Artificial Intelligence Approach for Optimization and Refinement of Psychological Analysis

Alfredo Garrachon Ruiz
Illinois Institute of Technology — Politécnica de Madrid
agarrachon@hawk.iit.edu — a.garrachon@alumnos.upm.es

December 1, 2023

Abstract—**This paper explores the application of deep learning and machine learning techniques to enhance psychological testing. Centered on investigating the feasibility of reducing the number of questions in a psychological test without compromising the validity of the results.**

The conclusions of this projects aim to answer the following reflection: *it is preferable to maintain the number of original questions with ideally more valid results or to reduce the number of questions so that the user is more likely to take the test and maintain attention during the test despite theoretically not having full accuracy*.

To answer this, deep learning and machine learning techniques are employed. The methodology includes researching the theory and basis of the most commonly used psychological tests, training different models with data open source data from a psychological test, using different feature selection techniques to identify the most relevant questions, and then training optimized models with a reduced number of features to compare the outcomes.

Index Terms—**Machine Learning, Deep Learning, Feature Selection, Psychological Tests**

I. Introduction

In the current era of technological advancements and growing interest in understanding human psychology, the optimization of psychological assessment methods has become a valuable goal. Personality tests are an essential tool in fields ranging from clinical psychology and personnel selection to scientific research [1] [2]. However, these tests often require respondents to answer a large number of questions, which can be perceived as a barrier to mass participation.

In this context, an intriguing problem arises: *Is it possible to significantly reduce the number of questions in a psychological test without compromising the validity of the results?*. To address this question, this project aims to employ deep learning and machine learning techniques to test it. The methodology consists of three main stages: first, training different machine learning and deep learning models with data collected from a open sourced psychological test. Second, subsequently apply different feature selection techniques to identify the most informative questions. And third, train some of the selected model with different sets of reduced features and compare the results with the original data.

The overarching goal of this research is clear and relevant: to simplify personality tests, making them more accessible and appealing to a wider audience. Reducing questions in tests could have a significant impact on participation and attention span, leading to a better validity of the psychological assessment overall [3], benefiting both psychology professionals and individuals seeking to gain insights into their own personalities. This project represents an interdisciplinary effort that integrates psychology, computer science, and artificial intelligence to tackle a challenge that could help the way we assess and understand the human mind.

a) Related Work: The world of psychology and sociology is a hot topic in the world of artificial intelligence. With the advent of LLMs (Large Language Models) many experiments are being done related to behavior and personality identification. A research team at Stanford is conducting tests on a simulation with several agents personified by LLMs to analyze their behavior [4]. There is also work in which they are trying to analyze the LLMs' own personalities as similar to those of humans, with inconclusive results [5] [6] [7]. * There is also research on training LLMs to be able to analyze the personality of real users, proposing new ways of evaluation and classification in contrast to the more classical approach of this project [8]. While the generalist premise is the same as that of this paper (analyzing personality traits), the approach of this project allows to compare results with historical data by following the same premise and assessing the same traits as the classical tests.

II. Problem Description

A. Statement

Analyzing an individual's personality and evaluating it into more abstract groups has been a hot topic since the beginning of the psychology. It provides valuable insights into human behavior, helping us better understand ourselves and others. By identifying common personality traits and tendencies, we can enhance communication, empathy, and cooperation, ultimately fostering healthier relationships and workplaces. This analysis can also guide personal development efforts, helping individuals leverage their strengths and address weaknesses.

Moreover, it is not only useful in the theory of better understanding the human mind, it also has benefits in practical applications. In fields like psychology and psychiatry, it aids in diagnosing and treating mental health conditions, as certain personality traits are linked to specific disorders [9]. Also, in social and group environments it is a very important aspect. By knowing the theoretical behavior and personality of a person, it is possible to estimate the most optimal social environment for this person. This has clear applications such as at work to match an employee with a work team or in the search for an apartment when having to live with strangers [10].

The current problem with this idea is that it takes a long time to evaluate a person into different psychological traits. These processes are done by tests that have been developed by prestigious psychologists that in general have more than 50 questions that must be answered directly by the person.

To put into context, one of the most famous taxonomy used in this field is the "Big Five Personality Traits". It is made up of 5 personality traits and most approaches require at least 50 questions to analize correctly the user [11]. Other very widespread ones are the recent Minnesota Multiphasic Personality Inventory (MMPI) [12] or the Sixteen Personality Factor Questionnaire (16PF) [13]. The latter, developed by Raymond B. Cattel after years of research, consists of 16 different personalities traits and normally are required at least 160 questions .

This is a high workload for a person, which he/she may refuse to do. Also, if this person is forced to take it, this pressure can lead to alterations in the answers and therefore errors in the psychological classification.

Therefore, the aim of this project is to optimize these tests and find ways to make them simpler. The overall idea is to explore how to reduce the number of questions required to analize a person psychologically without reducing significantly the accuracy of the test.

