Does Conscientiousness Matter? Examining the Influence of Personality Traits on Human-AI Decision Making

EMILIA WIŚNIOS* and MICHAŁ TYROLSKI*, University of Warsaw, Poland
STANISŁAW GIZIŃSKI, University of Warsaw, Poland
HUBERT BANIECKI, University of Warsaw, Poland
PRZEMYSŁAW BIECEK, Warsaw University of Technology, Poland

As the use of machine learning models for decision-making becomes more widespread, it is important to understand how individual differences, particularly personality traits, may impact their effectiveness. This paper presents the results of a study investigating the relationship between personality traits and reliance on machine learning predictions. Specifically, we explored the influence of conscientiousness on individuals' propensity for over-reliance on machine learning suggestions. In addition, we introduce a novel metric of self-reliance, which provides insights into the extent to which individuals unconsciously rely on machine learning models' predictions. Our study makes significant contributions to the field, including a proof of concept analysis of the relationship between personality traits and machine learning reliance.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Machine learning**.

Additional Key Words and Phrases: Explainable AI, Augmented Intelligence, Human Personality

1 INTRODUCTION

Human-AI decision making refers to the process by which individuals work in conjunction with artificial intelligence systems to make decisions in a given domain. This collaborative process involves humans and AI systems sharing and combining their expertise, intending to produce more accurate, efficient, and reliable decision-making outcomes than the human or the AI system could achieve on their own. However, recent studies, among others [6], have revealed incorrect predictions stemming from a model's inability to perform well with data outside its training distribution, as well as the over-reliance on human-AI decision-making. According to that study, participants were more likely to follow AI recommendations when they were supplemented with predicted outcomes, compared to situations with no explanation or feature-based explanations. However, this increased trust in AI led to the phenomenon of over-reliance, especially when the AI recommendation was incorrect. The use of predicted outcomes as explanations also reduced participants' ability to distinguish between correct and incorrect AI recommendations. These findings underscore the importance of carefully considering the type of explanation in the design of human-AI decision-making systems. Human nature, which encompasses demographic characteristics such as age, gender, and education level, as well as psychological traits like personality type, may play a critical role in decision-making. Our study represents the first attempt, to our knowledge, to explore the impact of personality on the human-AI decision-making process.
Our main contributions are:

(1) A proof of concept study analyzing the relationship between human personality traits and reliance on machine learning predictions based on the questionnaire with 8 questions answered by 72 users.
(2) Results show that people with higher conscientiousness are more likely to over-rely on machine learning models.

*Both authors contributed equally to this research.

(3) We introduce the metric of *self-reliance* and discover that participants unconsciously use models' suggestions, even if they say otherwise.

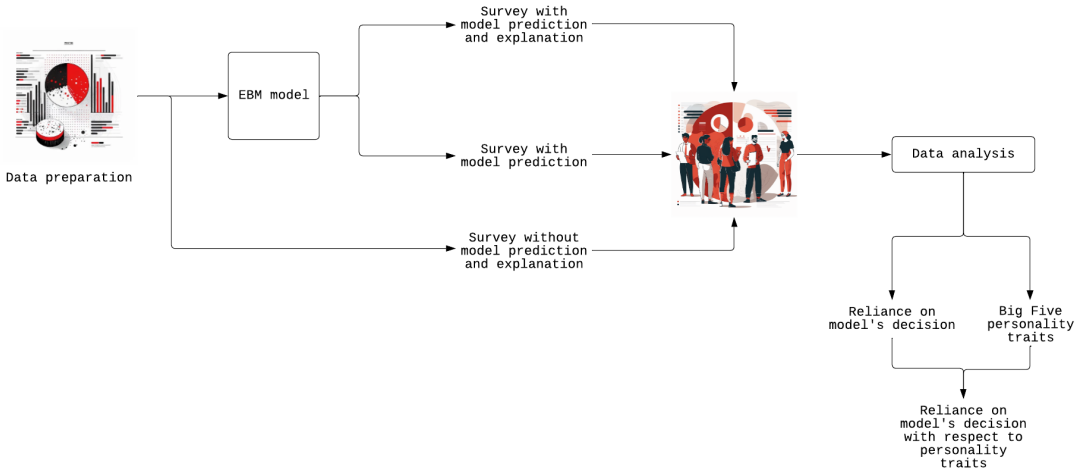

Fig. 1. In this study, we aimed to examine the impact of personality traits on decision-making through a user study. Our findings revealed significant correlations between certain personality traits, namely conscientiousness and openness to experience, and reliance on artificial intelligence (AI)-generated decisions. Additionally, we propose a novel metric called *self-reliance*, which quantifies the degree to which participants rely on their own judgment as opposed to the AI-generated model.

