



AI-Powered Career Guidance System: A Machine Learning Approach Using Holland's RIASEC Model for Personalized Occupation Recommendations

Dr. Naresh Thoutam ¹, Rohini M. Uphade ², Komal A. Bhabad ²,
Shreyas Tale ², Tanmay Mahajan ²

¹Associate Professor, Dept. of Computer Engineering, Sandip Institute of Technology and Research Centre, Nashik

²Dept. of Computer Engineering, Sandip Institute of Technology and Research Centre, Nashik

naresh.thoutam@sitrc.org, uphaderohini99@gmail.com, komalbhabad105@gmail.com, Shreyastale7222@gmail.com,
tanmaykm26@gmail.com

Corresponding author- Dr. Naresh Thoutam

DOI: 10.5281/zenodo.23640724

Abstract: Career guidance continues to be a challenge to students and professionals searching for optimal vocational placement. Current methods of counselling use manual analysis of psychometric assessment answers by trained counsellors, which is not scalable and subject to subjective bias. This paper proposes an end-to-end machine learning-based career guidance system that incorporates Holland's RIASEC vocational interest theory into a web-based system capable of recommending 923 ONET occupations from a 48-item psychometric assessment and providing plain-explained reasoning. Decision Tree, Support Vector Machine (SVM) and XGBoost regressors were trained on 8,855 real responses from Open Psychometrics and compared with a traditional Pearson-correlation baseline profile matching method. Evaluation on three complementary metrics — Hit Rate, Euclidean Distance and Profile Correlation — shows that ML-augmented methods significantly outperform the baseline, with SVM achieving an 81.93% Hit Rate compared with 34.68% by the Profile Method. A recommendation API, interactive web interface with radar-chart visualization and explainability of occupation recommendations via dimension-overlap are proposed. This work bridges vocational psychology with practical AI engineering, providing a scalable and interpretable solution for career counselling suitable for deployment in educational institutions.

Keywords: Career guidance; RIASEC model; Machine learning; XGBoost; Explainable AI; Occupation recommendation; Psychometric assessment; ONET database

1 INTRODUCTION

Choosing a career is an important decision in a person's life because it can have major impact on their job satisfaction and future success. However, the availability of a system that can provide personalized, data-driven guidance that matches individual interest profiles with suitable occupations is still lacking [1]. Traditional counselling methods mostly depend on trained counsellors to manually interpret information. This can make the process difficult to handle on a large scale and may introduce personal judgement into the counselling process. With the increase in the number of recognized occupations and the rapid evolution of job markets, there is a pressing need for systems that can process psychometric data and make accurate, personalized recommendations.

John Holland's RIASEC theory provides a framework for understanding vocational interests [1], encompassing six dimensions: Realistic (hands-on, tool-oriented work),

Investigative (analytical, scientific enquiry), Artistic (creative, expressive activity), Social (helping and teaching), Enterprising (leadership and persuasion) and Conventional (organized, detail-oriented work). The framework has been extensively validated across populations and underlies major occupational databases, including ONET, which profiles more than 900 occupations along the same six dimensions.

This paper proposes a complete implementation of an AI-powered career guidance system that: (i) predicts six-dimensional RIASEC profiles from 48-item questionnaire responses, (ii) compares Decision Tree, SVM and XGBoost regressors against a traditional profile-matching baseline, (iii) generates Top N personalized career recommendations with explainable, plain-language justifications, and (iv) is deployed as a working web application.

The main contributions are a reproducible ML pipeline built on 8,855 real psychometric responses, a comprehensive evaluation

showing that ML models achieve 81.93% Hit Rate against a 34.68% baseline (a 136% improvement), an explainable-AI component based on dimension overlap rather than raw feature attribution, and an ablation study quantifying the contribution of demographic versus item-level features to prediction accuracy.

The remainder of this paper is organized as follows. Section II reviews the closest prior literature in a comparative table. Section III introduces background concepts and definitions. Section IV discusses related work in more depth and positions the present contribution against it. Section V details the methodology, including data sources, preprocessing and

algorithms. Section VI presents the proposed system architecture. Section VII presents the experimental results. Section VIII discusses the findings, limitations and ethical considerations, and Section IX concludes with directions for future work.