B. Personality test and data

For this project, the "Big Five Personality Traits" taxonomy [11] has been chosen. It has been developed through the research and development of several psychologists and academics. Citing just a few names, these eminences include Gordon Allport, Raymond Cattell, Warren Norman and Lewis Goldberg among others. It is a widely accepted and comprehensive model used to describe and understand human personality. It classifies personality into five fundamental dimensions or traits, which capture different aspects of an individual's personality:

- **Openness to Experience:** This trait assesses a person's willingness to embrace novelty, creativity, and new experiences. People high in openness tend to be imaginative, curious, open to change, and interested in art and ideas.
- **Conscientiousness:** Conscientiousness measures a person's level of organization, self-discipline, and goal-directed behavior. Those high in conscientiousness are typically reliable, punctual, detail-oriented, and responsible.
- **Extroversion:** Extroversion is about an individual's level of sociability, enthusiasm, and preference for social interaction. People with high extroversion are outgoing, talkative, and thrive in social settings.
- **Agreeableness:** Agreeableness reflects an individual's tendency to be cooperative, kind, and considerate. Highly agreeable individuals are empathetic, compassionate, and seek harmony in relationships.
- **Neuroticism:** This trait gauges emotional stability and the tendency to experience negative emotions. Individuals with high neuroticism may be more prone to anxiety, worry, and mood swings.

In order to classify a user within these 5 personality traits, the ”50-item IPIP version of the Big Five Markers” test has been used [14]. This test is included in the IPIP (International Personality Item Pool) which contains a large number of freely usable and modifiable test types to classify the user within different types of metrics with respect to his or her personality.

These choices have been determined by a variety of factors. The choice of the ”Big Five Personality Traits” is mainly determined by the wide acceptance and use of this personality division by the scientific community. Thus, thanks to its widespread use, there is a huge amount of free data available for experimentation and model training.

The choice of the IPIP test specifically comes for a very similar reason. As this is an open source test, it is possible to find and use the questions that make up this test, as well as databases with the answers of users who have agreed to provide the data for scientific experimentation.

The data used during this project come from the ”Open-Source Psychometrics Project” database [15] which contains a large number of open source datasets with answers to various Psychometrics tests. Among them is the dataset used with the responses of more than 1 million people to the 50-item IPIP version of the Big Five Markers test.

III. Development

A. Data Processing

The previously mentioned database has been used for this project. It contains the answers to the test from a total of 1,015,342 people. The test consists of 50 different questions, so the dimensions of the desired initial dataset is 1,015,342 rows and 50 columns.

Exploring the dataset we find some extra columns besides the test answers themselves. Some examples of these columns are ”timestamp” or ”location”. These are not relevant extra information for the scope of this project, so it has been decided to remove them and keep only the answers to the tests.

The values of the answers to the questions vary in a numerical range from 1 to 5. Where 1 means ”Very Inaccurate” and 5 means ”Very Accurate”. In a first analysis of the database NaN values have been found. The rows with these values have been eliminated since they represented a marginal amount compared to the total data.

This dataset does not incorporate the target variables needed to train the models, which would be the score for each one of the personality traits. To calculate it, the IPIP guide has been followed. The process is based on, for each of the personalities, selecting the questions corresponding to these traits and taking into account their positive or negative rating when calculating the score for each of the traits.

Adding these personality traits as 5 extra columns in the dataset we obtain the target variables. In the Table I we can observe some information about the data that would be useful later on to understand the metrics.

B. Base Models

Once we have the clean data, several models have been trained to compare their performance with each other. On the one hand there are three different the machine learning models and on the other hand a neural network belonging to the field of deep learning. The same metrics have been used for both machine learning and deep learning models in order to compare the results.

a) Machine Learning Models: The models of machine learning are Random Forest Regressor [16], Gradient Boosting Regressor [17] and Support Vector Regression [18] from ”Sklearn” [19] . These have been the best performers in their field. the choice of these models comes from research in the machine learning world on regression models. These 3 models have been chosen for their good reputation in terms of results and for using different internal techniques to predict the final values.

A recursive optimization of the hyperparameters has been performed and the metrics RMSE, MSE, MAE and R^2 have been obtained for each of the target variables as well as for the whole set. The disaggregated metrics are in Appendix A and the global metrics are in Table II and will be used to compare the models.

b) Neural Network Model: This neural network (built from PyTorch [20]) does not need to be too complex since the basic learning algorithm is based on a weighting of different scores.

