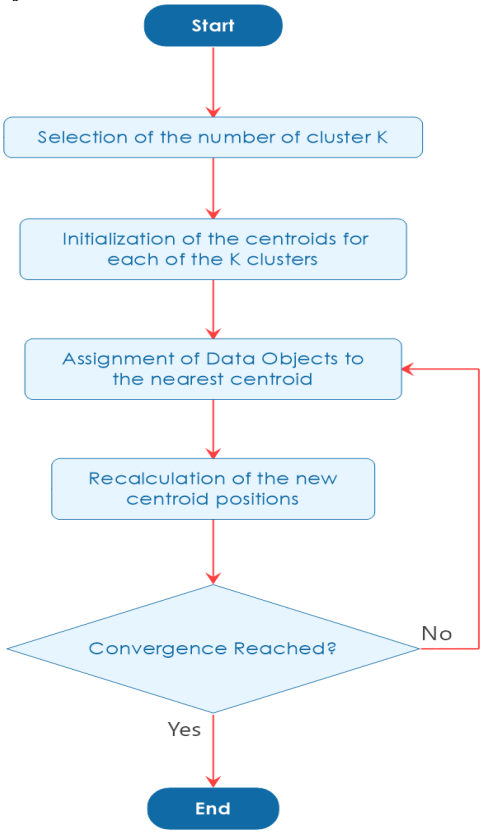

Figure 1: K-means algorithm [23]

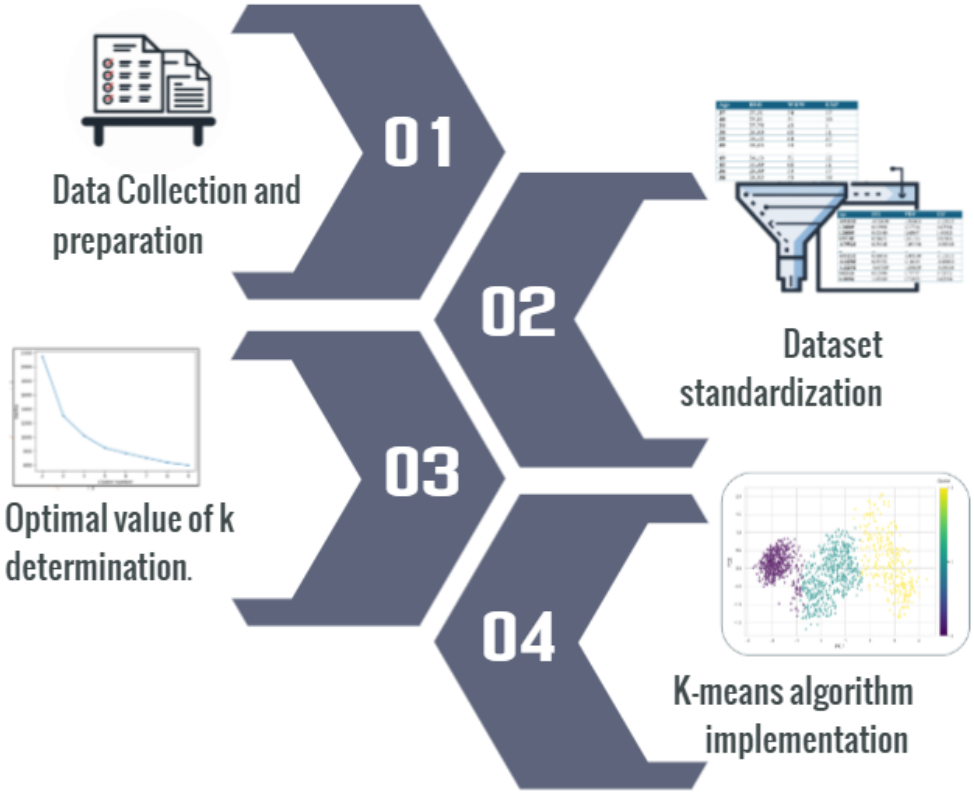

Figure 2: Driver segmentation process

Table 1: Description of the dataset

Column	Description
Age	This variable refers to the driver's age.
BMI	It specifies the body mass index.
WHW	It indicates the weekly driving time.
EXP	This variable reflects the driver's total years of driving experience.
FLHO	This column contains the frequency of heavy lifting per day (from 0 to 8).
WEH	Number of hours of physical activity per week (0 to 8).
ST	It indicates whether the seat is adjustable in height, depth or tilt.
STR	It specifies whether the steering wheel is adjustable or not.
ARM	It specifies whether the armrests are present or not.
TRP	This column indicates the nature of the journeys: short, medium, or long.
RD	It indicates the condition of the roads (poor, average or good).