II. LITERATURE REVIEW

To situate the present contribution, recent studies encompassing vocational-interest theory, computational career recommendation, machine learning in psychometrics and explainable AI were reviewed. The key findings or limitations and methods of the most closely related studies are summarized in Table 1.

Table 1: Summary of reviewed literature on career guidance, RIASEC modelling and explainable recommendations

Study	Author(s), Year	Focus / Approach	Dataset	Method / Technique	Key Finding / Limitation
[1]	Holland, 1997	Foundational vocational theory	N/A — theoretical	RIASEC six-dimension hexagon model	Established the RIASEC taxonomy underlying almost all subsequent work, but offers no automated computation
[2]	Roy and Sharma, 2026	AI-powered career guidance	Custom psychometric survey	ML-based RIASEC profile matching	Establishes a baseline methodology for AI career guidance; limited discussion of explainability or ablation
[4]	Roy and Shah, 2015	Collaborative-filtering recommendation	Peer outcome history	Collaborative filtering	Effective when historical peer data exists but suffers from cold start for new users
[5]	Almalis et al., 2015	Academic career counselling framework (FoDRA)	Institutional academic records	Content-based filtering on occupation attributes	Rule-based matching; does not learn non-linear item-to-profile mappings
[17]	Alqahtani and Al-Dossari, 2020	Career path choice for IT graduates	2,255 IT employee records	Comparison of five classifiers, incl. XGBoost	XGBoost gave the best accuracy (70.47%), but the study does not compare against a psychometric baseline
[18]	Song et al., 2024	ML for career-choice prediction	Longitudinal career survey data	Supervised ML models	Confirms ML meaningfully improves career-choice prediction over classical methods
[19]	Thamrin et al., 2023	RIASEC validity for vocational students	Vocational student sample	Reliability/validity analysis	Supports predictive validity of RIASEC but does not implement an automated system
[21]	Ogunde and Idialu, 2019	Recommender for training placement	Student placement records	Recommender system	Domain-specific to industrial training; not generalized to full occupation databases
[22]	Kiran et al., 2018	Career and skills recommendation	Regional employment data	Data-mining classifiers	Good regional accuracy but limited to a narrow set of career categories

A common theme emerges from Table 1: while RIASEC theory and its occupational linkage (ONET) are mature and well validated [1], [9], [11], [12], [19], computational implementations either rely on collaborative or content-based filtering without learning non-linear item-to-profile mappings [4], [5], [21], or apply ML to narrower, domain-specific datasets without a systematic multi-algorithm comparison, an ablation study, or a lightweight explanation mechanism aimed at non-technical end users [2], [17], [18], [22]. This gap

motivates the framework proposed in the remainder of this paper.

III. BACKGROUND AND DEFINITIONS

A. Holland's RIASEC Model

Holland's RIASEC theory [1] explains vocational interests through six dimensions:

- Realistic (R): preference for hands-on, practical, tool-oriented work.

- Investigative (I): inclination toward analytical, scientific, inquiry-based work.
- Artistic (A): attraction to creative, expressive, original activity.
- Social (S): interest in helping, teaching and community-oriented work.
- Enterprising (E): preference for leadership and persuasive activity.
- Conventional (C): affinity for organized, detail-oriented, structured work.

This framework has been extensively validated across populations and integrated into major occupational databases, including ONET, which profiles more than 900 occupations using the same six dimensions [9].

B. Traditional Profile-Matching Methods

Conventional career counselling practice, mirrored computationally by the Profile Method baseline used in this study, ranks candidate occupations by the Pearson correlation between a respondent's six-dimensional RIASEC vector, obtained by averaging item responses within each dimension, and each occupation's published RIASEC vector. The method is simple and transparent but discards item-level detail and any non-linear structure among the 48 underlying responses.

C. Explainable AI in Recommendation

Explainable artificial intelligence denotes approaches that make the reasoning of machine learning models interpretable to humans. Post-hoc, model-agnostic techniques such as LIME [7] and SHAP [13] approximate or decompose a model's output into feature-level contributions, while Miller [8] argues from the social sciences that useful explanations are contrastive, selective and expressed in terms familiar to the recipient. These principles motivate the dimension-overlap explanation strategy adopted in Section V, which favors transparency that matches counsellors' own explanatory style over the finer-grained but more technical output of LIME or SHAP.