Normalization of the features on the one hand and of the target variables on the other hand has been carried out. Then the dataset was divided into train and test with the classical 0.8 to 0.2 ratio, passed to

	Mean	Standard Deviation
Extroversion	30.23	3.85
Agreeableness	31.54	4.02
Conscientiousness	31.25	4.27
Neuroticism	30.30	6.75
Openness to Experience	32.64	4.27

TABLE I: Mean and Standard Deviation of the traits

	Random Forest	Gradient Boosting	Support Vector
MSE	1.2825	0.3312	0.4749
RMSE	1.1336	0.5727	0.6806
MAE	0.8037	0.4212	0.3105
R^2	0.9391	0.9848	0.9861

TABLE II: Machine Learning Models Metrics Comparison

tensor form and a batched DataLoader was built for training.

The neural network has as input a dimension of 50 corresponding to the features and an output dimension of 5 corresponding to the target variables. This neural network consists of 4 linear layers with a ReLU activation function which is generally a good choice for this type of architecture because it introduces non-linearity without significant computational cost [21].

With normalization layers embedded after each linear layer we can stabilize training, mitigate some of the issues with ReLU (such as the dying neuron problem) and generally improve performance. At the last linear layer there is no activation function, since in regressions tasks is not necessary as we are predicting continuous values.

The Mean Squared Error Loss has been used as in this type of task gives good results overall and Adam as the optimizer because it incorporates key concepts both from RSMprop (adaptative Learning rate) and from Stochastic Gradient Descent (momentum and gradiant descend) [22]. Also a learning rate scheduler is included to enhance the performance of Adam by adapting the learning rate as training progresses, potentially leading to better convergence.

It should be noted that several tests have been made changing both the architecture (making it deeper or adding dropout layer for example) as optimizers and various parameters, but this design has been the one that gets better results.

After defining all the necessary parameters, training was performed for 50 epochs. This number of epochs was defined after following the evolution of the losses, following the paradigm of Early Stopping. Different metrics such as RMSE, MSE, MAE and R^2 in both training and validation have been performed and plotted. Also the denormalized results (RMSE, MAE and R^2) were computed for a better later on comparison in the true scale of the output.

Observing the results (both Table II and Table III) we can make a first assessment stating that the neural model performs overall better than the machine learning models. It also has been able to adapt well to the algorithm. The original algorithm is not very complex and we have a lot of data, so these results were expected.

Taking the reference data from the initial analysis of the database. We observe that the MAE on the real scale is 0.0688 and the mean value of the trait ”Extroversion” is 30.23 with a standard deviation of 3.85. So the model is wrong on average 0.0688 in a value of 30.23, which is an error of 0.22% that can be considered negligible.

From now on we will focus on the neural model to compare the results using feature selection since this has been the one that has obtained the best results.

C. Feature Selection

As mentioned above, the goal of this project is to reduce the number of questions required to complete a psychology test. The questions in the dataset correspond to the features. That is why we are going to make a selection of the most important features in order to reduce the number of questions required.

First of all, a correlation matrix has been constructed between all the features of the dataset (see Figure 8). In this we can clearly see how each of the 10 questions corresponding to a personality trait are

	Base Neural Normalized	Base Neural Denormalized
MSE	0.0004	0.0004
RMSE	0.0251	0.0917
MAE	0.0183	0.0688
R^2	0.9998	0.9994

TABLE III: Base Neural Metrics Comparison

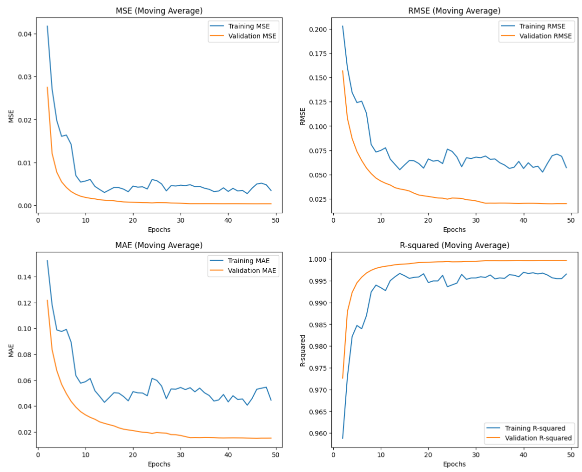

Fig. 1: Base Neural Normalized

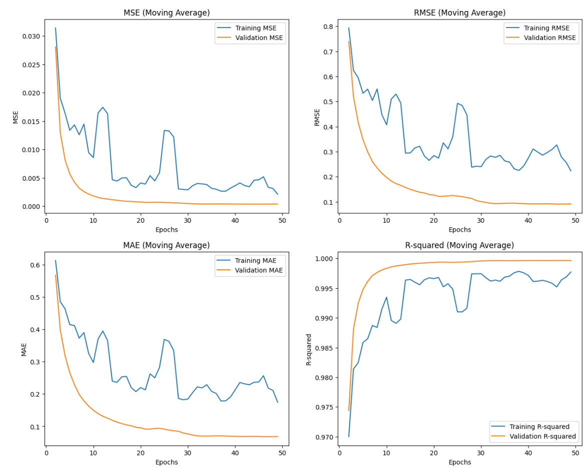

Fig. 2: Base Neural Denormalized

highly correlated between them. Thus, we can also observe how there are some relationships between questions oriented to different traits and how some questions of the same trait are not so closely related. This gives us the intuition that there is indeed room to work on to reduce the number of questions without compromising the quality of the results.

