Data and Metadata. 2026; 5:1287
doi: 10.56294/dm20261287

ORIGINAL

Development of a new predictive hiring system with multi-model voting sets and advanced stacking techniques for assessing semantic soft skills

Desarrollo de un nuevo sistema predictivo de contratación con conjuntos de votación multimodales y técnicas avanzadas de apilamiento para evaluar las habilidades sociales semánticas

Asmaa Lamjid[1], Anass Ariss[1], Jamal Mabrouki[2], Karim El Bouchti[1], Soumia Ziti[1]

[1]Department of Computer Science, Intelligent Processing Systems & Security Team, Faculty of Sciences, Mohammed V University in Rabat. Morocco.
[2]Laboratory of Spectroscopy, Molecular Modelling, Materials, Nanomaterial, Water and Environment, CERNE2D, Mohammed V, University in Rabat, Faculty of Science, Rabat. Morocco.

Cite as: Lamjid A, Ariss A, Mabrouki J, El Bouchti K, Ziti S. Development of a new predictive hiring system with multi-model voting sets and advanced stacking techniques for assessing semantic soft skills. Data and Metadata. 2026; 5:1287. https://doi.org/10.56294/dm20261287

Submitted: 17-08-2025 **Revised:** 19-10-2025 **Accepted:** 03-12-2025 **Published:** 01-01-2026

Editor: Dr. Adrián Alejandro Vitón Castillo

Corresponding author: Asmaa Lamjid

ABSTRACT

Human resources face a major challenge in extracting and identifying the semantic correspondence of data, in particular the soft skills most recruiters seek from heterogeneous data. The complexity lies in identifying the relationships between textual descriptions in CVs, keywords, descriptions in professional networks, and relevant soft skills such as communication, persuasion skills, negotiation, relationship building, empathy, teamwork, conflict resolution, emotional intelligence, time management, work ethics, after analysis and research, we chose these soft skills as input data because they encompass all the soft skills that a recruiter might look for in a candidate for any position. The present study introduces a predictive hiring system that assesses candidate performance based on soft skills extracted from three main sources, namely resumes, professional social network profiles, and psychometric assessments. A dataset of over one million candidate records was processed. Data analysis relied on state-of-the-art NLP techniques, including word embeddings and contextual language models, in order to build a semantic database linking keywords, phrases, and descriptions to targeted soft skills. Machine and deep learning models were applied, followed by an ensemble approach integrating KNN, Decision Tree, and Random Forest. To overcome prediction accuracy and overfitting limitations, a meta-model XGBoost was developed, achieving superior results with an accuracy of 98 %. The results demonstrate that the proposed meta-model outperforms baseline approaches, delivering high predictive accuracy and robust generalization. These findings highlight the potential of combining semantic analysis with advanced machine learning to support more reliable and scalable predictive recruitment systems.

Keywords: Predictive Hiring, Soft Skills; Machine Learning; Embeddings Models; Semantic Data Base; AI in HR; Automation of the Recruitment Process; BERT.

RESUMEN

Los recursos humanos se enfrentan a un gran reto a la hora de extraer e identificar la correspondencia semántica de los datos, en particular las habilidades sociales que la mayoría de los reclutadores buscan en datos heterogéneos. La complejidad radica en identificar las relaciones entre las descripciones textuales de los CV, las palabras clave, las descripciones en las redes profesionales y las habilidades sociales relevantes,

como la comunicación, la persuasión, la negociación, la creación de relaciones, la empatía, el trabajo en equipo, la resolución de conflictos, la inteligencia emocional, la gestión del tiempo y la ética laboral. Tras el análisis y la investigación, elegimos estas habilidades sociales como datos de entrada porque abarcan todas las habilidades sociales que un reclutador podría buscar en un candidato para cualquier puesto. El presente estudio introduce un sistema de contratación predictivo que evalúa el rendimiento de los candidatos basándose en las habilidades sociales extraídas de tres fuentes principales, a saber, los currículums, los perfiles de las redes sociales profesionales y las evaluaciones psicométricas. Se procesó un conjunto de datos de más de un millón de registros de candidatos. El análisis de los datos se basó en técnicas de PLN de última generación, incluyendo incrustaciones de palabras y modelos de lenguaje contextual, con el fin de construir una base de datos semántica que vinculara palabras clave, frases y descripciones con las habilidades sociales específicas. Se aplicaron modelos de aprendizaje automático y profundo, seguidos de un enfoque conjunto que integraba KNN, árbol de decisión y bosque aleatorio. Para superar las limitaciones de precisión de la predicción y el sobreajuste, se desarrolló un metamodelo XGBoost, que logró resultados superiores con una precisión del 98 %. Los resultados demuestran que el metamodelo propuesto supera a los enfoques de referencia, ya que ofrece una alta precisión predictiva y una generalización sólida. Estos hallazgos ponen de relieve el potencial de combinar el análisis semántico con el aprendizaje automático avanzado para respaldar sistemas de reclutamiento predictivos más fiables y escalables.

Palabras clave: Contratación Predictiva ; Habilidades Sociales ; Aprendizaje Automático ; Modelos de Incrustación ; Base de Datos Semántica ; IA en RH ; Automatización del Proceso de Selección ; BERT.

INTRODUCTION

The demand for effective and innovative recruitment solutions has risen in recent years as organizations search for candidates who possess both the necessary technical expertise and soft skills that enhance team dynamics and contribute to overall success. Although automated recruitment systems have been implemented for several decades—initially through rule-based Applicant Tracking Systems (ATS), these earlier methods mainly relied on keyword matching, surface-level text analysis, and rigid scoring procedures. Consequently, they were unable to capture deeper semantic relationships within candidate data or to reliably infer soft-skill indicators from heterogeneous sources. Despite their long-standing use, such systems often failed to reduce bias or to replace the need for extensive manual assessments. These limitations emphasize the demand for an enhanced predictive hiring framework that is powered by AI to perform contextual semantic analysis, integrate multiple sources of information, and generate robust predictions of candidate on-job performance. This context necessitates the development of the enhanced predictive hiring system proposed in this study.