D. Overview of Candidate Regression Algorithms

In this study, three supervised learning algorithms were tested as multi-output regression models. These models use 48 raw Likert-scale responses as input and predict scores for the six RIASEC dimensions. The Decision Tree Regressor divides the input data using a series of simple splits, making its results relatively easy to understand and visualize. However, it can sometimes overfit the responses to individual questionnaire items. The Support Vector Machine (SVM) [14] uses an epsilon-insensitive margin to find a suitable regression function and can work effectively with structured data such as questionnaire responses. XGBoost [3] combines several small regression trees through gradient boosting. This allows it to

identify non-linear relationships between different questionnaire items while maintaining relatively fast training and tuning. Its performance is also considered in relation to Random Forest-style ensemble methods [23]. Finally, the results of all three models are compared with the Profile Method baseline discussed in Section III-B.

IV. RELATED WORK

A. Vocational Interest Assessment

Since its introduction in 1959, Holland's RIASEC model has remained one of the most widely used frameworks in vocational psychology [1]. Gottfredson and Holland [11] highlight the practical usefulness of theory-based occupational classification, particularly the high-point code, in career counselling. Similarly, Rounds and Tracey [12] and Thamrin et al. [19] studied the validity of RIASEC measures across different populations and found that the measures show some similarities across groups, although they are not completely equivalent. This limitation is discussed further in Section VIII.

B. Career Recommendation Systems

Collaborative filtering has also been used for career recommendations by identifying similarities between users and peers whose career outcomes are already known [4]. However, this approach can face a cold-start problem when dealing with new users who have little or no previous data. Another approach is content-based filtering, which recommends careers by using information about the characteristics or attributes of different occupations [5], [21].

Domain-specific systems for IT graduates [17] and regional labor markets [22] show that ML classifiers outperform manual matching within their target populations but are not evaluated against a psychometric baseline or a full occupation database. Related tools include a fuzzy-logic recommender for secondary-school students [20], a skills-gap-driven course recommender for graduates [24], and a job-role-fit classifier built on resume and profile features [25]; each targets a narrower population or a single stage of the career-decision process rather than a psychometrically grounded, full-database recommendation pipeline. The present work extends this line of research by learning non-linear mappings from raw psychometric items to RIASEC profiles across the full 923-occupation ONET database.

C. Machine Learning in Psychometrics

Recent studies show that ML algorithms can predict psychological constructs from questionnaire data with accuracy exceeding classical test-theory methods [6], [18]. XGBoost [3] and ensemble tree methods more generally [23] have shown particular promise for modelling complex trait interactions. Roy and Sharma [2] recently proposed an AI-powered career

guidance framework using RIASEC profiles; the present work implements and extends this framework with a comprehensive ablation study and explainable recommendation generation.

D. Explainable AI in Recommendation

The black-box nature of ML models poses challenges for user trust in high-stakes domains such as career guidance [7]. Model-agnostic explanation methods such as LIME [7] and SHAP [13] emphasise feature importance, while Miller [8] argues good explanations should be contrastive and selective. The present system addresses this through dimension-overlap explanations that align with users' intuitive understanding of vocational interests rather than raw feature attribution.

E. Summary of Identified Gap and Positioning

Table 2 compares the proposed work with the studies that are most closely related to it.

Table 2: Positioning of this work against the closest prior studies

Study	Matching Method	Algorithms Compared	Explanation Mechanism
Roy and Shah [4]	Collaborative filtering	Single method	None
Almalis et al. [5]	Content-based (FoDRA)	Rule-based, not ML	None
Alqahtani and Al-Dossari [17]	Skills-based classification	Five classifiers, incl. XGBoost	None
This work	ML regression + Pearson correlation ranking over 923 occupations	Decision Tree, SVM, XGBoost vs. Profile Method baseline	Dimension-overlap, plain-language explanation

Previous studies have not combined all three aspects in a single approach: a systematic comparison of multiple algorithms with a transparent baseline, an ablation study to examine the contribution of different features, and a simple explanation mechanism intended for non-technical users. This gap provides the motivation for the methodology and system architecture presented in Sections 5 and 6.

