International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/IJSRCSEIT

Optimizing Mental Health Detection in Indian Armed Forces Personnel through Feature Engineering Driven Dataset Reduction, Addressing Suicide, Depression, and Stress

Sudipto Roy[1], Jigyasu Dubey[2]

[1]Research Scholar, Department of Computer Science, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India

[2]Head of Department of Computer Science, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India

ARTICLEINFO

Article History:

Accepted: 25 Feb 2024

Published: 07 March 2024

Publication Issue

Volume 10, Issue 2

March-April-2024

Page Number

70-81

ABSTRACT

Within the realm of machine learning, the construction of high-quality datasets stands as a crucial factor profoundly influencing model performance. This research aims to furnish a comprehensive guide for enhancing the accuracy and efficiency of dataset construction. It achieves this by integrating multi-variate reduction techniques and innovative feature engineering strategies, implemented within the Python programming ecosystem. As the landscape of datasets becomes increasingly diverse and complex, the imperative to optimize precision grows more critical. This study explores the judicious application of dimensionality reduction methods, such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), alongside various feature selection approaches to strategically streamline datasets while preserving vital information. In conjunction with these reduction techniques, the research introduces novel feature engineering methods to amplify the discriminative power of remaining features, thereby enriching the dataset's representational capacity. The exploration spans a spectrum of multi-variate reduction techniques and delves into feature engineering methodologies, including polynomial feature creation, interaction term generation, and domain-specific transformation functions. Practical implementations of these techniques are demonstrated through Python, showcasing their applicability across diverse domains. Empirical evaluations on real-world datasets underscore the efficacy of the proposed methodology, revealing superior accuracy and efficiency compared to conventional dataset construction approaches. The insights derived from this research contribute significantly to the broader discourse in machine learning, presenting a generic yet potent framework for enhancing precision in datasets. Beyond deepening our understanding of multi-variate reduction and feature

engineering, the findings offer a practical guide for researchers and practitioners seeking to optimize precision in various machine learning applications.

Keywords : Machine Learning, Psychometric Test, Feature Engineering, Exploratory Data Analysis, Dimensionality Reduction, Principal Components Analysis.

I. INTRODUCTION

A standard psychological test serves as a qualitative or quantitative measure of a personal trait in individuals, administered through questionnaires or verbal interaction. The most fundamental aspect in the development of any psychological evaluation is Psychometric Test Validation, which aims to understand the differences in individual mannerisms. It not only defines underlying concepts but also explores possible use cases and applications of psychological tests [1]. The influence of Machine Learning and associated feature engineering in psychological assessment is evident, notably highlighted in the case of Cambridge Analytica during the 2016 American presidential election. Researchers at the Psychometric Centre of the University of Cambridge collected social media data from over 50,000 Facebook users to predict their online personalities, primarily using statistical tools with minimal integration of machine learning. While this manual analysis revealed certain traits within the population, the full potential of the available data remained untapped due to the limitations of manual statistical tools. The leaked data was subsequently utilized by Cambridge Analytica, employing various feature engineering techniques on the dataset to create a custom-made campaign targeting specific population profiles.

The application of Machine Learning-based feature engineering extends beyond academic boundaries, gradually transforming into complementary analytical tools alongside statistical approaches. In various fields, Machine Learning-based feature engineering is being implemented to solve day-to-day issues, such as classifying psychiatric disorders through imaging data analysis, prescribing medicines for clinical use, forensic sciences, and genetic engineering. However, the extensive use of Machine Learning-based feature engineering in clinical psychological experiments remains undocumented. The scope for such utilization becomes apparent when considering the similarities between modeling cognitive and brain functions in the mathematical domain and neural network-based Machine Learning models derived from cognitive psychology. While data analysis in psychological science traditionally relies on inferential statistical tools, their limitations in replication within analytical tools designed for behavioral research create a significant opportunity for implementing more efficient feature engineering models in experimental psychological data analysis [2].