Based on this correlation matrix, all the correlations between each of the features have been run through in an effort to eliminate features that are highly correlated with each other. With a correlation threshold of 0.8, no features have been eliminated, so we can conclude that there is not a very high correlation between any of the features.

With that in mind, three different feature selection techniques have been used. Each of them is independent of the others. At the end of each experiment we will observe the features common to each of the 3 techniques and these will be the ones chosen in the final dataset.

a) SelectKBest: SelectKBest is a feature selection tool located in the library of scikit-learn [23]. It is particularly beneficial for regression models, focusing on isolating the most influential features from a dataset. It primarily employs the f-regression test, a variant of ANOVA (Analysis of Variance), to determine the linear correlation between each independent feature and the dependent target variable. This process involves computing the F-statistic and corresponding p-values, which collectively assess the statistical significance of each feature. Features with higher F-statistics and lower p-values are considered more relevant, indicating a stronger association with the target variable.

b) Recursive Feature Elimination: Recursive Feature Elimination (RFE) is another tool of scikit-learn [24]. It relies on models that assign importance to each feature, like linear regression or tree-based models. Starting with all features, RFE systematically prunes the least significant ones in each iteration, based on the model's evaluation criteria, until only the desired most impactful features remain.

After a series of experiments a linear model has been chosen instead of a tree-based model, more specifically an Elastic Net Regression model has been used [25]. This decision is due to the computational load involved in recursively training tree-based models with the amount of data belonging to the dataset. In a runtime environment such as Google Collab using even 10 percentage of the dataset data

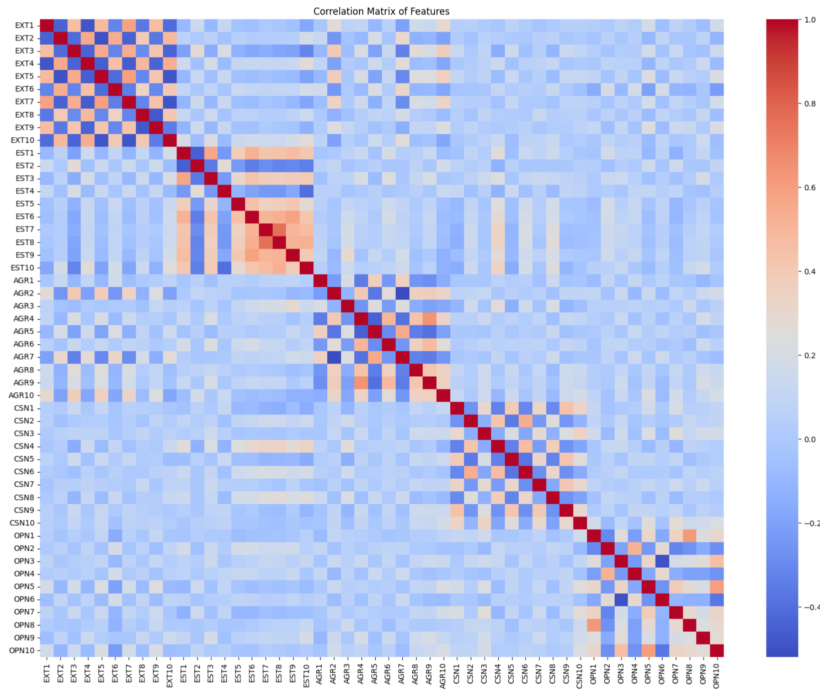

Fig. 3: Correlation Matrix

(100,000 entries), it is not possible to use a tree-based regression model, resulting in a logout for exceeding the time limit.

c) Feature Importance: This technique uses a similar dynamic to the previous one but without using recursive model training. In this scenario it is possible to train a Random Forest Regressor with all the data of the dataset for each of the target variables. The importance of each of the features in the prediction is extracted from the model directly, sorted and filtered to the desired number of features [16].

With the three feature selection techniques defined, we have proceeded to make 3 different feature reductions. The methodology followed is based on choosing the most important desired features for each of the target variables in each of the techniques presented above. Then, an intersection between the results of each technique was performed and the features common to the 3 were obtained.

Thus, in the first selection of features, we have specified to filter the 10 most important features for each of the target variables. Intersecting all the results we obtain 40 common features. In the second selection we have specified to filter 8 features obtaining as a result 32 common features and in the third selection filtering 5 features per target variable we obtain 16 features as a result.