2 RELATED WORK

Several recent studies have examined the effects of AI explanations on human-AI collaboration. Bansan et al. [2] found that explanations increased the likelihood of human acceptance of AI recommendations, regardless of their accuracy. Another study by Liu et al. [7] found that interactive explanations improved perceptions of AI usefulness but may also reinforce human biases and lead to limited performance improvement. Jakubik et al. [6] investigated the effects of providing predicted outcomes as explanations and found that people relied more on AI recommendations when given predicted outcomes, even when they were incorrect.

Additional studies explored the interplay between human and AI expertise levels [13], and how cognitive biases can impact collaborative decision-making [11]. One study found that participants were able to calibrate their reliance on an AI assistant based on its level of expertise relative to their own [12]. Another study by Rastogi et al. [11] developed a mathematical model to understand the impact of cognitive biases on decision-making, with a focus on the anchoring bias. Finally, a study [8] examined how people adjust their reliance on machine learning models when performance feedback is limited, finding that high confidence in human-model agreement affects reliance on the model if people receive no information about its model's performance.

Table 1. Experimental conditions of our study design.

Condition	Explanation
Group 1: Prediction and local explanation	Study participants were provided with features' values, the model's prediction, its confidence, and local explanations produced by the model.
Group 2: Prediction and no explanation	Study participants were provided with features' values, the model's prediction, and its confidence.
Group 3: No prediction and no explanation	Study participants were provided only with features' values.

3 METHODOLOGY

Dataset. In our study, we utilized the Adult Census Income dataset obtained from Kaggle. The data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker [5], and the objective of the prediction task is to determine whether an individual earns an annual income of over $50,000. We selected the following features for our analysis: age, race, gender, native country, marital status, relationship, work class, occupation, years of education, hours of work per week, capital gain, and capital loss. Categorical features were then transformed through one-hot-encoding. Univariate feature selection with mutual information as a metric was performed, resulting in the selection of 30 out of 85 features for the final dataset.

Model. We employed the use of Explainable Boosting Machines (EBMs) [3] from the InterpretML [9] package for model training. EBMs are a variant of gradient boosting algorithms that incorporate feature importance and local interpretability methods to provide explanations for individual predictions. This approach allows for the identification of the most important features in the model, as well as the understanding of how these features influence the model's predictions. EBMs have been shown to achieve comparable or even superior performance compared to traditional gradient-boosting methods. Given the class imbalance in the dataset, we chose to optimize the model tracking the F1 metric. The final model achieved a score of 85.9% accuracy and 68.7% F1, demonstrating an effective balance between precision and recall.

Experimental Design. To perform the study, we carried out a scenario-based online experiment (see Fig. 1). Participants were shown a description of an individual and asked to determine if their yearly income exceeded $50,000. We divided the participants into three groups: a control group and two groups that received AI predictions, with one group also receiving a local explanation from the EBM model (see Fig. 4). The group compositions are listed in Table 1. The features and range of values were introduced at the start of the study, with a detailed description provided in Appendix B. To ensure a balanced distribution of answers, the groups were randomly assigned. However, one group tend to quit the survey before the end, and thus had a higher dropout rate, so only the least represented group was included in the final days of the experiment. All cases were consistent across groups (see Appendix A), but with varying levels of information provided. Additionally, all participants completed common demographic and personality questions, including age, gender, education level, and a shortened version of the Big Five personality test (10-Item Personality Inventory; [10]) obtained from the following website.[1] Moreover, we asked in order participants to score their own machine-learning knowledge to check if this statistically influences results. The study also included a consent form and an introduction to the task.

[1] https://scienceofbehaviorchange.org/measures/10-item-personality-inventory/(last accessed: 30.01.2023)

Subject Recruitment. A total of 74 subjects were recruited for the study. Dissemination of information regarding the study was accomplished through various channels, including students' mailing lists, LinkedIn and Twitter profiles, as well as slack channels of various research groups. Despite efforts to obtain a diverse participant pool, limitations in time and resources resulted in a majority of participants having mathematical and computer science backgrounds. All subjects participated voluntarily and received no remuneration for their involvement. Final participants distribution was 22 for group with prediction and explanation, 21 for group with prediction-only group and 31 for the control group.