Machine Learning-based feature engineering offers the advantage of replicating designed models on new, unseen data. Despite being forerunners in theory building, the adoption of Machine Learning in the field of psychology has been slow, with analyses relying on age-old methods such as probability-based statistical tools and size effect measures. This paper refrains from delving into the current advancements in cognitive process modeling using Deep Neural Networks in Machine Learning. Instead, it introduces a novel model that emphasizes the benefits of utilizing Machine Learning-based feature engineering over previously collated data on psychological experiments, in comparison to traditional statistical tool analysis. Our

focus will primarily center on two objectives: replicating models for use on fresh data and achieving single-subject level model predictions [3].

II. Characteristics of the survey instrument

In this section, we deal with the various mental disorder or psychological disorder, as one may say it, that we intend to find out using a machine learning algorithm. Since this is the first time in the recorded history of psychological disorder test Construction, using machine learning it is important to mention that such kind of test construction has been limited to the number of questions as is possible to reduce manually. The psychologist in the previous era of non-machine learning times had to limit the number of questions they asked to a maximum that was feasible for manual deduction using the known statistical methods. In this paper, I have tried to invent a framework into place which will use the machine learning algorithms to reduce the number of questions in a questionnaire to have the minimum number of questions to get to the desired result. Though it is important to mention that it is a laborious task to manually classify the data into the various psychological disorder categories to help our machine learning algorithm to be able to reduce the questionnaire to a minimum level based on machine learning methodologies which is again based on different statistical deductions [4]. However, we had to understand to set up the initial data set and to ensure that we are not missing on to any of the important questions that may lead us to deducting the mental state of our individual, we have resorted to manual classification. This is the only time that manual classification has been done based on the response of the test subjects. Thereafter the statistical analysis was carried out using machine learning algorithms.

Through our literature review we have come to understand that stress depression and suicide cannot be a reason that can originate from any single physical or mental symptoms. The mental state or the psychological state of a human being to run into stress depression or suicide thoughts is cumulative of several different factors. Such factors can arise out of not only the physical environment in which individual is working, it is also cumulative product of his family relationship, his sleep patterns, his drinking habits and so on and so forth. There may be several reasons for a person to have Suicidal Thoughts [5]. Such kind of Suicidal Thoughts may arise out of different environmental mental or working condition of an individual. It is it is well understood that stress leads to depression and further to suicide. In this research we are trying to find suicidal thoughts in an individual who has gone into stress there after leading to depression and finally having suicide thoughts. In this research work we have limited our scope to stress, depression and suicide thoughts and are research is based on the same line of thought. Any other reason that leads to suicidal thoughts into a person has not been made part of this research and is out of the scope.

Questionnaire that we have built covers or is able to measure an individual based on demographic assessment, organizational environment assessment, anxiety assessment, standard of living assessment, marital harmony assessment, post deployment and organizational trauma dependency on alcohol and insomnia. Since I am using machine learning algorithms it gives me an opportunity to be able to deal in several different mental health issues and does allowing me to be able to conclude about this stress level of an individual and whether such stress level is leading to depression or suicide ideations. There are total of 301 questions total of 8 different domains. No psychological tool developed till date has been able to consider such a diverse field of questionnaire due to the limitation of the human mind to comprehend such many questions at one point of time [6]. I have tried to put forward a framework which will be useful for reduction of questionnaire into only relevant aspects, for any psychological tool development, with the help of Artificial Intelligence and machine learning algorithms.

III. Important Feature Selection for the Dataset

In data analysis, it is occasionally possible to characterize an item using a subset of the numerous features without losing any information. One of the most important steps in data analysis is finding these feature subsets, also known as feature selection, variable selection, or feature subset selection. An overview of the primary feature selection techniques used by us using machine learning (ML) is given in this section. Feature selection is a crucial step in the process of building machine learning models, and it offers several benefits:

Faster Training and Inference.

Working with a reduced set of features accelerates both the training and inference phases of a machine learning model. The computational resources required decrease, making the model more efficient and scalable [7].

Enhanced Model Interpretability.

Models with fewer features are often more interpretable. It becomes easier to understand the relationships between input features and the model's predictions, making it more straightforward for users to trust, explain, and act upon the model's decisions [8].

Reduced Overfitting.

Feature selection helps mitigate the risk of overfitting, where a model performs well on the training data but fails to generalize to new, unseen data. By removing irrelevant features [9], the model is less likely to capture noise in the training data.

Simpler Model.