Furthermore, improving the performance of prediction models presents a significant technical challenge. Individual machine learning models have limitations regarding accuracy and generalization capacity, which can raise issues related to robustness and the risk of overfitting, especially in a large database with over a million profiles. To overcome these limitations and enhance prediction accuracy, we initially experimented with an ensemble learning approach based on voting. However, this method yielded less satisfactory results.

To address these shortcomings, a meta-model was developed that combines foundational algorithms like K-Nearest Neighbors (KNN), Decision Trees, and Random Forests with advanced meta-models such as XGBoost, Logistic Regression, and LightGBM. Among these, XGBoost emerged as the top performer, demonstrating exceptional capability in handling large-scale and dynamic datasets. Through rigorous evaluation using standardized metrics, the proposed system shows great potential to transform predictive recruitment by automating the process and significantly improving the accuracy of candidate performance predictions.

This paper is organized as follows: it first details the data sourcing and preparation process, followed by the development of the semantic database and the stages of model engineering and execution. Next, the evaluation results are presented, along with a discussion of the challenges encountered and the solutions implemented. The findings highlight the feasibility and impact of integrating semantic analysis and machine learning to advance predictive hiring technologies.

Related works

The integration of artificial intelligence and machine learning in predictive recruitment, and in particular the assessment of soft skills, represents a dynamic and promising area of research. address the problematics of semantic matching in Transformer-based pre-trained models.[1] address the problematics of semantic matching in Transformer-based pre-trained models. The authors propose a Dual Path Modeling Framework to enhance the model's ability to perceive subtle differences in sentence pairs by separately modeling affinity and difference semantics. However, the article's limitations concern the absence of an evaluation in a context of real and

heterogeneous data as encountered in recruitment. Lin Li et al.[2] propose a Code-Enhanced fine-grained semantic matching method for Tag Recommendation in software information sites (CETR) to learn the matching score between tags and software objects. In the CETR, code-enhanced semantic interaction is designed to capture fine-grained semantic relevance between tags and software objects. Yanmin chen et al.[3] propose a Self-attention Relational Sentence Semantic Matching (SR-SSM) framework, which utilizes the LSTM network to obtain the original sentence representation, and incorporates the information of adjacent sentences to obtain context-aware sentence representation through the self-attention mechanism.

Kaili Wang et al.[4] proposed a Multi-granularity Knowledge Enhancement (MGKE) model that improves the accuracy of short text semantic matching and introduces an attention mechanism to capture the hidden information at both character and word granularity which enhances the semantic information of the sentences. Furthermore, Han-Jia Ye et al.[5] introduced an evaluation criterion by predicting the comparison's correctness after assigning the learned embeddings to their optimal conditions, which measures how much Weakly Supervised Conditional Similarity Learning (WS-CSL) could cover latent semantics as the supervised model. Zanxia Jin et al.[6] proposed a Text-VQA method to alleviate OCR errors via OCR token evolution, which introduces a vocabulary predictor with character-level semantic matching, which enables the model to recover the correct word from the vocabulary even with misspelled OCR tokens. Besides, Chen Xu et al.[7] formalized the task of semantic sentence matching as a graph-matching problem, where each sentence is represented as a directed graph based on its syntactic structure. Their proposed method, called Interacted Syntax Graphs (ISG), combines the syntactic alignments of two sentences and their semantic matching signals into a single association graph. Subsequently, they adapted neural quadratic assignment programming (QAP) to extract syntactic matching patterns from this association graph. This approach allows for a detailed and interactive examination of the syntactic structures during the matching process.

Other works focus on developing an intelligent matchmaking model between the company's needs and the candidates applying for a specific offer. This matchmaking is based on the candidates' technical skills and other aspects such as age, gender, and years of experience to predict the candidate's performance and who will be the recruit in a given position. Mustafa Agaoglu[8] discussed several features and variables that were essential in determining employee performance. An attribute usage analysis was performed using many classifiers. The C5.0 algorithm showed a maximum usage of attributes compared to other classifiers like SVM and CART. Guo et al.[9] undertook research and developed an automated model that mapped a candidate's resume to a potential job posting. Factors like job title, study area, competitive level, and education degree were accumulated from online job postings to develop this predictive job model. In a study by Cavnar and Trenkle[10] n-grams approach was employed to classify job subjects. The dataset used constitutes 778 articles from 5 different newsgroups. The text data was available in seven class labels of the questionnaire. The classification accuracy could have been higher, ranging from 30 % to 80 %. Suryadi et al.[11] discussed the effect of a parent's career on choices made by students. This study selected a sample of 278 students, using a purposive sampling technique. Multiple regression and Confirmatory Factor Analysis (CFA) were performed to analyze the data.

Sushruta Mishra et al.[12] developed an intelligent predictive model to decide upon a candidate's suitability for an applied It based job using the KNN (K-Nearest Neighbours) algorithm combined with a hard-voting approach is employed.

Sridevi G.M et al.[13] developed an Artificial Intelligence (AI) system to measure and predict a suitable candidate from an available Candidate Resume (CR) database. The Jaccard similarity is measured between these clusters, and a suitability measure is proposed based on the cluster parameters. The prediction of candidate suitability is performed using the three classifiers: linear regression, decision tree, Adaboost, and XGBoost. Various features are formed by employing the bag of words technique to carry out the classification tasks. Ayishathahira C H et al.[14] developed a system for resume parsing using deep learning models using the convolutional neural network (CNN), Bi-LSTM (Bidirectional Long Short-Term Memory), and Conditional Random Field (CRF) to classify a resume into three segments and extract 23 fields. Marcu Florentina[15] application the web scraping to extract a massive amount of data from websites using the UiPath automation tool.