Table 2: Adopted representation of non-numerical criteria

Criteria	Representation
The available fit options for the vehicle seat (ST)	A score of 0 is assigned if the seat is not adjusted. On the other hand, when there are adjustment options, a score of 1 is given for each available option (height, tilt, depth). The final score is the sum of all points earned.
Available Adjustment Options for the Vehicle Steering Wheel (STR)	The score is one if there are adjustment options and zero if there are no adjustment options.
The presence of armrests (ARM)	The score awarded is 1 if the armrests are present and 0 if not.

The nature of the trips (TRP)	The scores awarded are as follows: 0 for long trips, 1 for medium trips, and 2 for short trips.
Road conditions (RD)	The scores awarded are as follows: 0 for roads in poor condition, 1 for roads in average condition, and 2 for roads in good condition.

Table 3: Extract from the generated dataset

Age	BMI	WHW	EXP	FLHO	WEH	ST	STR	ARM	TRP	RD
37	27,31	54	17	1	5	3	1	1	1	1
40	25,01	51	10	2	8	3	1	0	0	0
23	27,75	43	1	1	0	1	1	0	0	0
30	26,89	60	11	0	1	3	1	1	1	1
59	34,15	64	27	0	8	3	1	1	0	0
39	29,85	43	17	4	5	2	1	1	1	0
…		…	…	…	…	…	…	…	…	…
49	34,13	51	12	8	4	2	1	0	1	1
45	33,49	60	11	8	4	0	0	0	0	0
36	28,49	53	17	0	2	0	0	0	0	0
38	28,92	53	19	2	4	1	0	0	0	0

To simplify the resulting dataset, the columns ST, STR, ARM, TRP, and RD have been removed and replaced by a new column called "WCS", which contains a score reflecting the drivers' working conditions. This score was obtained by adding up different coefficients related to the vehicle's ergonomics, the length of the journeys, and the state of the roads. Table 4 shows an excerpt from the resulting dataset.

Table 4: Simplified dataset extract

Age	BMI	WHW	EXP	FLHO	WEH	WCS
37	27,31	54	17	1	5	7
40	25,01	51	10	2	8	4
23	27,75	43	1	1	0	2
30	26,89	60	11	0	1	7
59	34,15	64	27	0	8	5
39	29,85	43	17	4	5	5
…	…	…	…	…	…	…
49	34,13	51	12	8	4	5
45	33,49	60	11	8	4	0
36	28,49	53	17	0	2	0
38	28,92	53	19	2	4	1

The second step is to increase the dataset volume to 1385 rows. To expand the original dataset of 277 observations, an algorithmic approach was adopted to generate new entries while ensuring their consistency and realism. This method relied on the controlled reproduction of existing profiles using rules derived from the statistical distributions observed in the original data. Specifically, value ranges were defined for each variable based on their initial distribution, and additional instances were generated randomly within these intervals, while preserving the correlations identified among variables. For example, age groups, BMI categories, driving experience levels, and weekly driving hours were maintained, and plausible combinations were selected based on their observed frequencies. This process resulted in an enriched dataset of 1,385 synthetic individuals, whose overall characteristics preserved the statistical structure of the original dataset. Validation checks were then performed to ensure internal consistency and representativeness, in order to guarantee the reliability of the analyses conducted on the expanded dataset.

The dataset was then standardized; this is essential since the model is based on measuring distances between data,

and dimensions play a crucial role in its implementation. Table 5 provides an overview of the standardized data.

Table 5: Normalized dataset

Age	BMI	WHW	EXP	FLHO	WEH	WCS
-0.599713	-0.402375	0.032550	0.109777	-1.046203	1.255522	1.338201
-0.366148	-0.914460	-0.219196	-0.612325	0.636639	2.755444	-0.000644
-1.689681	-0.304411	-0.890519	-1.540742	-1.046203	-1.244348	-0.893207
-1.144697	-0.495887	0.536041	0.509168	-1.455766	-0.744374	1.338201
1.113095	1.120521	0.871703	1.141351	-1.455766	2.755444	0.445638
…	…	…	…	…	…	…
-0.444003	0.163144	-0.890519	0.109777	0.182488	1.255522	0.445638
0.334546	1.116068	-0.219196	-0.406010	1.820741	0.755548	0.445638
0.023126	0.973574	0.536041	-0.509168	1.820741	0.755548	-1.785770
-0.677568	-0.139654	-0.051366	0.109777	-1.455766	-0.244400	-1.785770
-0.521858	-0.043916	-0.051366	0.316092	-0.636639	0.755548	-1.339488

To implement the K-means algorithm, it is necessary to determine the optimal value of k. Figure 3 illustrates the result obtained by the Elbow method. According to the results obtained by applying the Elbow method, the optimal number of clusters is 3. Three clusters were obtained by applying the K-means algorithm; the number of drivers per cluster is presented in Table 6.