With the features selected for each of the scenarios, 3 different datasets have been elaborated with the subset of features with which 3 different neural networks will be retrained in order to compare results.

D. Optimized Neural Models

As stated earlier, the neural architecture has been chose to compare the results. In order to have a coherent comparison between the optimized models and the base model, we will use the same network architecture as well as optimizers, losses and hyperparameters. Thus, for each of the networks we will show both normalized and denormalized results (RMSE, MAE and R^2).

Thanks to the denormalized results we can have a better understanding of the comparison between models and the loss of accuracy with respect to the original algorithm. Metric graphs and tables summarizing the metrics with which the 3 models of 40, 32 and 16 features have been evaluated can be found in the Appendix B.

IV. COMPARISON

With the metrics obtained from the 4 models we can compare them. Table IV summarizes the results of the denormalized metrics of each model. The decision to present the denormalized metrics is based on being able to have them in the real scale of the target variables so it will be easier to compare them with the original data.

The mean square error gives us an idea of how the error increases as the features are reduced, as expected. Both RMSE and MAE can give us more idea of how the error is with respect to the real data when they are on the same scale.

Taking as an example the trait ”Agreeableness”. Looking at the original data we see that it has a mean of 31.54 and a standard deviation of 4.02. The MAE increases as we reduce the number of features, but even in the extreme case with only 16 features we have a MAE of 1.9513. This value shows that on average the model is wrong by 1.9513 around the original mean of 31.54. This value is less than half the standard deviation, so even in the worst case it still gives good results. Comparing with the other traits gives us a similar idea, in the worst case with the trait ”Extroversion” the MAE is around half of its standard deviation of 3.85.

Focusing o the R^2 score, which indicates the proportion of variance in the dependent variable that is predictable from the independent variables, decreases notably from 0.9994 (50 features) to 0.7236 (16 features). This suggests that as the number of features is reduced, the model becomes significantly less capable of explaining the variability in the target.

V. CONCLUSION

After having reduced the number of questions of the original test from 50 to 16, we have been able to corroborate how the accuracy and overall performance decreases as expected. Even so, by reducing the number of questions to less than a third of the original ones, we still have decent metric values with an average error over the prediction of less than half the original standard deviation. This still gives vericity to the optimized psychology tests and let us extract relevant data from them

The key to this result lies in the following reflection: *it is preferable to maintain the number of original questions with ideally more valid results or to reduce the number of questions so that the user is more likely to take the test and maintain attention during the test despite theoretically not having full accuracy.*

In a test with many questions, the user tends to lose attention as he/she progresses through the test, so the results in this case lose validity despite being the original test developed by psychologists. Thus, a test with reduced questions may give more valid results in real situations even though it theoretically has a worse final performance.

This reflection depends largely on the situation in which the test is being taken and on the person taking the test. A person who takes the test of his own free will in pursuit of a goal will maintain better attention and obtain more valid results in the original test. On the other hand, a person who is being forced to take the test will tend to lose attention much sooner, giving sincere and thoughtful answers only to the first few questions. In this case, taking a 16-question test may give better results since it preserves much of the validity of the original.

This project is intended for easy extrapolation to other psychology tests and free use by anyone interested. That is why the code, resources and results have been published as open source in its corresponding GitHub repository [26].

VI. NEXT WORK

During the course of this project there have been some technical limitations in terms of resources available to test new hypotheses and strategies. These may include training a model based on the transformer architecture and observing whether it performs better against the selected neural network.

Thus, a more ambitious idea would be to test this hypothesis on large language models. These offer a much more advanced understanding of the language. Allowing not only to reduce the questions based on the importance of their answers, but also to understand the meaning of the questions and to be able to merge several questions into one that encompasses common meaning.

	50-features	40-features	32-features	16-features
MSE	0.0004	0.0972	0.1925	0.3448
RMSE	0.0917	1.3243	1.8839	2.5445
MAE	0.0688	0.9981	1.4384	1.9513
$\mathbf{R^2}$	0.9994	0.9247	0.8432	0.7236

TABLE IV: 16-features Neural Metrics Comparison

References

[1] A. Pinon, "Why most psychologists should assess and report personality," *Frontiers in Psychology*, vol. 10, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01982

[2] W. Camara, J. Nathan, and A. Puente, "Psychological test usage: Implications in professional psychology," *Professional Psychology: Research and Practice*, vol. 31, pp. 141–154, 04 2000.

[3] E. Nasiri, M. Khalilzad, Z. Hakimzadeh, A. Isari, S. Faryabi-Yousefabad, S. Sadigh-Eteghad, and A. Naseri, "A comprehensive review of attention tests: can we assess what we exactly do not understand?" *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, vol. 59, 02 2023.

[4] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," 2023.

[5] G. Serapio-García, M. Safdari, C. Crepy, L. Sun, S. Fitz, P. Romero, M. Abdulhai, A. Faust, and M. Matarić, "Personality traits in large language models," 2023.