The work conducted by Ivo Wingsa et al.[16] presents a closely related approach to our proposed model. They propose a system that effectively extracts hard and soft skills from candidates' resumes and job descriptions using token classification. In another study, Lamjid A et al.[17] introduce a novel predictive micro-model specifically tailored for Information Technology consultants, utilizing soft skills as the basis for prediction. Furthermore, Tongshan Chang et al.[18] introduce a real-time project module that employs machine learning to assist higher education institutions in enrollment tasks, emphasizing the crucial role of information mining in job recruitment.

In other study, Bodhvi Gaur et al.[19] focuses on the challenge of extracting educational institutions' names and degrees from resume education sections. The authors propose a semi-supervised approach using a deep neural network model trained on a small annotated dataset. The model predicts entities in unlabeled sections, corrected by a module, and achieves 92,06 % accuracy through iterative training updates. Moreover, Silvia

Fareri et al.[20] introduced SkillNER, a data-driven method designed to automatically extract soft skills from text. This system employs named entity recognition (NER) and is trained using a support vector machine (SVM) on a corpus of over 5 000 scientific papers. The authors developed SkillNER by evaluating its performance against various training models and validating the findings with input from a team of psychologists. Additionally, SkillNER was tested in a real-world case study, utilizing job descriptions from ESCO (European Skill/Competence Qualifications and Occupations) as the source material. Asmaa Lamjid et al.[21] Introduced a novel predictive model that leverages Artificial Intelligence in the hiring process. By analyzing soft skills extracted from CVs, cover letters, websites, professional social media, and psychometric tests, the model accurately predicts potential candidates suitable for specific job roles. This system eliminates poor hiring decisions, reduces time and effort, minimizes recruitment costs, and mitigates turnover risks. Implementing the proposed model employs various predictive machine learning classifiers with key input soft skills, including creativity, collaboration, empathy, curiosity, and critical thinking.[22,23,24]

METHOD

Model Description

Overview of the model

Type of study: this is a non-observational computational modeling study, aimed at developing and evaluating an artificial intelligence driven predictive hiring system to extract and analyze soft skills from heterogeneous candidate data. The study is experimental in its nature, as it focuses on building, training, and validation of machine learning and meta-learning models for performance prediction.

Universe and sample: the study's universe consists of candidate data coming from online job portals, professional networking sites, and psychometric assessment providers. The sample contains over a million anonymized candidate profiles in the form of their résumés, professional descriptions, and psychometric test outcomes. These have been collected from diverse recruitment environments, ensuring representation across sectors, experience levels, and linguistic contexts.

Variables: the current study considered two broad categories of variables: the input variables consisting of textual elements extracted from résumés and online professional profiles, such as keywords, key phrases, and descriptive segments, besides soft-skill indicators sourced from psychometric assessments, and other metadata with respect to the candidates' backgrounds, if available. The output variables consisted of the predicted soft-skill classification outputs of the model; the scores on candidate performance inferred from the system's analytics; and the confidence levels of each such prediction. These variables collectively provide the base for the semantic modeling and evaluation framework adopted in this research.