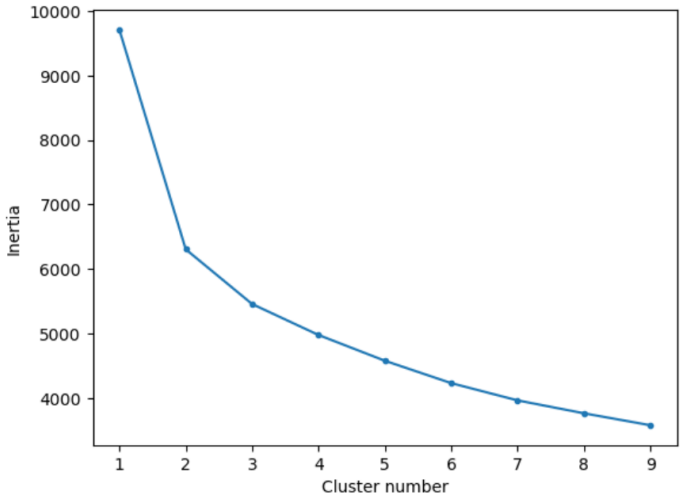

Figure 3: Elbow method

Table 6: Number of drivers per cluster

Cluster	Driver
0	527
1	575
2	285

To gain deeper insight into the structure of the resulting clusters, principal component analysis (PCA) was applied. Although PCA is a dimensionality reduction technique, in this case it was used solely for visualization purposes. By projecting the high-dimensional data into two principal components, we were able to plot the clusters on a 2D chart, allowing a more straightforward interpretation of their separation and internal organization. The visualization obtained by PCA is shown in Figure 4.

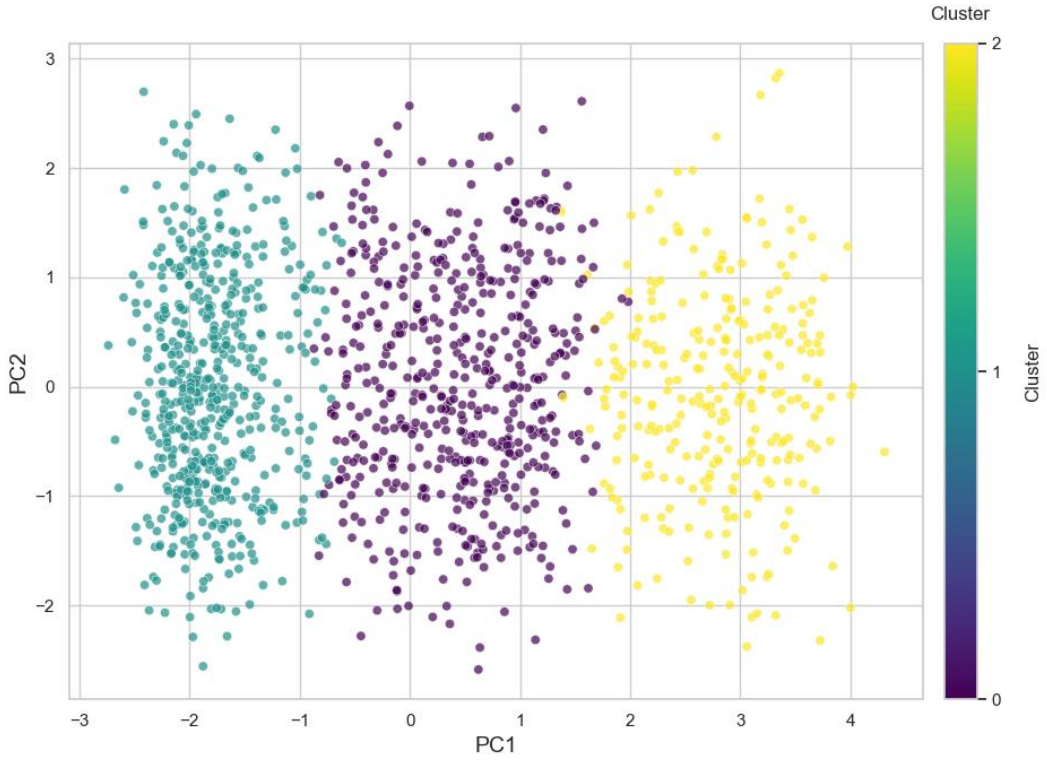

Figure 4 K-Means clusters visualization using PCA

3.2 Results analysis

The K-means model proved effective in performing efficient driver segmentation. To evaluate the quality of the clustering, two internal validation metrics were used: the Silhouette Score and the Calinski-Harabasz Index. The results yielded a Silhouette Score of 0.218 and a Calinski-Harabasz Index of 538.93. These values suggest that the clustering structure is reasonably coherent, with moderate cohesion and separation (as indicated by the Silhouette Score), and strong inter-cluster distinction (as reflected by the high Calinski-Harabasz Index). These findings support the robustness and validity of the clustering results obtained using the K-means algorithm. Each cluster represents a segmentation of drivers; the analysis and understanding of the characteristics of each cluster will lead to the implementation of preventive measures that are more precisely targeted at high-risk groups. To further illustrate the outcome of the clustering process, Figure 5 shows the size distribution of each cluster.