Selecting a subset of relevant features contributes to a simpler model. Simplicity is often desirable as it reduces the risk of model complexity and the potential for overfitting. Simpler models are easier to maintain, understand, and deploy.

Addressing the Curse of Dimensionality.

The curse of dimensionality refers to challenges that arise when working with high-dimensional data. Feature selection helps mitigate these challenges by reducing the number of features, making the data more manageable and improving the model's ability to generalize [10].

Better Handling of Multicollinearity.

Feature selection has helped address multicollinearity, a situation where features are highly correlated. Including highly correlated features introduces instability in the model, and selecting a subset of these features has improved stability and interpretability [11].

Facilitates Model Deployment.

Models with a reduced set of features are often more practical for deployment in real-world scenarios. They require fewer resources and are more likely to run efficiently on various platforms.

Easier Debugging and Maintenance.

Working with a smaller set of features simplifies the debugging and maintenance processes. Identifying issues, troubleshooting, and updating the model become more straightforward when dealing with fewer input variables.

Facilitates Feature Engineering.

Feature selection has been guiding the feature engineering process by highlighting the most influential features. This helps data scientists and

domain experts focus their efforts on enhancing and transforming the most relevant aspects of the data [12].

Development of a Psychometric Instrument through the Implementation of Feature Engineering Techniques

By carefully applying feature engineering techniques, we have enhanced the utility and interpretability of machine learning models in the context of psychological test construction. This has led to more accurate predictions and a better understanding of psychological phenomena based on the information collected through interview. The total attributes come to 301. Our first focus was to conduct a survey on the 301 attributes. Then we got the samples collected from 440 individuals classified into different mental health state of stress level, depression, or suicidal ideations by the team of experts. The dataset that we are dealing with is discrete, numerical, and cardinal for the input features. In the context of a questionnaire, input features have been categorized into numerical, discrete, and cardinal types [13]. Numerical features are those that represent measurable quantities and can take on a range of numerical values. These features are continuous and can take any numerical value within a certain range. Discrete features take on distinct, separate values and typically represent counts or categories. Discrete features can only take on specific, separate values and often represent counts or categories. Cardinal features are like numerical features but have a meaningful zero point, indicating an absence of the quantity being measured. Cardinal features have a true zero point, distinguishing them from numerical features [14].

DISORDER PREDICTION	NO OF ATTRIBUTES
Demographic and Organizational Environmental Assessment	62
Occupational Stress Assessment	21
Living Standard Assessment	17
Work Related Anxiety Assessment	30
Marital Discord Assessment	16
Post Trauma Assessment	25
Alcohol dependency assessment	12
Insomnia Assessment	18
Individual Depression Assessment	49
Suicide Ideation/ Suicidal Assessment	51

Fig. 1. Questions per individual domains

In the Phase 1 (PH-1) data building process, a commendable total of 440 participants have been included, reflecting a comprehensive and inclusive approach. Emphasizing the ethos of the Indian Army, the dataset encompasses all ranks, symbolizing a collective representation of the diverse personnel within the organization. This inclusive data collection underscores a commitment to understanding the holistic experiences and perspectives of individuals across various roles, fostering a more nuanced analysis and informed decision-making process within the Indian Army. With a total of 440 participants and 301 data points per participant, the dataset comprises an impressive 132,440 data points. This extensive collection of data underscores a thorough and detailed exploration of various aspects, reflecting a commitment to obtaining a comprehensive understanding of the variables involved. Such a substantial dataset holds the potential to yield valuable insights and contribute significantly to any analysis or research endeavor [15]. The data collection methodology employed for gathering information from all ranks in the Indian Army involved a dual approach, combining personal interviews and the use of Google Forms. This strategic combination reflects a commitment to thoroughness and inclusivity, aligning with the disciplined and comprehensive conduct expected within the Indian Army.

Criteria for Choosing Optimal Features

The selection of the best features in a machine learning model involves a trade-off between various factors such as time, memory usage, and accuracy. Each of these considerations plays a crucial role in determining the efficiency and effectiveness of the model. We will elaborate on the significance of each parameter:

Time.