4 EXPERIMENTAL RESULTS

4.1 Big Five

Our first series of experiments cover the investigation of the influence of big five personality types on participations' responses. Based on survey answers, we assign to each person values of binary features, namely: agreeableness, conscientiousness, extraversion, openness to experience, and emotional stability. Due to sharing common features, those parts can partially overlap. We calculate the correlation in answers between each of the features and average responses from all parts. Additionally, we set hypothesis

$$H_0 = \textbf{There is no difference between responses of all participants vs group G}$$

$$H_1 = \textbf{There is a significant difference between responses}$$

For each case, we calculated the p-value. We set a threshold of 0.05. Results are shown in Table 2. We see that participants with conscientiousness attributes tend to behave differently than the rest, both for group 1 and group 2. Additionally, participants which are open to experience tend to behave differently in case the model is wrong.

Table 2. Correlations and p-values between average reliance (avg) or average reliance when model wrong (fp/fn) in groups where model prediction was shown. Table legend: AGR - agreeableness, CONSC - conscientiousness, EXTR - extraversion, OP EXP - openness to experience, EWM STAB - emotional stability. P-values < 0.05 are marked on the green, P-values < 0.1 on yellow. Fp/fn indicates that model prediction differs from ground truth.

	AGR		CONSC		EXTR		OPEXP		EMSTAB	
Group	corr	p-val	corr	p-val	corr	p-val	corr	p-val	corr	p-val
1 avg	-0.259	0.244	0.141	0.528	-0.102	0.650	0.002	0.991	0.145	0.518
1 fp/fn	0.139	0.535	0.532	0.010	0.084	0.708	-0.441	0.039	0.025	0.909
2 avg	-0.387	0.091	0.352	0.127	-0.138	0.561	-0.367	0.110	0.115	0.629
2 fp/fn	-0.290	0.2137	0.559	0.010	-0.026	0.912	-0.226	0.337	0.115	0.629

4.2 Reliance and Self-Reliance

Definition 1 (Self-reliance) *Let us consider the number of cases where a study participant stated that the model was useful in decision-making to be represented by the variable n, and the total number of decisions made to be represented by N. The concept of self-reliance can then be defined as* $1 - \frac{n}{N}$ *which represents the proportion of people who relied on their own judgment, as opposed to relying on the model.*

In order to investigate the relationship between average self-reliance and average performance of study participants in Group 1 (which received

Table 3. Average accuracy and self-reliance for group 1

	Accuracy	Self-reliance
FP	0.05	0.45
FN	0.5	0.73
TP	0.91	0.3
TN	0.75	0.39

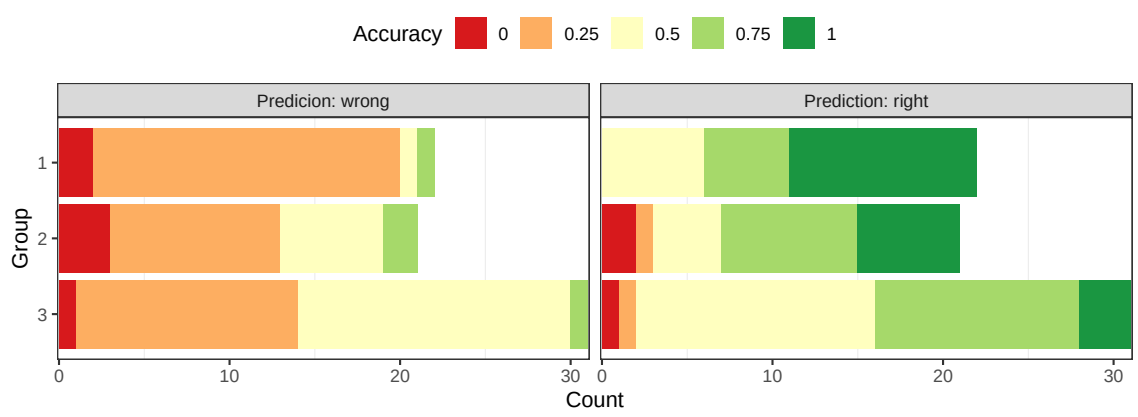

Fig. 2. Average accuracy for each experimental group. For both prediction wrong and right, the scores are calculated for four questions.