Of the three clusters, cluster 1 (C1) has the highest proportion of drivers at 41%, followed by cluster 0 (C0) with 38%, and cluster 2 (C2) with 21%.

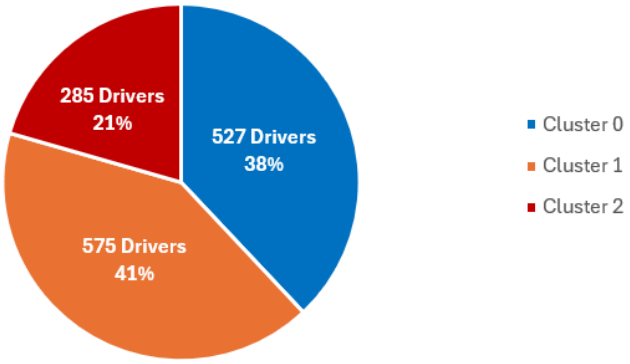

Figure 5: Clusters size

Table 7 shows the average values of each variable for each cluster. Age, BMI, WHW, and EXP appear to be the most influential factors in cluster formation, with significant differences between the averages of the clusters. In contrast, the FLHO, WEH, and WCS variables appear to have less influence on cluster formation, as their means are similar across clusters.

Table 7: Mean values of the characteristics of each cluster

Cluster	Age	BMI	WHW	EXP	FLHO	WEH	WCS
0	41.28	29.26	50.09	12.58	3.53	2.55	4.11
1	56.72	32.51	64.17	25.55	3.50	2.52	3.97
2	26.77	21.97	38.78	2.72	3.69	2.30	3.84

The distribution of the three clusters according to age, BMI, weekly driving time, and experience was represented in Figures 6, 7, 8, and 9. The x-axis of these figures represents the different clusters of drivers, while the y-axis represents the values of age, BMI, weekly driving time, and experience. Each rectangle has a height corresponding to the range of variation of the y values, and the line inside the rectangle indicates the average value.