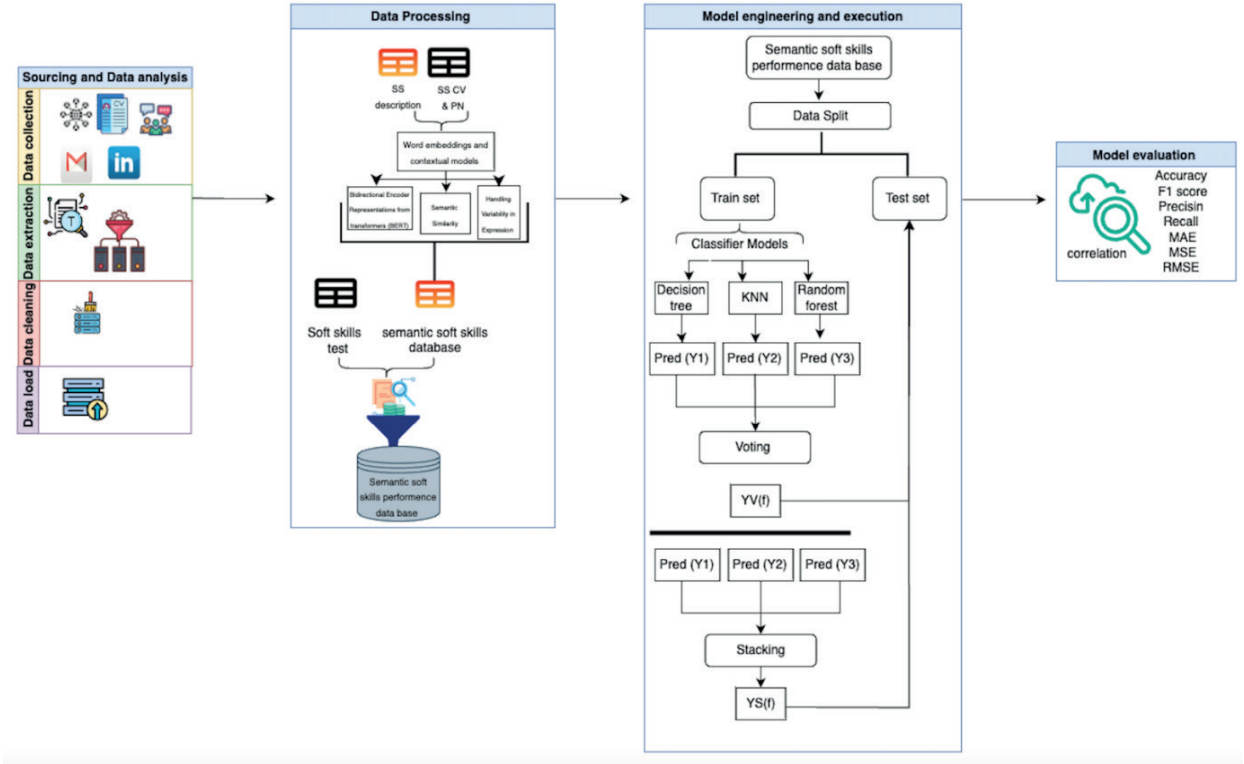

Figure 1. Innovative predictive hiring soft skills model

Data collection and Processing: the proposed model consists of four key stages, as illustrated in figure 1 first, we start with sourcing and data analysis, which involves collecting CVs and candidate descriptions from diverse sources such as job portals, social media platforms, and professional networking sites. During this phase, essential data, particularly soft skills, are extracted. This step is crucial for efficiently structuring the database in subsequent stages. Subsequently, data preprocessing is instrumental in the success of the entire model; during this phase, we establish a semantic database designed to uncover the intricate relationships between keywords, phrases, and descriptive texts. This innovative approach significantly enhances the system's robustness and accuracy in identifying behavioral skills across diverse languages and sources, ensuring a reliable model for the following assessments. The third and pivotal phase is the Model Engineering & Execution phase. In this phase, we integrate machine learning algorithms with a meta-model, such as XGBoost. This combination demonstrated exceptional performance in predicting candidates' skills and overall performance. The results highlight this system's effectiveness and relevance for assessing skills and selecting the best candidates. Finally, we reach the model evaluation stage, which marks the conclusion of our multifaceted process. Here, we subject our predictive model to rigorous testing and analysis. The outcomes of this evaluation reinforce the system's effectiveness and relevance in assessing skills and selecting optimal profiles. This thorough scrutiny not only guarantees the model's reliability but also its practicality in real-world hiring scenarios. Ultimately, the outcomes confirm the ability of this model to support the formation of effective, well-rounded teams by ensuring that only the best candidates are identified and selected.

Ethical Standards: all data used in the research were completely anonymized before processing. This study follows ethical guidelines regarding data protection, privacy preservation, and responsible use of artificial intelligence. No Personally Identifiable Information is accessed or retained, ensuring the research meets the standard ethical and data governance requirements.

Revealing Insights: A Thorough Analysis of Model Details

Sourcing and Data analysis

This phase focuses on four essential phases: sourcing, data extraction, cleaning, and load.

a) Sourcing: sourcing is not just an initial step but a crucial one in the recruitment process. It's an active approach aimed at broadening the pool of candidates and identifying profiles that match the company's needs, often before they have applied. This phase is the foundation for the subsequent phases, as it caters to the unique requirements of each company's recruitment process. Every company seeks distinct profiles and specialties while adhering to selection criteria based on skills, abilities, and motivations. When a recruiter faces a vast pool of candidates, the predictive recruitment model becomes invaluable. It is a tool that ensures the selection of a limited number of candidates possessing all the necessary qualifications and soft skills essential for the job, giving the HR professionals a sense of control and confidence in the security and certainty of their decisions in the recruitment process. Once the candidates have been identified in the previous phase (sourcing), the Data extraction phase begins.

b) Data extraction: collecting each candidate's CV and profiles on professional social networks and descriptions from various sources, considering linguistic diversity and variations in CV presentation style. Then extracting all the relevant soft skills and descriptions from these sources, which we organize in databases filtered according to their origin. Each candidate undergoes a psychometric test adapted to their experience level and professional field. These tests, developed in collaboration with coaches and psychiatrists specializing in human aspects, provide an in-depth, personalized assessment of the candidate's behavioral skills. This assessment ensures the highest validity and reliability, making the psychometric test a crucial part of the recruitment process. The data extraction process involves extracting soft skills from resumes using resume parsing techniques, gathering information from professional networks through data scraping and storing it in the required format, and ranking soft skills for each candidate based on psychometric tests using intelligent question-answering methods. The data collected from the process above comprises information on the candidate's soft skills, such as communication, persuasion skills, negotiation, relationship building, empathy, teamwork, conflict resolution, emotional intelligence, time management, and work ethic. These ten soft skills are not just essential, but they are vital for every employee and candidate to possess in every job, develop, and demonstrate for success and career advancement. After the extraction phase comes the last two phases of this stage, which are data cleaning and data loading

c) Data cleaning: following the data extraction phase, the resulting file may contain unwanted lines, punctuation, bullets, and other irrelevant elements. In data cleaning, we employ advanced techniques such as string replacement methods and regular expressions to eliminate these elements from the data, ensuring its thoroughness and quality.

d) Data load: data load is a precise process that involves transferring the data sets obtained from the previous phases into different soft skills databases. The SS CV and PN database will store soft skills

extracted from resumes and professional network data. On the other hand, soft skills identified through the psychometric test will be stored in an SS TEST database, and the description extract will be stored in an SS description database. It is worth mentioning that soft skills derived from the psychometrical test will be converted into weights ranging from 1 to 4.

Data processing

At this stage, we developed a semantic database named the semantic soft skills dataset based on the data extracted in the previous stage, the soft skills description database (SS description database), and the CV and PN databases.

The database leverages advanced text-processing techniques to analyze large volumes of unstructured data and identify key phrases related to soft skills. To elaborate this database, word embeddings and contextual models were employed to detect and analyze the semantic Similarity and relationships between words, phrases, and descriptions.

The process of word embeddings and contextual models is outlined as follows:

a) Capture the semantic Similarity and relationship between words: we utilize a technique called semantic Similarity

For example, “teamwork” and “collaboration” may not be identical words but will have similar embeddings, helping the model identify them as related skills.

b) Analyze the context in which a word or phrase appears: we employ context-aware analysis (BERT). For instance:

“I lead the scoping workshops with the customer” implies soft skills in “team leadership,” “collaboration,” and “communication.”

c) Handling Variability in Expression:

We can find a description of the same skill in different ways. For example: “I lead the scoping workshops with the customer.”

and “ I manage scoping meetings with customers.” Both imply “leadership, communication, and collaboration “ as soft skills. Contextual models can understand this variability and map these phrases to the same skill.