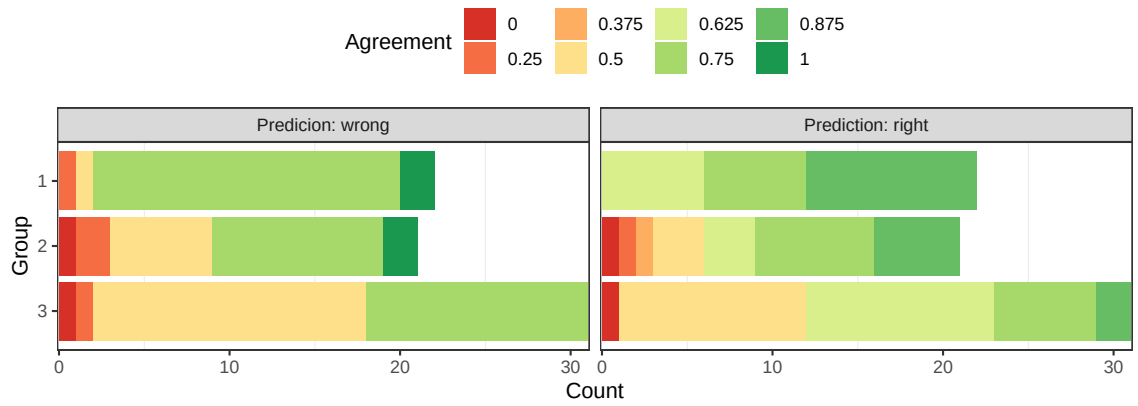

Fig. 3. Average agreement for each experimental group. For both prediction wrong and right, the scores are calculated for eight questions.

model predictions and explanations), we divided the calculations into four separate groups (see Fig. 2, 3), considering that the results may vary significantly in the case of incorrect predictions. Our study showed that the lowest level of self-reliance was associated with the highest accuracy, while the highest accuracy was associated with self-reliance at a level of 45% (see: Table 3).

5 LIMITATIONS AND FUTURE WORK

Dataset. The data utilized in our study is outdated, potentially introducing biases in individuals' decision-making. To address this limitation, we plan to incorporate the latest version of the data set in our future research [4].

Experiment Scale. Our user study obtained 72 responses, which provides preliminary insights. To enhance the reliability of our findings and introduce more variability, future work will focus on increasing the sample size. Enlarging the participant pool would facilitate the exploration of diverse groups, and provide more robust results.

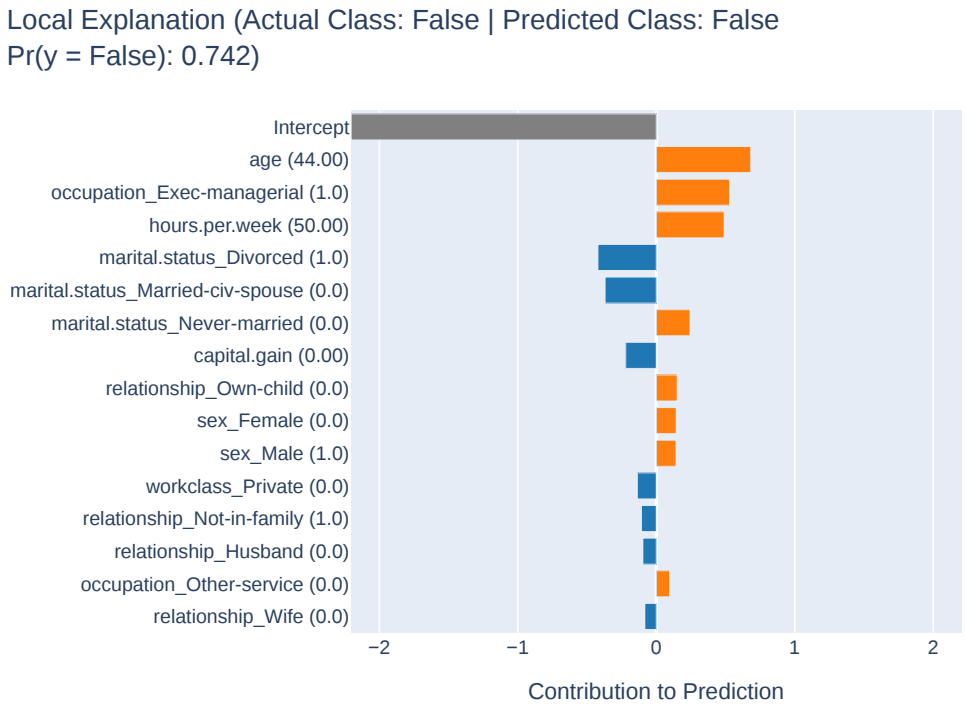

Fig. 4. Local explanation of prediction for person 4. Plots for study participants didn't include actual class information.

Types of Explanations. In this study, static and local explanations of the model's predictions were employed. However, recent studies [1, 7] have demonstrated that utilizing different types of explanations may yield divergent results. To explore this further, future work will incorporate global explanations and enhance the interactive nature of the explanations provided.