Time efficiency is a critical consideration, especially in scenarios where model training and prediction need to be performed in a timely manner. Faster training and prediction times are advantageous, particularly in real-time applications or when dealing with large datasets [16]. Feature selection methods that are computationally efficient, such as filter methods, may be preferred to minimize the time required for model development. However, more sophisticated methods like recursive feature elimination or wrapper methods might take longer but can potentially yield higher accuracy.

Memory Used.

Memory efficiency is essential, particularly when working with large datasets or in resource-constrained environments. Optimizing memory usage ensures that the model can scale appropriately and can be deployed in environments with limited resources. Certain feature selection methods may require storing intermediate data structures or matrices, leading to increased memory usage [17]. Choosing methods that are memory-efficient, or optimizing the implementation of feature selection algorithms, becomes crucial for scalability.

Accuracy.

Accuracy is a fundamental metric, representing how well the model performs in making predictions. The goal is to select features that contribute meaningfully to the predictive power of the model [18]. Some feature selection methods may sacrifice a small amount of accuracy in favour of computational efficiency or reduced memory usage. It is important to strike a balance between accuracy and resource utilization, selecting a method that provides a good compromise for the specific use case. The selection of the best features involves carefully considering the trade-offs between time, memory usage, and accuracy. The optimal choice of feature selection method depends on the specific requirements of the application, the characteristics of the dataset, and the available computational resources. In our case we have selected the Accuracy as the guiding parameter [19].

Feature Selection Techniques

In my research endeavour, I successfully implemented a feature selection technique using Python, harnessing the capabilities of prominent libraries such as scikit-learn, pandas, and NumPy. The chosen platform for this implementation was PyCharm, offering a robust integrated development environment (IDE) that facilitated seamless coding and testing of the feature selection methodology [20]. The utilization of scikit-learn, a powerful machine learning library, allowed me to efficiently apply various feature selection algorithms. Leveraging its extensive functionalities, I navigated through diverse techniques to identify the most relevant features for my specific context. Pandas, a versatile data manipulation library, played a pivotal role in handling and preprocessing the dataset. Its intuitive data structures facilitated the extraction, cleaning, and transformation of features, contributing to the overall efficacy of the feature selection process. NumPy, known for its numerical computing capabilities, complemented the implementation by enabling efficient handling of arrays and matrices. This was particularly beneficial when dealing with large datasets, ensuring optimized performance and computational efficiency.

For the collaborative and cloud-based aspects of my work, I seamlessly integrated Google Colab into the workflow. This platform, with its accessibility and powerful resource allocation, provided an ideal environment for executing and fine-tuning the feature selection methodology [21]. The collaborative nature of Google Colab allowed for easy sharing and collaboration on the codebase. The culmination of these tools and platforms not only facilitated the successful implementation of the feature selection technique but also streamlined the overall research process. The integration of PyCharm and Google Colab, coupled with the capabilities of scikit-learn, Pandas, and NumPy, reflects a comprehensive and well-thought-out approach to feature selection in the Python programming ecosystem.

During my research, I employed a supervised feature selection technique to enhance the performance of my feature selection. This approach, rooted in the paradigm of supervised learning, involves leveraging labelled data to strategically choose the most relevant features for predictive modelling [22]. By utilizing this technique, I aimed to optimize the feature set by considering the direct impact of each feature on the model's ability to make accurate predictions. The supervision aspect comes into play as the algorithm assesses the correlation between features and the target variable, discerning which attributes contribute the most to the desired outcome. This method not only enhances the model's predictive accuracy but also aids in interpretability by focusing on features that have a discernible influence on the target variable. Leveraging the wealth of information contained in labelled data, I navigated through various supervised feature selection algorithms to identify and retain the most informative attributes. The application of this technique aligns with the broader goal of refining the model's ability to generalize well to new, unseen data while mitigating the risk of overfitting. The chosen approach reflects a strategic fusion of domain knowledge, data labelling, and algorithmic prowess to craft a more nuanced and effective machine learning model for my specific research objectives.

The supervised feature selection technique I employed adds a layer of intelligence to the model building process. It involves the systematic evaluation of features based on their relevance to the target variable, ensuring that the selected attributes contribute meaningfully to the model's understanding of the underlying patterns in the data. One of the key advantages of this approach is its ability to handle complex relationships within the dataset [23]. By considering the labelled nature of the data, the algorithm is equipped to discern intricate dependencies between features and the target variable, leading to a more nuanced and accurate model.