[6] J. tse Huang, W. Wang, E. J. Li, M. H. Lam, S. Ren, Y. Yuan, W. Jiao, Z. Tu, and M. R. Lyu, "Who is chatgpt? benchmarking llms' psychological portrayal using psychobench," 2023.

[7] A. Gupta, X. Song, and G. Anumanchipalli, "Investigating the applicability of self-assessment tests for personality measurement of large language models," 2023.

[8] H. Rao, C. Leung, and C. Miao, "Can chatgpt assess human personalities? a general evaluation framework," 2023.

[9] B. Lubin, R. M. Larsen, and J. D. Matarazzo, "Patterns of psychological test usage in the united states: 1935–1982." *American Psychologist*, vol. 39, pp. 451–454, 1984. [Online]. Available: https://api.semanticscholar.org/CorpusID:145228998

[10] M. Shekhawat, S. Deshmukh, G. Monroy, A. Tiwari, X. He, H. Shin, Y. Hong, and H. Lu, "Usability test of personality type within a roommate matching website: A case study," *Journal of International Technology and Information Management*, vol. 25, no. 1, 2016.

[11] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annual Review of Psychology*, vol. 41, no. 1, pp. 417–440, 1990. [Online]. Available: https://doi.org/10.1146/annurev.ps.41.020190.002221

[12] A. L. Barbian-Shimberg, *Minnesota Multiphasic Personality Inventory (MMPI)*. John Wiley Sons, Ltd, 2020, pp. 293–308. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118970843.ch121

[13] H. E. P. Cattell and A. D. Mead, *The Sixteen Personality Factor Questionnaire (16PF).*, ser. The SAGE handbook of personality theory and assessment, Vol 2: Personality measurement and testing. Thousand Oaks, CA, US: Sage Publications, Inc, 2008, pp. 135–159. [Online]. Available: https://doi.org/10.4135/9781849200479.n7

[14] "Administering ipip measures, with a 50-item sample questionnaire," https://ipip.ori.org/new_ipip-50-item-scale.htm.

[15] "Open-source psychometrics," https://openpsychometrics.org/_rawdata/.

[16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] J. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[18] Y. Wang, X. Su, X. Wu, Y. Zhang, R. He, X. Li, E. Han, and D. Tao, "Support vector regression: A comprehensive review from traditional svr to deep learning-based approaches," *Information Sciences*, vol. 523, pp. 237–262, 2020.

[19] scikit-learn contributors. (Year of Access) scikit-

learn: Machine learning in python. Accessed on Date of Access. [Online]. Available: https://scikit-learn.org/stable/
[20] PyTorch contributors. (Year of Access for PyTorch) Pytorch: An open source machine learning framework that accelerates the path from research prototyping to production deployment. Accessed on Date of Access for PyTorch. [Online]. Available: https://pytorch.org/
[21] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML'10)*, 2010.
[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[23] scikit-learn contributors. (2023) scikit-learn: Selectkbest. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html
[24] ——. (2023) scikit-learn: Recursive feature elimination. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html
[25] H. Zou and T. Hastie, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.
[26] A. Garrachon, "Psychological analysis optimization," https://github.com/Garrachonr/PsychologicalAnalysisOptimization, 2023.