These steps allow the extraction of all possible soft skills from different sources, whatever their description or synonyms. These skills are combined with the psychometric test results (SS TEST database) to obtain a coherent semantic soft skills performance database. To reinforce this correspondence, each soft skill’s index performance was calculated according to its frequency and importance in the source data. A range on a scale of 1 to 4 was assigned to the performance index (1 represents a low-performance level, 2 represents a medium performance level, 2 represents a high-performance level, and 4 indicates an exceptional performance level), taking into account the results of the psychometric tests to reflect the relevance of the soft skills extracted. The database provides a comprehensive resource for identifying trends and patterns in soft skills representation across professional profiles by systematically collecting and categorizing these keywords. More importantly, this database serves as a foundation for predictive modeling, enabling the efficient forecasting of soft skills in candidates or professionals, and is an essential resource for the subsequent modeling phases.

In conclusion, the database used for the ML pipeline consists of integer-based soft skills inputs ranging from 1 to 4. This structured approach ensures precise and meaningful analysis, empowering us to unlock the full potential of these essential interpersonal skills.

Model engineering and execution

The next phase involved developing a robust predictive system based on machine learning techniques. An ensemble learning approach was first implemented, combining multiple models to improve predictive performance. This involved combining three basic models: KNN (K-Nearest Neighbors), Decision Tree, and Random Forest. These models were combined using the voting method (soft and hard voting), where the final decision was based on the majority results of the three models. However, the initial performance could have been more satisfactory. The predictions lacked precision, notably due to the large size of the database (more than one million candidates with ten soft skills as inputs). To overcome these limitations, a meta-model was introduced. This combines the previous basic models with advanced algorithms, such as XGBoost, Logistic Regression, LightGBM using the stacking method. After several iterations, XGBoost demonstrated the best results regarding robustness and accuracy.

The meta-model developed also has a dynamic capability, enabling it to accept input data in real-time and update predictions automatically. This flexibility ensures that the system can adapt to new data.

Evaluation of the model

The system was developed in a multi-stage technical pipeline that integrated data extraction, semantic

processing, and predictive modeling. First, candidate information that was retrieved from résumés, professional platforms, and psychometric assessments was cleaned, normalized, and encoded by tokenization, lemmatization, and contextual embedding techniques. Then, a semantic database was constructed to capture the relationships between soft-skill indicators and their descriptive expressions across heterogeneous sources. Several machine learning models, such as K-Nearest Neighbors, Decision Trees, and Random Forests, were implemented during the engineering phase, followed by a meta-model that combined XGBoost, Logistic Regression, and LightGBM to enhance predictive accuracy and generalization. Finally, the completed system was validated with standard evaluation metrics. For classification tasks, it used accuracy, precision, recall, F1-Score, and the Matthews Correlation Coefficient (MCC) to assess predictive quality. For regression tasks, it used Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to measure the deviation between predicted and observed values. A further correlation analysis was carried out to investigate the relationship between the predicted soft skills and real professional performance, confirming the relevance and superiority of the XGBoost meta-model, with higher accuracy, stronger generalization capabilities, and more stable performance than baseline models.

RESULTS

Presentation of the results

This section describes the methods and algorithms for predicting the performance index, including voting and stacking using the KNN, random forest, and decision tree as a basic model and logistic regression, LightGBM, and XGBoost as a méta Model. It also discusses the results of the two methods in this section.

Voting Soft skills predictive hiring model

Predictive Hiring Soft Skills: Implementing models Before Voting

Before presenting and analyzing the results obtained from comparing the KNN, Decision Tree, and Random Forest models, it determined the optimal parameters for each algorithm. For the KNN model, following the method described in the article by Lamjid et al.[21], which helped us ascertain the optimal value of K. In our case, the optimal value was K = 25. For both the Decision Tree and Random Forest models, we optimized the parameters by adjusting the random state, with the best performance achieved at a value of 42. These optimizations ensure maximum performance for each model in our specific context.

Table 1. Models results before voting

SSPML Evaluations metrics	KNN	DECISON TREE	RANDOM FOREST
Accuracy test	0,88	0,86	0,89
Accuracy train	0,90	0,91	0,92
Precision	0,88	0,86	0,89
Recall	0,88	0,86	0,89
F1 score	0,88	0,86	0,89
MCC	0,76	0,71	0,80
MSE	0,11	0,14	0,09
MAE	0,11	0,14	0,09
RMSE	0,34	0,14	0,30
Correlation (R_2)	0,54	0,44	0,61

The random forest outperforms KNN and decision tree across all metrics, achieving the highest accuracy, 89 %. However, the individual results of the models present suboptimal outcome limitations in terms of accuracy and generalisability, in particular in the context of a large database containing over a million profiles. To overcome these limitations, we combined the three models to improve the system's robustness, prediction, accuracy, and performance, taking advantage of ensemble learning methods based on voting.

Predictive Hiring Soft Skills: Implementing models after voting

Voting is an ensemble learning technique that combines the predictions of multiple individual models to improve overall performance and robustness. In this approach, the outputs of base models are aggregated, either through hard voting (majority rule for classification) or soft voting (averaging predicted probabilities). By leveraging the strengths of diverse models, voting helps to reduce the impact of individual model biases and enhances the accuracy and generalization of the system. It is effective when the base models are diverse and

complementary.

Hard Voting Function:

$\hat{y} = \text{mode}\{M_1(x), M_2(x), \ldots, M_n(x)\}$

Where M, M2… are the models and x is the sample

Soft voting Function:

$\hat{y} = \text{argmax}\ ((1/n)\ \Sigma_{i=1}^{n}\ p_i(x)\ )$

Where pi(x) is the probability of the class given by model i.

a) Evaluation of soft voting and hard voting models before optimization

The voting soft skill model was evaluated for its training and testing accuracies, which yielded a testing accuracy for hard voting of 91 % and testing accuracy for soft voting of 86 %. The accuracy assessments demonstrate that the hard voting model can effectively classify instances. However, the testing accuracy of the soft voting model needs to be improved to our goal, which is to exceed the 90 % accuracy threshold. To accomplish this, we employ an optimization technique that focuses on choosing the optimum weights to enhance the overall accuracy of soft voting and minimize false predictions.