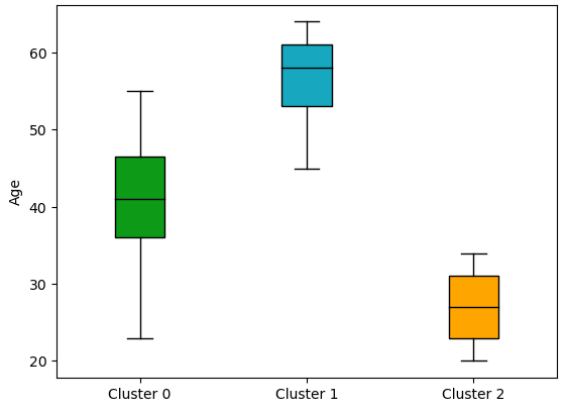

Figure 6: Clusters distribution according to age

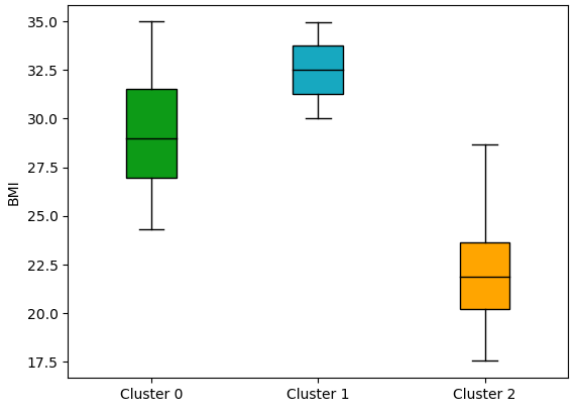

Figure 7: Clusters distribution according to BMI

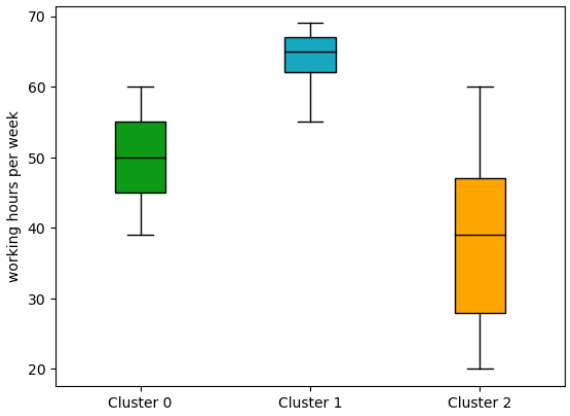

Figure 8: Clusters distribution according to WHW

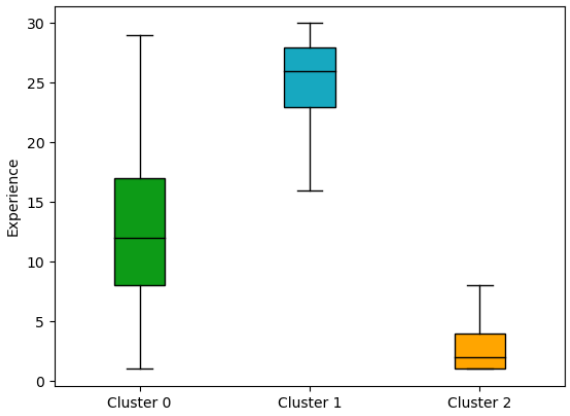

Figure 9: Clusters distribution according to experience

Cluster 0 comprises 38% of the total drivers and includes middle-aged drivers (approximately 41 years old) with a relatively high body mass index of 29.26. Their volume of work is average, at approximately 50 hours per week, while their professional experience averages about 12 years. The habit of carrying heavy loads is moderately practiced (up to 4 times a day), and the intensity of physical activities in this group remains low. This group has the best working conditions, with an average score of 4.11. Therefore, this cluster can be labeled as a group of drivers at average risk of developing MSDs.

Cluster 1 comprises 41% of the total drivers, consisting of older drivers with an average age of 56 years and a higher body mass index of 32.51. They work an average of 64.17 hours per week, which is the highest volume of work among the groups, and have an average work experience of approximately 25 years. Their frequency of heavy lifting is moderate, and their weekly physical activity is still low. The average working conditions score for this group is around 3.97, slightly lower than for cluster 0, indicating that the perception of working conditions is somewhat worse due to the high workload. It follows, therefore, that this group could be classified as drivers who are at high risk of developing MSDs.

Cluster 2 includes 21% of the total number of drivers. It comprises young drivers with an average age of about 27 years and a BMI of less than 21.97, indicating better overall physical health. Their weekly workload was the lowest of the three groups, averaging about 39 hours. They also have the least experience, with an average of only

about three years. Although their workload is lower than that of others, the frequency of heavy lifting is a little higher, and their weekly physical activity is the lowest. Finally, the working conditions score for this group is around 4, the weakest among the three groups. All these factors give the impression of less favourable working conditions. Therefore, based on this result, this cluster can be labeled as a group of drivers at low risk of developing MSDs.

To further evaluate the suitability of K-means, we compared its clustering results with those obtained using hierarchical clustering, DBSCAN and GMM. Table 8 provides comparison results based on internal validation metrics. Figures 10, 11 and 12 illustrate the clusters generated by each method using PCA-based visualizations.

Among the tested methods, K-means yielded the highest scores (Silhouette Score = 0.218, Calinski-Harabasz Index = 538.93), indicating relatively well-defined and compact clusters. GMM and hierarchical clustering showed slightly lower performance, with less distinct boundaries between clusters (GMM Silhouette Score = 0.163; Hierarchical Silhouette Score = 0.183). DBSCAN, while known for its robustness to noise and ability to detect clusters of arbitrary shape, did not yield usable results under the tested parameter settings, as it failed to produce a sufficient number of meaningful clusters without classifying a majority of the points as noise.

These results suggest that K-means is the most suitable method for this dataset, providing a balance between simplicity, interpretability, and clustering performance. The clusters obtained through K-means were subsequently analyzed to identify distinct driver profiles and to guide targeted preventive measures.

Table 9 presents the final dataset after the corresponding group label has been assigned to each driver.