Appendix

Appendix A: Result Metrics Machine Learning Models

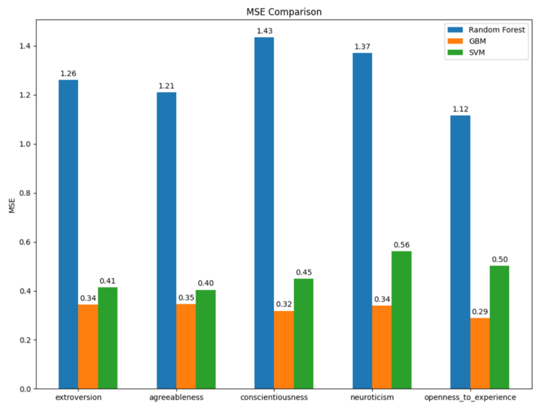

Fig. 4: Mean Square Error Machine Learning Models

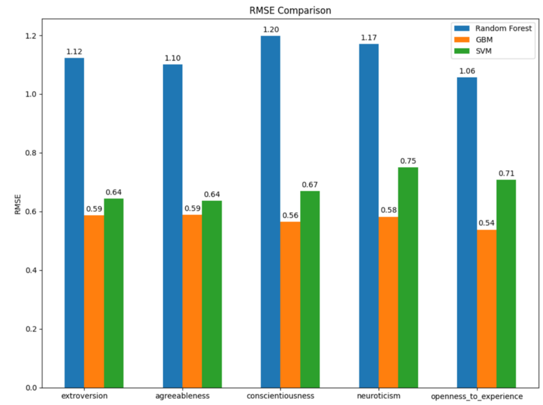

Fig. 5: Root Mean Square Error Machine Learning Models

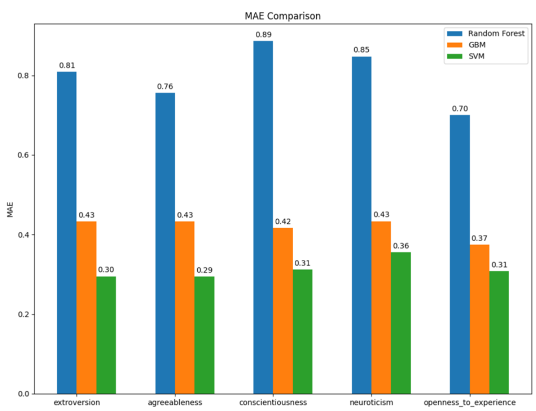

Fig. 6: Mean Absolute Error Machine Learning Models

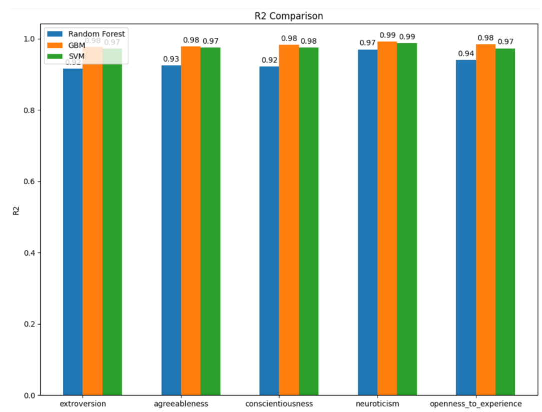

Fig. 7: R^2 Machine Learning Models

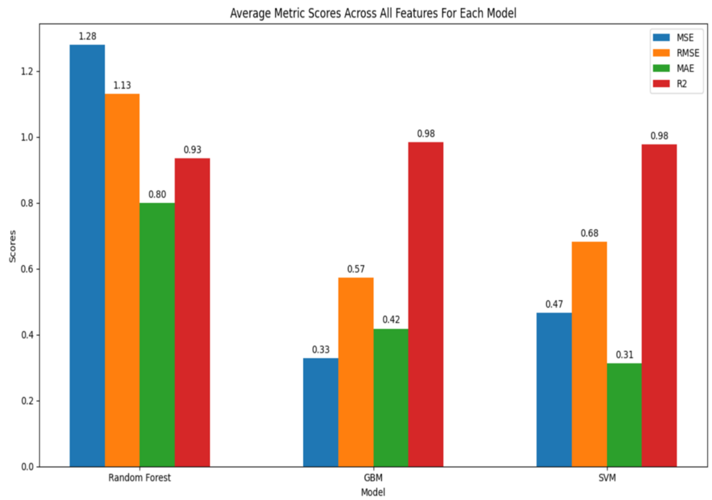

Fig. 8: Average Metrics Machine Learning Models

Appendix B: Result Metrics Optimized Models

	40-features Neural Normalized	40-features Neural Denormalized
MSE	0.0985	0.0972
RMSE	0.3183	1.3243
MAE	0.2324	0.9981
$\mathbf{R^2}$	0.8973	0.9247

TABLE V: 40-features Neural Metrics Comparison

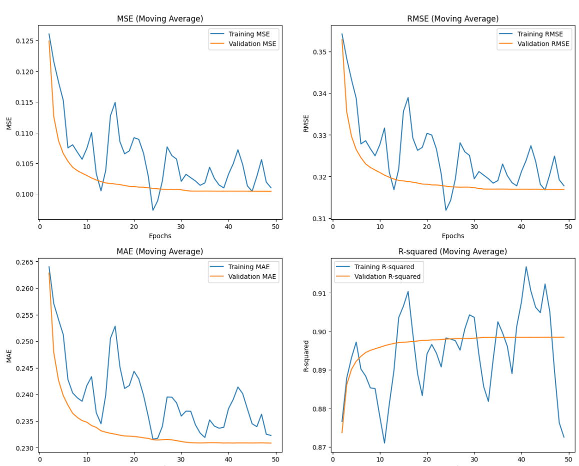

Fig. 9: 40-features Neural Normalized

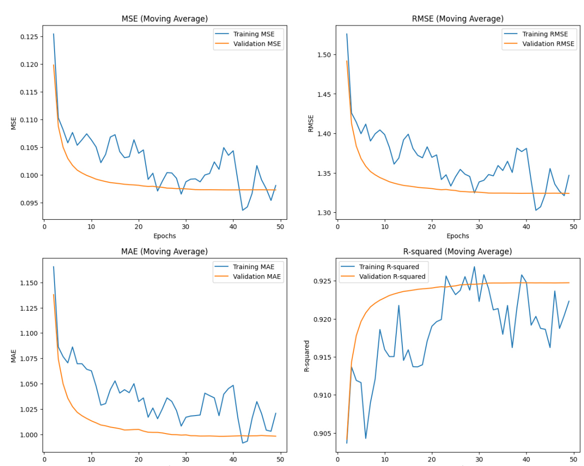

Fig. 10: 40-features Neural Denormalized

	32-features Neural Normalized	32-features Neural Denormalized
MSE	0.1437	0.1925
RMSE	0.3802	1.8839
MAE	0.2826	1.4384
R^2	0.8567	0.8432