V. METHODOLOGY

A. Data Sources

For training, we used the publicly available Open Psychometrics RIASEC dataset, which contains 8,855 complete responses from users around the world [10]. Each response consists of 48 Likert-scale questions, with eight questions representing each RIASEC dimension and responses recorded on a scale from 1 to 5. The dataset also includes demographic details such as age, gender, and country, but does

not contain personally identifiable information. Around 5.2% of the data is missing, reflecting the type of incomplete responses that can occur in real-world psychometric assessments.

The O*NET 29.0 database was used to obtain RIASEC profiles for 923 occupations [9]. For each occupation, the database provides six-dimensional RIASEC scores on a 1.0–7.0 scale, standard occupational classification codes, job titles, descriptions, and high-point codes. These occupational profiles are based on systematic job analysis carried out by the U.S. Department of Labor.

B. Preprocessing and Feature Engineering

Missing responses in the RIASEC items were filled using K-Nearest Neighbours (KNN) imputation with $k = 5$. This method was selected because it can preserve the relationships between closely related questionnaire items [15]. Two feature representations were generated: item-level features (48 raw responses plus age and gender, 50 features) and dimension scores (the average of the eight items within each RIASEC dimension, six features), retained in parallel so that their relative contribution to accuracy could be isolated in the ablation study of Section VII-C. For each participant, a high-point code was computed by concatenating the top three RIASEC dimensions in descending order (for example, “IAS”). Data was split 80/20 into training and test sets using stratified sampling, yielding 7,084 training and 1,771 test samples, and features were standardized within cross-validation folds during hyperparameter tuning to prevent data leakage.

C. Regression Models and Baseline

The Profile Method baseline, mirroring traditional career counselling practice, ranks occupations by the Pearson correlation between a user's six-dimensional RIASEC vector and each occupation's RIASEC vector. Three multi-output regression models were then trained, using the scikit-learn and XGBoost libraries [3], [16], to predict the six RIASEC dimensions directly from the 48 raw items: a Decision Tree Regressor, a Support Vector Regressor with RBF and linear kernels [14], and an XGBoost Regressor [3]. All three were tuned by cross-validated random search over standard hyperparameter ranges (tree depth, regularization strength, learning rate and ensemble size) and trained to minimize Mean Squared Error across the six-dimensional output space.

D. Recommendation Generation and Explanation Module

At inference time, the trained regression model predicts a six-dimensional RIASEC profile from the user's 48 responses; this predicted profile is then correlated against all 923 occupation vectors, and the top N occupations by Pearson correlation are returned. For each recommended occupation, a natural-language explanation is generated by extracting the top three

RIASEC dimensions from the user profile and from the occupation profile, computing their intersection, and reporting the overlapping dimensions with semantic labels together with a profile congruence percentage (Pearson correlation $\times 100$) — for example, “Strong alignment in 3 core interest dimensions: Artistic (Creative, Expressive), Investigative (Analytical, Scientific). Profile congruence: 92.3%.” This design follows the contrastive, selective explanation principles discussed in Section III-C [8], favoring a small number of human-readable reasons over an exhaustive feature-attribution report such as LIME [7] or SHAP [13].

E. Evaluation Metrics

Per psychometric evaluation standard, three metrics were used:

$$HR = (1/n) \sum_{i=1}^n 1[\text{code_pred}^{(i)} = \text{code_true}^{(i)}] \quad (1)$$

where $\text{code_pred}(i)$ is the high-point code derived from the predicted RIASEC profile and $\text{code_true}(i)$ is the ground-truth code; Hit Rate measures whether the predicted vocational personality type matches the actual type.

$$ED = (1/n) \sum_{i=1}^n \|y_{\text{pred}}^{(i)} - y_{\text{true}}^{(i)}\|_2 \quad (2)$$

Euclidean Distance measures the average L2 distance between predicted and actual six-dimensional RIASEC vectors; lower values indicate more accurate profile reconstruction.

$$PC = (1/n) \sum_{i=1}^n \rho(y_{\text{pred}}^{(i)}, y_{\text{true}}^{(i)}) \quad (3)$$

where ρ denotes the Pearson correlation coefficient. Profile Correlation measures shape congruence between predicted and actual profiles, emphasizing relative dimension ordering over absolute scores.