b) Evaluation of soft voting and hard voting models after optimization

Table 2. Soft skills voting model's results before optimization

Evaluations metrics	**Soft skills voting models**	
	Hard Voting	**Soft Voting**
Accuracy test	0,91	0,86
Accuracy train	0,96	0,91
Precision	0,91	0,86
Recall	0,91	0,86
F1 score	0,91	0,86
MCC	0,83	0,73
MSE	0,08	0,13
MAE	0,08	0,13
RMSE	0,29	0,37
Correlation (R_2)	0,90	0,47

To improve the predictive performance and robustness of the ensemble models, we applied targeted optimization techniques adapted to each voting strategy. We employed a weight-tuning procedure via grid search combined with 5-fold cross-validation for the soft voting model. The best configuration was obtained with the weight vector [1, 1, 2], reflecting a greater influence of the Random Forest classifier within the ensemble. The base classifiers were optimized individually prior to ensemble integration: the Decision Tree used a maximum depth of 10 and Gini impurity as the splitting criterion; the KNN model was configured with k=7 and used the Euclidean distance metric; and the Random Forest consisted of 150 estimators, each with a maximum depth of 12. These settings were selected based on their stability across folds and their ability to generalize. The optimal configuration [1, 1, 2]—assigning higher importance to the Random Forest classifier—yielded the best balance between precision and generalization, as reflected by an accuracy of 89 % on the test set.

In parallel, we refined the ensemble performance for the hard voting strategy by integrating stratified K-fold cross-validation from the earliest stages of model training. This ensured better stability and generalization of individual base models by reducing variance across folds and mitigating potential bias due to class imbalance. This approach resulted in a test accuracy of 91 %. Although, this score is already aligned with our preliminary objectives, further improvements in precision and recall remain necessary to achieve the level of predictive robustness required for reliable candidate profiling in recruitment contexts.

To achieve these objectives regarding prediction robustness, we exploited stacking techniques and developed several meta-models, combining the three basic models used in the previous steps. This approach aimed to compare these meta-models to identify the one offering the best performance in selecting the profiles best suited to the company's needs, with an accuracy of over 91 %. The meta-models developed include the XGBoost meta-model, the LightGBM meta-model, and the Logistic Regression meta-model.

Table 3. Soft skills voting model's results after optimization

Evaluations Metrics	Soft skills voting models	
	Hard Voting	Soft voting
Accuracy test	0,91	0,89
Accuracy train	0,96	0,97
Precision	0,91	0,89
Recall	0,91	0,89
F1 score	0,91	0,89
MCC	0,83	0,77
MSE	0,085	0,11
MAE	0,085	0,11
RMSE	0,29	0,34
Correlation (R_2)	0,72	0,58

Stacking Soft skills predictive hiring model

To enhance performance, the stacking soft skills model is built over three meta-models: the XGBoost meta-model, the LightGBM meta-model, and the Logistic Regression meta-model. These models are characterized by their flexibility, robustness, and efficiency, allowing them to capture complex interactions, prevent overfitting, and maintain high-performance levels.

Evaluation of Logistic regression soft skills meta model

Logistic regression is a model that uses log loss as a function to predict the probability of belonging to one or more classes. To achieve an accuracy of 91 % on the test set and 95 % on the training set, as mentioned in table 4, the optimal log loss value was 0,156.

To evaluate the robustness of the Logistic Regression soft skills stacking model, we used ten different evaluation metrics:

Table 4. Insights from the Logistic Regression Soft Skills Meta-Model Evaluation Metrics

Evaluations Metrics	Logistic regression meta model
Accuracy Test	0,91
Accuracy train	0,95
Precision	0,91
Recall	0,91
F1 score	0,91
MCC	0,81
MSE	0,09
MAE	0,09
RMSE	0,31
Correlation	0,725

Evaluation of LightGBM soft skills meta model

LightGBM is a boosting method optimized for fast calculations. It uses num_leaves, learning_rate, and mew_depth as parameters. As shown in table 3, the optimum parameters are equal to num_leaves = 42 and max_depth = [5,10,15] to achieve 93 % test accuracy.

To evaluate the robustness of the LightGBM soft skills stacking model, we used ten different evaluation metrics:

Table 5. Insights from the LightGBM Soft Skill Meta-Model Evaluation Metrics

Evaluations Metrics	LightGBM meta model
Accuracy Test	0,93
Accuracy train	0,96
Precision	0,93

Recall	0,93
F1 score	0,93
MCC	0,89
MSE	0,065
MAE	0,065
RMSE	0,245
Correlation	0,87

Evaluation of XGBoost soft skills meta model

XG boost is a boosting-based method designed to minimize prediction errors by adjusting the parameters n_estimators, learning rate, and max_depth for a robust model. As shown in the table 6, the optimum n_ estimators equal 100, max_depth equal [5,10,15], and the eval metric = log loss to acquire 98 % accuracy.

To evaluate the robustness of the XGBoost soft skills stacking model, we used ten different evaluation metrics:

Table 6. Insights from the XGBoost Soft Skills Meta-Model Evaluation Metrics

Evaluations Metrics	XGBoost meta model
Accuracy Test	0,98
Accuracy train	0,97
Precision	0,98
Recall	0,98
F1 score	0,98
MCC	0,97
MSE	0,02
MAE	0,02
RMSE	0,14
Correlation	0,94

Summary of the stacking soft skills models

Based on our insights from the tree stacking soft skills model, we can confidently conclude that the XGBoost stacking soft skill model demonstrates exceptional performance in predicting candidates. With an impressive accuracy rate of 0,98 for testing and 0,97 for training, the model consistently achieves a high accurate predictions. The model demonstrates high accuracy and low error rates, confirming its effectiveness as a valuable tool for evaluating candidates. Organizations can rely on this model to make accurate predictions, streamline their hiring processes, and secure the most suitable candidates for various positions.