Participant Expertise. The study participants in our research were primarily computer science students, who possess some level of expertise in explainable machine learning. However, to examine potential biases, future studies will introduce the research to two distinct participant groups: domain experts and individuals with no prior experience in machine learning. This approach will facilitate the assessment of potential expertise-related biases and provide a more comprehensive understanding of the model's effectiveness across different populations.

6 CONCLUSIONS

Our research sought to examine the impact of personality traits on individuals' reliance on machine learning predictions. Through our study, we made significant contributions to the field, including a proof of concept analysis of the relationship between personality traits and machine learning reliance. Our findings revealed that participants with higher conscientiousness demonstrated a greater propensity for over-reliance on machine learning predictions. Additionally,

we introduced the novel metric of self-reliance, which highlighted the extent to which participants unconsciously relied on the machine learning models' suggestions, even if they denied doing so. Our research demonstrates the need for further analysis of factors influencing susceptibility to explanations and could inform the development of more effective decision-making strategies that account for individual differences in personality.

REFERENCES

[1] Hubert Baniecki, Dariusz Parzych, and Przemyslaw Biecek. 2023. The grammar of interactive explanatory model analysis. *Data Mining and Knowledge Discovery* (2023), 1–37. https://doi.org/10.1007/s10618-023-00924-w

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[3] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613

[4] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 6478–6490. https://proceedings.neurips.cc/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf

[5] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[6] Johannes Jakubik, Jakob Schöffer, Vincent Hoge, Michael Vössing, and Niklas Kühl. 2022. An Empirical Evaluation of Predicted Outcomes as Explanations in Human-AI Decision-Making. https://doi.org/10.48550/ARXIV.2208.04181

[7] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (oct 2021), 1–45. https://doi.org/10.1145/3479552

[8] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 78, 16 pages. https://doi.org/10.1145/3411764.3445562

[9] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *CoRR* abs/1909.09223 (2019), 8 pages. arXiv:1909.09223 http://arxiv.org/abs/1909.09223

[10] Beatrice Rammstedt and Oliver P. John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality* 41, 1 (2007), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

[11] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 83 (apr 2022), 22 pages. https://doi.org/10.1145/3512930

[12] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[13] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 114, 28 pages. https://doi.org/10.1145/3491102.3517791

A DETAILED DESCRIPTION OF CASES FROM THE STUDY

Table 4. Detailed descriptions of cases used in the study with ground truth and model prediction. Additional legend for abbreviations in the table: relationship (rship.), occupation (occ.), years of education (yrs. edu.), confidence of the prediction (conf.), executive managerial (exec. mgmt.).

age	race	gender	country	martial Status	rship.	work class	occ.	yrs. edu.	hrs of work	cap. gain	cap. loss	ground truth	model's pred.	conf.
60	white	male	USA	married civ spouse	husband	other	exec. mgmt.	10	40	0	0	below 50k	above 50k	59.2%
32	other	male	other	married civ spouse	husband	private	prof speciality	14	40	0	0	below 50k	above 50k	57.6%
38	other	male	USA	married civ spouse	husband	private	prof speciality	13	70	0	0	above 50k	below 50k	50.8%
44	white	male	USA	married civ spouse	husband	private	other	10	60	0	0	above 50k	below 50k	52.3%
58	white	male	other	married civ spouse	husband	private	exec. mgmt.	11	40	0	0	above 50k	above 50k	61.4%
21	white	female	USA	never married	not in family	private	exec. mgmt.	10	40	99999	0	above 50k	above 50k	78.7%
59	white	male	USA	divorced	not in family	self emp	exec. mgmt.	9	60	0	0	below 50k	below 50k	67.9%
27	other	male	USA	never married	not in family	private	exec. mgmt.	10	40	0	0	below 50k	below 50k	98.3%

B DETAILED DESCRIPTION OF FEATURES

Table 5. Features and its description provided in the study.

Feature	Range or possible values
Age	17 to 90
Race	White or other
Gender	Female or Male
Native Country	Few most popular like USA, France, etc.
Martial status	Married civilian spouse, Divorced, Never-married, Separated, etc.
Relationship	Represents what this individual is relative to others. For example, an individual could be a husband. Each entry only has one relationship attribute and is somewhat redundant with marital status.
Work class	Private, Self-employed, Other
Occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
Years of Education	1 to 16
Hours of work per week	Average Weekly Hours in the United States averaged 34.40 Hours from 2006 until 2022
Capital gain	
Capital loss	