TABLE VI: 32-features Neural Metrics Comparison

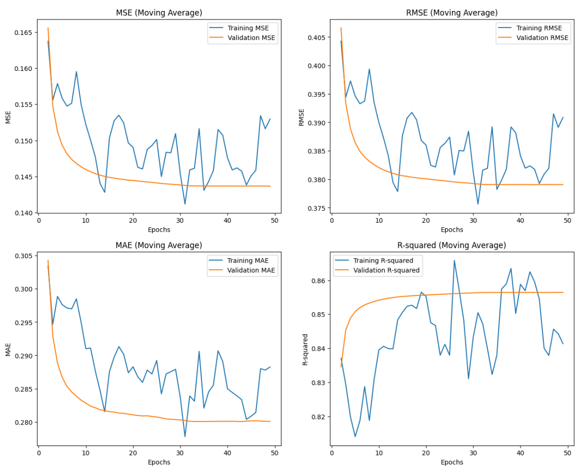

Fig. 11: 32-features Neural Normalized

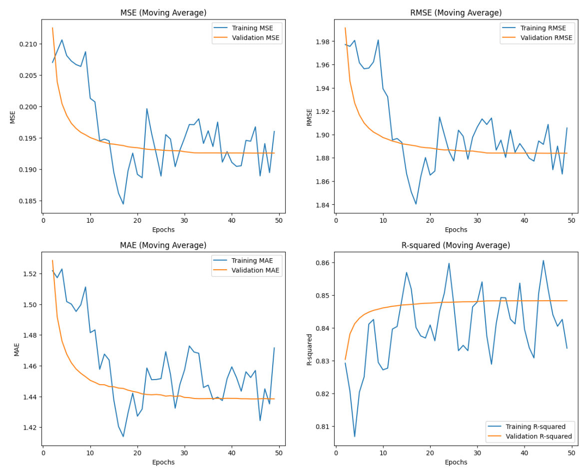

Fig. 12: 32-features Neural Denormalized

	16-features Neural Normalized	16-features Neural Denormalized
MSE	0.3429	0.3448
RMSE	0.5855	2.5445
MAE	0.4405	1.9513
R^2	0.6566	0.7236

TABLE VII: 16-features Neural Metrics Comparison

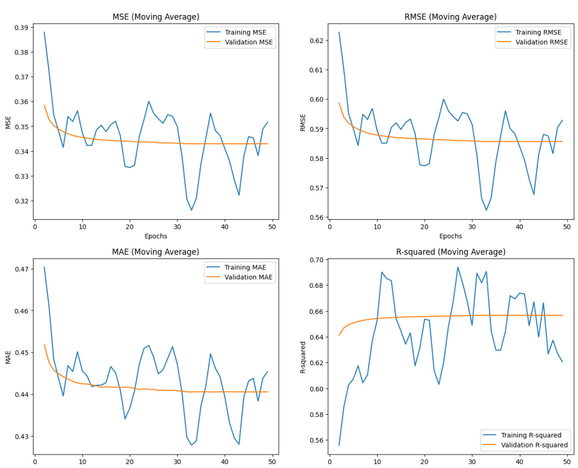

Fig. 13: 16-features Neural Normalized

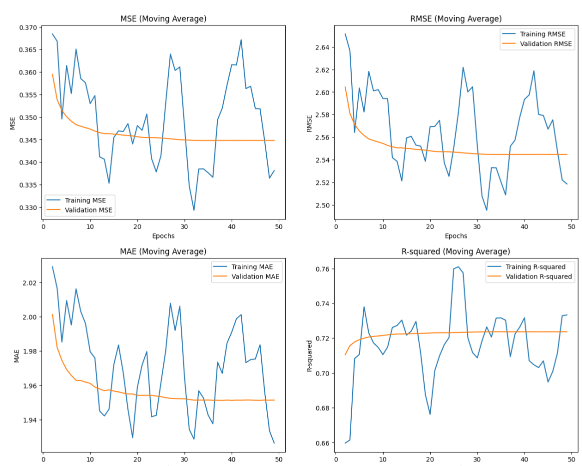

Fig. 14: 16-features Neural Denormalized