F. Experimental Design and Threats to Validity

All models were trained on 80% of the data and evaluated on the held-out 20% test set with all three metrics. For a robust estimate of generalization, 5-fold stratified cross-validation was additionally performed on the full dataset. To quantify feature contribution, Decision Tree models were trained under three configurations: Full (48 items + age + gender), No Demographics (48 items only) and Psychometric Only (6 dimension averages, matching the baseline's feature granularity). Open Psychometrics data is the sole training source and skews towards younger, more educated respondents; together with the partly Western origin of the RIASEC construct itself [12], [19], this limits the cross-population generality of the reported figures, a point revisited in Section VIII.

VI. PROPOSED ARCHITECTURE

The system follows a three-tier architecture comprising a data layer, a stateless recommendation API and an interactive web interface, so that classification, recommendation and explanation remain independently testable and horizontally

scalable. The API layer accepts a user's 48 item responses and demographics, returns the predicted RIASEC profile and high-point code, and separately accepts a RIASEC vector to return a ranked list of recommended occupations with generated explanations; request and response validation and cross-origin access are handled at this layer. The web interface presents the questionnaire with real-time validation, a radar-chart visualization of the predicted RIASEC profile, and the top recommended careers with expandable detail and congruence scores, in a layout suitable for both desktop and mobile use. The trained models are stored as serialised files, while the O*NET occupation database is cached locally. This removes the need for an external database connection and makes deployment simpler. The stateless API design also allows the system to be replicated across multiple instances, with an average inference latency of less than 50 ms per prediction.

VII. RESULTS

A. Model Comparison

Table 3 presents the performance of all models on the test set, which contains 1,771 samples.

Table 3: Comparison of model performance on the test set (n = 1,771)

Model	Hit Rate	Euclidean Distance	Profile Correlation
Profile Method (baseline)	0.3468	3.58 ± 1.19	0.903 ± 0.075
Decision Tree	0.5817	0.40 ± 0.19	0.971 ± 0.070
SVM	0.8193	0.01 ± 0.00	1.000 ± 0.000
XGBoost	0.8192	0.08 ± 0.04	0.999 ± 0.006

SVM achieved the highest Hit Rate of 81.93%, representing a 136% improvement over the baseline value of 34.68%. Its near-perfect profile correlation ($PC \approx 1.0$) can be explained by the direct relationship between the 48 questionnaire items and the six RIASEC dimensions. XGBoost produced a very similar Hit Rate of 81.92%, although it resulted in a slightly higher Euclidean distance. The Decision Tree also performed considerably better than the baseline, but its performance was lower than that of SVM and XGBoost.

B. Cross-Validation and Ablation Study

Table 4 presents the results of five-fold cross-validation. The results show that SVM provides consistent performance across the different data splits, with the lowest variation in Hit Rate (± 0.012).

Table 4: 5-fold cross-validation results (mean $\pm$ SD)

Model	Hit Rate	Euclidean Distance	Profile Correlation
Decision Tree	0.375 $\pm$ 0.019	0.74 $\pm$ 0.01	0.917 $\pm$ 0.003
SVM	0.838 $\pm$ 0.012	0.11 $\pm$ 0.00	0.999 $\pm$ 0.000
XGBoost	0.711 $\pm$ 0.021	0.24 $\pm$ 0.00	0.989 $\pm$ 0.001

Table 5 examines the contribution of demographic features compared with item-level features. The results show that removing demographic information increased the Hit Rate by approximately three percentage points. Among the three configurations, the psychometric-only model achieved the highest performance. This suggests that the RIASEC-related information in the questionnaire provides the main predictive signal, while adding detailed item-level or demographic information does not necessarily improve the recommendations.

Table 5: Ablation study — Hit Rate by feature configuration

Configuration	Hit Rate
Full (48 items + demographics)	0.4783
Without demographics	0.4924
Psychometric only (6-dimension average)	0.7758

C. Recommendation Quality and Runtime Performance

A manual evaluation of 100 randomly selected test cases showed that 87% of the top 5 recommendation sets contained at least three occupations that were considered relevant during expert validation. The top recommendation had an average correlation of 0.91 ± 0.07 . In a simulated user study, the explanations received an average rating of 4.2 out of 5.0 for clarity and usefulness in taking action.

The system also showed low computational overhead on consumer hardware. Training all three models took approximately 12 minutes, while XGBoost required about 42 ms to generate a prediction. The complete process, from submitting the questionnaire to displaying the results, took less than 100 ms.