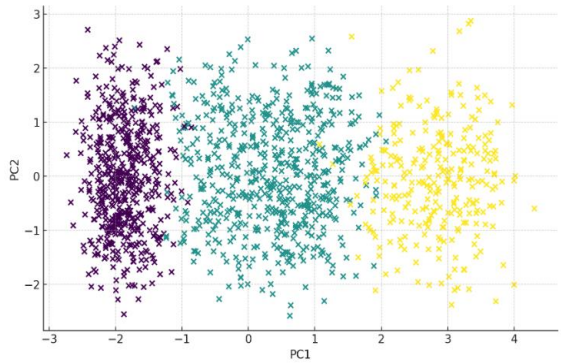

Figure 10: GMM clusters visualization using PCA

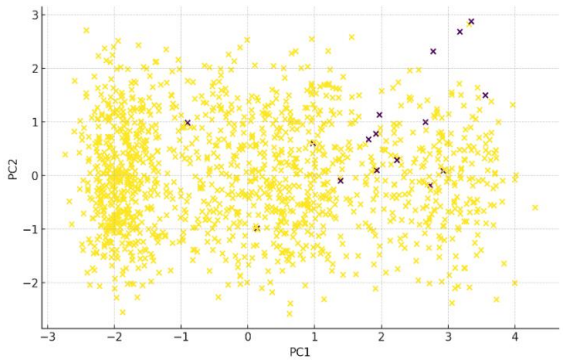

Figure 11: DBSCAN clusters visualization using PCA

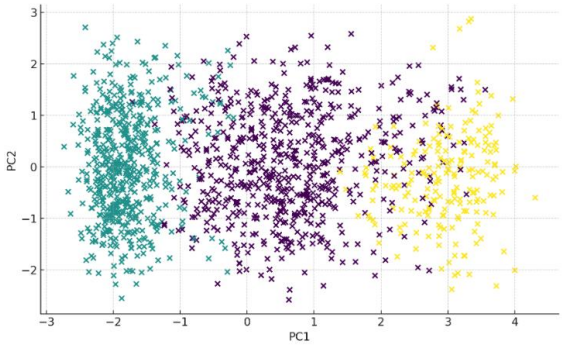

Figure 12: Hierarchical clusters visualization using PCA

Table 8: Comparison of clustering algorithms based on internal validation metrics

Algorithme	Parameters	Silhouette Score	Calinski-Harabasz Index
K-means	k = 3	0.218	538.93
Hierarchical clustering	n_clusters = 3	0.183	479.44
GMM	n_components = 3	0.163	440.86
DBSCAN	eps = 1.5, min_samples = 14	Non applicable	Non applicable

Table 9: Final dataset

Age	BMI	WHW	EXP	FLHO	WEH	WCS	Cluster
37	27,31	54	17	1	5	7	Medium risk
40	25,01	51	10	2	8	4	Medium risk
23	27,75	43	1	1	0	2	Low risk
30	26,89	60	11	0	1	7	Medium risk
59	34,15	64	27	0	8	5	High risk
39	29,85	43	17	4	5	5	Medium risk
…	…	…	…	…	…	…	…
49	34,13	51	12	8	4	5	Medium risk
45	33,49	60	11	8	4	0	Medium risk
36	28,49	53	17	0	2	0	Medium risk
38	28,92	53	19	2	4	1	Medium risk

4 Discussion

Musculoskeletal disorders (MSDs) are a major global health problem. They impact the quality of life and cause injuries among staff, leading to economic losses due to absenteeism and medical expenses [24].

This work used the K-means algorithm to group drivers into 3 clusters based on their susceptibility to developing MSDs. Although alternative clustering techniques were explored, K-means proved to offer the most interpretable and well-separated clusters for the dataset, supporting its use in the final driver segmentation analysis. An analysis of the clusters was conducted to identify the characteristics of the clusters obtained. Cluster 0 was considered to be the group of drivers at average risk of MSDs. MSD Cluster 1 is composed of drivers at high risk of MSDs, while members of Cluster 2 are at low risk of MSDs. These results estimate that only 21% of professional drivers have a low risk of developing MSDs, which confirms the relatively high prevalence of these disorders in this category of professionals. Therefore, it is imperative to impose preventive measures to preserve the proper functioning of these professionals' musculoskeletal systems. These measures may include periodic MSD screening programs and training programs on behaviors and exercises that effectively reduce MSDs [25].

5 Conclusion

This article highlights the application of machine learning algorithms to improve drivers' well-being and working conditions. This approach resulted in the development of a driver segmentation model based on their risk profile for musculoskeletal disorders (MSDs). By integrating key variables such as age, experience, body mass index (BMI), weekly work hours, and frequency of heavy lifting, this model provides an innovative approach to identifying the groups of drivers most at risk of MSDs. This model can be easily adapted to other datasets with similar characteristics, making it a flexible tool for various transportation companies.

Segmenting drivers into distinct risk groups provides decision-makers valuable information to design targeted interventions and improve working conditions. Indeed, this classification enables the prioritization of prevention actions and the optimization of resources by identifying the drivers most likely to benefit from training programs, ergonomic modifications, or workload reductions.

The results of this study pave the way for future research, including the use of reinforced learning for real-time recommendations. This system will enable the recommendation of adjustments to activity or posture based on real-time data from movement tracking devices and posture sensors. These recommendations would reduce risky behaviours even before a problem arises.