The table below represents the summary of the stacking soft skills models:

Table 7. Summary of the comparison results for stacking soft skills model

Evaluations Metrics	Stacking soft skills models		
	Logistic regression meta model	LightGBM meta model	XGBoost meta model
Accuracy Test	0,91	0,93	0,98
Accuracy train	0,95	0,96	0,97
Precision	0,91	0,93	0,98
Recall	0,91	0,93	0,98
F1 score	0,91	0,93	0,98
MCC	0,81	0,89	0,97
MSE	0,09	0,065	0,02
MAE	0,09	0,065	0,02
RMSE	0,31	0,245	0,14
Correlation	0,725	0,87	0,94

The model showed strong predictive performance, as reflected by low values of MSE, MAE, and RMSE combined with a high MCC and consistent correlation between training and testing results. These results confirm that the system is able not only to classify unseen data but also to avoid overfitting-a necessary condition for any real-world recruitment application. Methodologically, these results point to the fact that semantic modeling combined with meta-learning techniques provides a more robust framework than traditional machine-learning approaches. Compared to previous works that relied on keyword-based extraction or shallow classifiers, our model demonstrates much better stability and generalization, especially on soft-skill prediction. Common pitfalls of previous studies included limited semantic understanding, poor cross-dataset transferability, or inflated performance on training data. By contrast, the consistent metrics across different datasets in this work illustrate a more reliable and scalable solution. These observations emphasize the relevance of combining contextual embeddings with state-of-the-art ensemble methods and further underline the additional value of the approach within the broader context of predictive hiring systems.

DISCUSSION

After the emergence of artificial intelligence, recruitment professionals aim to automate the traditional recruitment process and recruit the best profiles with high performance. When referring to high-performing profiles, we highlight the importance of soft skills, which are challenging to replicate.

Table 8. Implementation of the soft skills models

P- Metrics	KNN	Decision tree	Random forest	Voting soft skills models			Stacking soft skills models		
				Soft Voting before optimization	Soft voting after optimization	Hard voting	Stacking Logistic regression	Stacking LightGBM	Stacking XGBoost
Accuracy Test	0,88	0,86	0,89	0,86	0,89	0,91	0,91	0,93	0,98
Accuracy train	0,90	0,91	0,92	0,91	0,97	0,96	0,95	0,96	0,97
Precision	0,88	0,86	0,89	0,86	0,89	0,91	0,91	0,93	0,98
Recall	0,88	0,86	0,89	0,86	0,89	0,91	0,91	0,93	0,98
F1- score	0,88	0,86	0,89	0,86	0,89	0,91	0,91	0,93	0,98
MCC	0,76	0,71	0,80	0,73	0,77	0,83	0,81	0,89	0,97
MSE	0,11	0,14	0,09	0,13	0,11	0,085	0,09	0,065	0,02
MAE	0,11	0,14	0,09	0,13	0,11	0,085	0,09	0,065	0,02
RMSE	0,34	0,14	0,30	0,37	0,34	0,29	0,31	0,245	0,14
Correlation	0,54	0,44	0,61	0,47	0,58	0,72	0,725	0,87	0,94

This predictive hiring system streamlines the recruitment process, reduces the time spent on lower-value tasks, and decreases costs while minimizing employee turnover. It aids in making more informed hiring decisions and identifying high-performing employees. The system gathers data from multiple sources and builds a comprehensive semantic database of soft skills, allowing companies to evaluate candidates with exceptional soft skills effectively.

The proposed method combines Decision Trees, K-Nearest Neighbors, and Random Forests within a stacking-based meta-learning architecture. Clearly, it outperforms traditional machine learning approaches commonly used in recruitment analytics. Contrary to the typically isolated classifiers or simple keyword-matching algorithms used in earlier studies, which led to limited semantic understanding and inconsistent generalization. Our method leverages the complementary strengths of ensemble learning and contextual feature modeling. Indeed, the stacking XGBoost model obtains a test accuracy of 98 %, significantly outperforming previously reported systems in the literature. Stacking LightGBM and logistic regression models achieves over 90 % accuracy, confirming the robustness of the ensemble framework. These results confirm that leveraging both semantic representations and meta-learning significantly enhances predictive reliability while improving transferability across heterogeneous profiles. Analytically, the performance gap between XGBoost and other models indicates that tree-boosting architectures are particularly effective at capturing nonlinear relationships in soft-skill indicators. Further enhancements can be considered to optimize loss functions, such as the logistic log-loss, to improve generalization. Overall, the method offers a substantial improvement over existing approaches and contributes a more scalable and accurate framework for automated soft-skill assessment.

Based on these findings, the Stacking XGBoost soft skills model emerged as the most reliable in predicting the performance index of selected candidates.

Although our findings show the promise that the suggested approach holds to enhance recruitment procedures through superior predictive power, it is as crucial to pay attention to the ethical and societal consequences

related to its implementation in the real world. The incorporation of artificial intelligence-driven systems in recruitment procedures brings core ethical concerns.

The approach developed in this study is uniquely designed to augment human decision-making, not replace recruiters. The proposed automation is meant to enhance the objectivity and efficiency of specific tasks, such as candidate pre-screening or analysis of the semantic content of the resume, but leave human expertise the final validation authority.

But the potential benefit of these tools - in terms of standardizing processes or reducing certain implicit biases[1] - cannot overshadow the risks inherent in algorithmic processing, specifically reproducing or even exacerbating existing biases embedded in the training dataset.[2] The quality of data, the modeling, and operational conditions under which they are put in place must remain subject to continuous monitoring, supported by supervision measures and regular audits.

Secondly, it has to be added that deployment of such technologies is a task of companies and HR professionals. Inability to disclose mechanisms for algorithmic decision-making or excessive delegation without adequate supervisory oversight can erode legitimacy of processes and compliance with current regulations. In this context, traceability of choices, explainability of models, and the use of appeal mechanisms appear to be a pre-requisite for the ethically sustainable use of AI in working settings.[3]

In fine, la mise en place de l'IA dans les processus de recrutement est à soutenir d'un encadrement méthodologique et éthique assurant la complémentarité homme-machine, la redevabilité des décisions, et le respect des règles fondamentales d'équité et de non-discrimination.