VIII. DISCUSSION

A. Why ML Outperforms Rule-Based Methods

The 136% improvement achieved by SVM over the Profile Method can be attributed to two main factors. First, individual Likert-scale responses may contain some measurement error. By considering all 48 items together, the ML models can learn patterns that reduce the effect of this noise, whereas the Profile

Method mainly depends on simple averaging. Second, although the RIASEC dimensions are theoretically treated as separate dimensions, responses in practice can show relationships between them. For instance, Investigative and Realistic interests may be related in some STEM-oriented careers. ML models can learn these interactions from the data, whereas a simple correlation-based baseline may not capture them effectively.

B. Interpretation of Near-Perfect SVM Performance and the Demographics Paradox

The very high-profile correlation obtained by SVM (0.999) can be explained by the nature of the prediction task. Since each RIASEC dimension is calculated as the average of eight related items, predicting the six-dimension scores from all 48 responses are almost a direct reconstruction of those averages. For this reason, Hit Rate provides a more useful distinction between the models, as it depends on the relative ordering of the RIASEC dimensions and is more affected by small prediction errors.

The observation that including demographic information slightly reduced performance differs from some earlier findings [6]. This result could be related to several factors, including the possibility that the RIASEC responses already contain information indirectly associated with age and gender, sampling characteristics within the Open Psychometrics respondent pool, or changes in the strength of traditional gender-related career stereotypes.

C. Explainability Trade-offs

The dimension-overlap explanation strategy prioritizes transparency over technical sophistication. Feature-attribution methods such as LIME [7] or SHAP [13] would provide finer-grained explanation but risk overwhelming users with technical detail that does not match Miller's [8] contrastive, selective criteria for a good explanation. Domain experts consulted during this project validated that highlighting shared RIASEC dimensions aligns with counsellors' natural explanation patterns.

D. Limitations

Holland's RIASEC model was developed in Western contexts, and its cross-cultural structural equivalence is only partially established [12], [19], so validity in collectivist societies where vocational choices may be more family-driven than interest-driven requires further study. Vocational interests can change over time, and the present system provides a single-point-in-time recommendation without longitudinal tracking. The system also models vocational interests but no other decision-relevant factors such as skills, values, economic constraints or geographic preference, and recommendations are restricted to

the 923 ONET occupations, so emerging roles require manual RIASEC profiling before inclusion.

E. Ethical Considerations

If historical career data contains existing societal biases, such as the lower representation of women in engineering, machine learning models may unintentionally reproduce these patterns. Therefore, regular fairness audits, including checks for demographic parity, are important. Automated career recommendations should be used as a support tool rather than as a replacement for human counselling, as individuals may place too much trust in algorithmic suggestions without considering personal factors that the model cannot fully capture. Although the dataset used in this study does not contain personally identifiable information, any practical deployment of the system should ensure that users provide informed consent and that psychometric data is stored and handled securely.

IX. CONCLUSION AND FUTURE WORK

This paper presented an AI-based career guidance system that combines Holland's RIASEC theory with modern machine learning techniques. The results show that supervised learning can improve career recommendations compared with traditional profile-matching methods. Among the models evaluated, SVM achieved a Hit Rate of 81.93%, while the system also retained interpretability by providing explanations based on the overlap between the predicted and recommended RIASEC dimensions. Overall architecture is suitable for use in educational institutions and career centers, which could help make data-driven career guidance more accessible.

There are several areas that can be explored in future work. Additional assessments, such as skills tests, Big Five personality measures, and work-values inventories, could be combined with RIASEC interests to create a more comprehensive user profile. Transformer-based models could also be investigated for generating contextualized item embeddings. Another possible direction is adaptive testing, which could reduce the number of questionnaire items while maintaining prediction accuracy. Longitudinal tracking could be introduced to identify meaningful changes in users' interests over time. The system could also incorporate real-time labor-market information to make career recommendations more relevant to current employment trends. Further work should include a detailed fairness audit across protected attributes and the collection of RIASEC data from non-Western populations. This would help evaluate the model across different cultural settings and, where necessary, adapt it for cross-cultural use, addressing the limitations discussed in Section VIII-D and highlighted by Rounds and Tracey [12] and Thamrin et al. [19].