These research perspectives could lead to a deeper understanding of MSD risk factors among drivers and develop more targeted and effective prevention strategies, thus improving both the workers' health and the transport companies' productivity.

Ethical considerations

This study was conducted in accordance with ethical principles applicable to low-risk research. Participants were verbally informed about the study’s objectives, the nature of the data collected, and their right to decline to answer specific questions or withdraw at any time without consequence. Verbal informed consent was obtained prior to participation.

No personally identifiable information was collected. All responses were fully anonymized, processed in aggregate form, and stored under strict confidentiality.

The study did not involve any sensitive data as defined by Moroccan Law No. 09-08 on the protection of personal data. Given the non-intrusive and anonymous nature of the research, no formal ethical committee approval was required in accordance with current ethical guidelines.

Acknowledgment

This research is supported by the Ministry of Higher Education, Scientific Research and Innovation, the Digital Development Agency (DDA), and the National Center for Scientific and Technical Research (CNRST) of Morocco

(Smart DLSP Project - AL KHAWARIZMI IA-PROGRAM).

References

[1] S. B. M. Tamrin, K. Yokoyama, N. Aziz, and S. Maeda, « Association of Risk Factors with Musculoskeletal Disorders among Male Commercial Bus Drivers in Malaysia », *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol. 24, nº 4, p. 369-385, 2014, doi: 10.1002/hfm.20387.

[2] L. Montoro, S. Useche, F. Alonso, and B. Cendales, « Work Environment, Stress, and Driving Anger: A Structural Equation Model for Predicting Traffic Sanctions of Public Transport Drivers », *International Journal of Environmental Research and Public Health*, vol. 15, nº 3, Art. nº 3, mars 2018, doi: 10.3390/ijerph15030497.

[3] A. A. Rufa'i and *al.*, « Prevalence and Risk Factors for Low Back Pain Among Professional Drivers in Kano, Nigeria », *Archives of Environmental & Occupational Health*, vol. 70, nº 5, p. 251-255, sept. 2015, doi: 10.1080/19338244.2013.845139.

[4] O. Pickard, P. Burton, H. Yamada, B. Schram, E. F. D. Canetti, and R. Orr, « Musculoskeletal Disorders Associated with Occupational Driving », *Int J Environ Res Public Health*, vol. 19, nº 11, p. 6837, juin 2022, doi: 10.3390/ijerph19116837.

[5] S. A. Arslan, M. R. Hadian, G. Olyaei, S. Talebian, M. S. Yekaninejad, and M. A. Hussain, « Comparative effect of driving side on low back pain due to Repetitive Ipsilateral Rotation », *Pakistan Journal of Medical Sciences*, vol. 35, nº 4, p. 1018, août 2019, doi: 10.12669/pjms.35.4.488.

[6] H. Ayari, M. Thomas, and S. Doré, « A Design of Experiments for Statistically Predicting Risk of Adverse Health Effects on Drivers Exposed to Vertical Vibrations », *International Journal of Occupational Safety and Ergonomics*, vol. 17, nº 3, p. 221-232, janv. 2011, doi: 10.1080/10803548.2011.11076888.

[7] A. V. Araújo, G. S. Arcanjo, H. Fernandes, and G. S. Arcanjo, « Ergonomic work analysis: A case study of bus drivers in the private collective transportation sector », *Work*, vol. 60, nº 1, p. 41-47, janv. 2018, doi: 10.3233/WOR-182718.

[8] S. Senthanar and P. L. Bigelow, « Factors associated with musculoskeletal pain and discomfort among Canadian truck drivers: A cross-sectional study of worker perspectives », *Journal of Transport & Health*, vol. 11, p. 244-252, déc. 2018, doi: 10.1016/j.jth.2018.08.013.

[9] M. Grabara, « The association between physical activity and musculoskeletal disorders—a cross-sectional study of teachers », *PeerJ*, vol. 11, p. e14872, févr. 2023, doi: 10.7717/peerj.14872.

[10] R. Govaerts and *al.*, « Prevalence and incidence of work-related musculoskeletal disorders in secondary industries of 21st century Europe: a systematic review and meta-analysis », *BMC Musculoskeletal Disorders*, vol. 22, nº 1, p. 751, août 2021, doi: 10.1186/s12891-021-04615-9.

[11] L. M. Matos and *al.*, « Proactive prevention of work-related musculoskeletal disorders using a motion capture system and time series machine learning », *Engineering Applications of Artificial Intelligence*, vol. 138, p. 109353, déc. 2024, doi: 10.1016/j.engappai.2024.109353.

[12] J.-M. Su, J.-H. Chang, N. L. D. Indrayani, and C.-J. Wang, « Machine learning approach to determine the decision rules in ergonomic assessment of working posture in sewing machine operators », *Journal of Safety Research*, vol. 87, p. 15-26, déc. 2023, doi: 10.1016/j.jsr.2023.08.008.