CONCLUSION

In conclusion, this study aimed to address the challenge of the reliable assessment of soft skills in a manner that is automated and data-driven, a quest still at the core of improving modern recruitment practices. By proposing a system with its underpinning from semantic modeling and artificial intelligence, the study presents that it is possible to move beyond traditional methods of evaluation toward more objective, scalable, and context-sensitive methodologies. From a technical contribution perspective, this work underlines the need for merging multidisciplinary views by integrating insights from artificial intelligence, behavioral sciences, and cognitive analysis in order to enhance the understanding of how soft skills will be manifested across heterogeneous candidate data. This research reinforces ideas on how predictive hiring systems must evolve toward models capable of capturing subtle human attributes while ensuring fairness, transparency, and adaptability.

Finally, the approach developed here has established a foundation for future explorations toward more comprehensive, ethically aligned, and scientifically grounded systems for soft-skill assessment in recruitment.

REFERENCES

1. Xue C, Liang D, Wang S, Wu W, Zhang J. Dual Path Modeling for Semantic Matching by Perceiving Subtle Conflicts. 2023.

2. Li L, Wang P, Zheng X, Xie Q. Code-Enhanced Fine-Grained Semantic Matching For Tag Recommendation In Software Information Sites. 2023.

3. Chen Y, Wang H, Sun R, Chen E. Context-Aware Semantic Matching with Self Attention Mechanism. 2022.

4. Wang K, Cao X, Yang X, Cao Y. Multi-granularity Text Semantic Matching Model Based on Knowledge Enhancement. 2022.

5. Ye HJ, Shi Y, Zhan DC. Identifying Ambiguous Similarity Conditions via Semantic Matching. 2022.

6. Jin Z, Shou MZ, Zhou F, Tsutsui S, Qin J, Yin XC. From Token to Word: OCR Token Evolution via Contrastive Learning and Semantic Matching for Text-VQA. 2022.

7. Xu C, Xu J, Dong Z, Wen J. Semantic Sentence Matching via Interacting Syntax Graphs. 2022.

8. Agaoglu M. Predicting Instructor Performance Using Data Mining Techniques in Higher Education. IEEE Access. 2016;4:1-1.

9. Guo S, Alamudun F, Hammond T. RésuMatcher: A personalized résumé-job matching system. Expert Syst Appl. 2016;60:169-82.

10. Cavnar WB, Trenkle JM. N-Gram-Based Text Categorization. Ann Arbor: Environmental Research Institute

of Michigan; 2019.

11. Suryadi B, Hayat B, Putra MDK. The Influence of Adolescent-Parent Career Congruence and Counselor Roles in Vocational Guidance on the Career Orientation of Students. Int J Instr. 2020;13(2):45-60.

12. Mishra S, Mallick PK, Tripathy HK, Jena L, Chae GS. Stacked KNN with hard voting predictive approach to assist hiring process in IT organizations. Int J Electr Eng Educ. 2021.

13. Sridevi G M, Suganthi SK. AI based suitability measurement and prediction between job description and job seeker profiles. Int J Inf Manag Data Insights. 2022;2(2):100109.

14. Ayishathahira CH, Sreejith C, Raseek C. Combination of Neural Networks and Conditional Random Fields for Efficient Resume Parsing. In: 2018 International CET Conference on Control, Communication, and Computing (IC4). 2018.

15. Marcu F. Web data extraction with robot process automation. Study on LinkedIn web scraping using UIPath Studio. Ann 'Constantin Brancusi' Univ Targu-Jiu Eng Ser. 2020.

16. Wings I, Nanda R, Adebayo KJ. A Context-Aware Approach for Extracting Hard and Soft Skills. Procedia Comput Sci. 2021;193:163-72.

17. Lamjid A, El Bouchti K, Ziti S, Mohamed RO, Labrim H, Riadsolh A, et al. Predictive Hiring System: Information Technology Consultants Soft Skills. In: International Conference on Advanced Intelligent Systems for Sustainable Development. 2022. p. 680-5.

18. Chang T, editor. Data mining: a magic technology for college recruitment. Paper of Overseas Chinese Association for Institutional Research; 2021.

19. Gaur B, Saluja GS, Sivakumar HB, Singh S. Semi-supervised deep learning based named entity recognition model to parse education section of resumes. Neural Comput Appl. 2021;33:5705–18.

20. Fareri S, Melluso N, Chiarello F, Fantoni G. SkillNER: Mining and mapping soft skills from any text. Expert Syst Appl. 2021;184.

21. Lamjid A, Ariss A, Ennejjai I, Mabrouki J, Ziti S. Enhancing the hiring process: A predictive system for soft skills assessment. Dep Comput Sci Fac Sci Mohammed V Univ. 2024.

22. Lal N, Benkraouda O. Exploring the Implementation of AI in Early Onset Interviews to Help Mitigate Bias. arXiv preprint arXiv:2501.09890. 2025.

23. Peng A, et al. Investigations of performance and bias in human-AI teamwork in hiring. Proc AAAI Conf Artif Intell. 2022;36(11).

24. Yanamala KKR. Transparency, privacy, and accountability in AI-enhanced HR processes. J Adv Comput Syst. 2023;3(3):10-8.

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Asmaa Lamjid, Anass Ariss.
Data curation: Asmaa Lamjid, Anass Ariss, Jamal Mabrouki, Soumia Ziti.
Formal analysis: Asmaa Lamjid, Anass Ariss, Karim El Bouchti, Soumia Ziti.
Research: Asmaa Lamjid, Anass Ariss, Jamal Mabrouki, Soumia Ziti.
Methodology: Asmaa Lamjid, Anass Ariss, Jamal Mabrouki, Karim El Bouchti, Soumia Ziti.
Supervision: Soumia Ziti.

Validation: Asmaa Lamjid, Anass Ariss, Soumia Ziti.
Drafting - original draft: Asmaa Lamjid, Anass Ariss.
Writing - proofreading and editing: Asmaa Lamjid, Anass Ariss, Soumia Ziti.