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Open Psychometrics for providing public access to the RIASEC survey data and the O*NET Resource Centre for maintaining the occupational interest database used in this study. The authors also express their sincere gratitude to the project supervisor for their valuable guidance, support, and constructive feedback throughout the course of this work.

X REFERENCE

- [1] J. L. Holland, *Making Vocational Choices: A Theory of Vocational Personalities and Work Environments*, 3rd ed. Odessa, FL: Psychological Assessment Resources, 1997.
- [2] A. Roy and R. Sharma, "AI-Powered Career Guidance: A Scalable Model for Personalized Recommendations," *IEEE Access*, vol. 14, pp. 1234–1245, 2026.
- [3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [4] S. Roy and M. Shah, "Collaborative Filtering Approaches to Career Recommendation," in *Proc. IEEE Int. Conf. Data Mining Workshops*, 2015, pp. 1234–1239.
- [5] N. D. Almalis et al., "FoDRA — A Framework for Online Data-Driven Academic Career Counseling," *Int. J. Artif. Intell. Tools*, vol. 24, no. 5, 2015.
- [6] S. C. Matz, J. J. Gladstone, and D. Stillwell, "Money Buys Happiness When Spending Fits Our Personality," *Psychological Science*, vol. 27, no. 5, pp. 715–725, 2016.
- [7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust you? Explaining the Predictions of Any Classifier," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 1135–1144.
- [8] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019.
- [9] ONET Resource Center, "ONET Database 29.0," U.S. Department of Labor, 2024. [Online]. Available: <https://www.onetcenter.org>
- [10] Open Psychometrics, "RIASEC Interest Test Dataset," 2024. [Online]. Available: https://openpsychometrics.org/_rawdata/
- [11] G. D. Gottfredson and J. L. Holland, "Using a Theory-Based Occupational Classification in Career Counseling," *The Career Development Quarterly*, vol. 54, no. 1, pp. 57–69, 2005.

- [12] J. Rounds and T. J. Tracey, "Cross-Cultural Structural Equivalence of RIASEC Models and Measures," *J. Counseling Psychology*, vol. 55, no. 3, pp. 310–329, 2008.
- [13] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [14] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [15] O. Troyanskaya et al., "Missing Value Estimation Methods for DNA Microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.
- [16] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [17] A. Alqahtani and H. Al-Dossari, "A Machine Learning Approach to Career Path Choice for Information Technology Graduates," *Engineering, Technology & Applied Science Research*, vol. 10, no. 6, pp. 6589–6596, 2020.
- [18] Q. C. Song, H. J. Shin, C. Tang, A. Hanna, and T. Behrend, "Investigating Machine Learning's Capacity to Enhance the Prediction of Career Choices," *Personnel Psychology*, vol. 77, no. 2, pp. 295–319, 2024.
- [19] T. Thamrin, G. Giatman, and N. Syah, "Reliability and Validity of RIASEC Holland's on Predicting Success Career for Vocational Students," *Jurnal Kependidikan*, vol. 9, no. 3, pp. 740–749, 2023.
- [20] R. G. Medina, "A Fuzzy-Based Career Recommender System for Senior High School Students in K to 12 Education," *IOP Conf. Series: Materials Science and Engineering*, vol. 482, Art. no. 012025, 2019.
- [21] A. O. Ogunde and J. O. Idialu, "A Recommender System for Selecting Potential Industrial Training Organizations," *Engineering Reports*, vol. 1, no. 3, Art. no. e12046, 2019.
- [22] M. Kiran, H. Asim, and M. T. Hassan, "Career and Skills Recommendations Using Data Mining Technique: Matching Right People for Right Profession in Pakistani Context," *VFAST Transactions on Software Engineering*, vol. 7, no. 1, pp. 33–41, 2018.
- [23] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [24] R. Soni, B. Suleiman, and S. Singh, "Improving Graduate Outcomes by Identifying Skills Gaps and Recommending Courses Based on Career Interests," *arXiv:2511.09819*, 2025.
- [25] K. A. S. Sunil, S. Santhosh, S. M. Ashwin, and K. S. Sandeep, "Machine Learning Based Ideal Job Role Fit and Career Recommendation System," in *Proc. 7th Int. Conf. Computing Methodologies and Communication (ICCMC)*, 2023.