[13] J. Zhao, E. Obonyo, and S. G. Bilén, « Wearable Inertial Measurement Unit Sensing System for Musculoskeletal Disorders Prevention in Construction », *Sensors*, vol. 21, nº 4, Art. nº 4, janv. 2021, doi: 10.3390/s21041324.

[14] M. Cohen and *al.*, « Artificial intelligence vs. radiologist: accuracy of wrist fracture detection on radiographs », *Eur Radiol*, vol. 33, nº 6, p. 3974-3983, juin 2023, doi: 10.1007/s00330-022-09349-3.

[15] H. Hess and *al.*, « Deep-Learning-Based Segmentation of the Shoulder from MRI with Inference Accuracy Prediction », *Diagnostics*, vol. 13, nº 10, Art. nº 10, janv. 2023, doi: 10.3390/diagnostics13101668.

[16] V. A. Georgeanu, M. Mămuleanu, S. Ghiea, and D. Selișteanu, « Malignant Bone Tumors Diagnosis Using Magnetic Resonance Imaging Based on Deep Learning Algorithms », *Medicina*, vol. 58, nº 5, Art. nº 5, mai 2022, doi: 10.3390/medicina58050636.

[17] F. Zmudzki et R. J. E. M. Smeets, « Machine learning clinical decision support for interdisciplinary multimodal chronic musculoskeletal pain treatment », Front. Pain Res., vol. 4, mai 2023, doi: 10.3389/fpain.2023.1177070.

[18] A. Obukhov and *al.*, « Examination of the Accuracy of Movement Tracking Systems for Monitoring Exercise for Musculoskeletal Rehabilitation », *Sensors*, vol. 23, nº 19, Art. nº 19, janv. 2023, doi: 10.3390/s23198058.

[19] S. A. Balakrishnan, E. F. Sundarsingh, V. S. Ramalingam, and A. N, « Conformal Microwave Sensor for Enhanced Driving Posture Monitoring and Thermal Comfort in Automotive Sector », *IEEE*

Journal of Electromagnetics, RF and Microwaves in Medicine and Biology, p. 1-8, 2024, doi: 10.1109/JERM.2024.3405185.

[20] M. Aliabadi, E. Darvishi, M. Farhadian, R. Rahmani, M. Shafiee Motlagh, and N. Mahdavi, « An investigation of musculoskeletal discomforts among mining truck drivers with respect to human vibration and awkward body posture using random forest algorithm », *Human Factors and Ergonomics In Manufacturing*, vol. 32, nº 6, p. 482-493, nov. 2022, doi: 10.1002/hfm.20965.

[21] M. E. Celebi and K. Aydin, Éd., *Unsupervised Learning Algorithms*. Cham: Springer International Publishing, 2016. doi: 10.1007/978-3-319-24211-8.

[22] T. Kodinariya and P. Makwana, « Review on Determining of Cluster in K-means Clustering », *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, p. 90-95, janv. 2013.

[23] A. A. Abdulnassar and L. R. Nair, « Performance analysis of Kmeans with modified initial centroid selection algorithms and developed Kmeans9+ model », *Measurement: Sensors*, vol. 25, p. 100666, févr. 2023, doi: 10.1016/j.measen.2023.100666.

[24] N. Hasheminejad, M. Amirmahani, and S. Tahernejad, « Biomechanical evaluation of midwifery tasks and its relationship with the prevalence of musculoskeletal disorders », *Heliyon*, vol. 9, nº 9, sept. 2023, doi: 10.1016/j.heliyon.2023.e19442.

[25] E. Rezaei, F. Shahmahmoudi, F. Makki, F. Salehinejad, H. Marzban, and Z. Zangiabadi, « Musculoskeletal disorders among taxi drivers: a systematic review and meta-analysis », *BMC Musculoskelet Disord*, vol. 25, nº 1, p. 663, août 2024, doi: 10.1186/s12891-024-07771-w.

[26] P. K. Hanumegowda et S. Gnanasekaran, « Prediction of Work-Related Risk Factors among Bus Drivers Using Machine Learning », International Journal of Environmental Research and Public Health, vol. 19, no 22, Art. no 22, janv. 2022, doi: 10.3390/ijerph192215179.

[27] M. Raza, R. K. Bhushan, A. A. Khan, A. M. Ali, A. Khamaj, et M. M. Alam, « Prevalence of Musculoskeletal Disorders in Heavy Vehicle Drivers and Office Workers: A Comparative Analysis Using a Machine Learning Approach », Healthcare, vol. 12, no 24, Art. no 24, janv. 2024, doi: 10.3390/healthcare12242560.