Furthermore, this technique facilitates a balance between feature reduction and model performance. It allows for the identification of a compact set of features that not only preserves the integrity of the predictive task but also simplifies the model, reducing the risk of overfitting and enhancing interpretability [24]. The iterative nature of supervised feature selection, driven by continuous feedback from the model's performance on labelled data, underscores its adaptability. This adaptability ensures that the selected features remain relevant even as the dataset or research context evolves. In summary, the supervised feature selection technique adopted in my research signifies a strategic integration of labelled information to guide the model toward a more precise, interpretable, and generalizable representation of the underlying data patterns. It is a dynamic and informed approach that harnesses the power of supervision to refine the feature set for optimal predictive performance.

Figure No 2 illustrates a comprehensive overview of various feature selection techniques used in supervised machine learning. These methods are essential for improving model accuracy, performance, and interpretability by removing irrelevant, redundant, or noisy data. At the top level, the image categorizes the

techniques into three primary methods: Filter Methods, Embedded Methods, and Wrapper Methods. Filter Methods are generally the first step in feature selection [25]. They include removing constant values that provide no information, quasi-constant values that vary very little and are therefore of little predictive value, and duplicate values which are redundant. Additionally, correlation treatment is applied to identify and handle features that are highly correlated with each other, which can affect the model's performance due to multicollinearity. Techniques such as Pearson, Spearman, and Kendall correlation coefficients are used to measure the linear and monotonic relationships between variables. Embedded Methods integrate feature selection as part of the model training process. Regularization methods like L1 (Lasso) and L2 (Ridge) shrinkage are used to penalize the inclusion of irrelevant features. Furthermore, models such as Random Forest Classifier (RFC), Extra Trees Classifier, Gradient Boost, ADA Boost, and Decision Trees inherently perform feature selection by determining feature importance during model training. Wrapper Methods are a search approach where different combinations of features are tested and compared to select the best subset for model performance. This includes forward feature selection, which starts with no features and adds them one by one, backward feature selection, which starts with all features and removes them one by one, exhaustive feature selection, which evaluates all possible combinations of features, and recursive feature elimination, which recursively removes attributes and builds a model on those attributes that remain. The diagram also shows that within these methods, statistical treatment plays a role, using tests such as ANOVA, Chi-Square, and Fisher Score to evaluate the importance of features [26]. This structured approach to feature selection is pivotal to model optimization, as it leads to simpler, faster, and more effective models that are less prone to overfitting, making them more generalizable to new data.

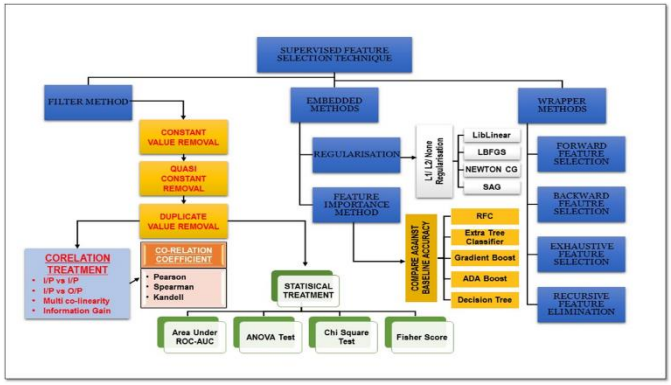

Fig. 2. Supervised Feature Selection Techniques used

Determination of Initial Accuracy Level

In delving into the intricacies of my research methodology, I began by addressing the challenge of an unbalanced dataset, recognizing that the conventional accuracy metric might not be the most reliable indicator of model performance. To establish a baseline, I opted for a straightforward yet effective approach—implementing a Simple Tree-Based Classification with Recursive Feature Elimination (RFE). The dataset, consisting of 440 samples and 301 features, presented a scenario where class imbalances needed careful consideration. Leveraging the simplicity of a tree-based classifier allowed for an initial exploration of the dataset's patterns. The Recursive Feature Elimination technique added a layer of sophistication by iteratively removing less significant features, refining the model's understanding of the underlying relationships. In preparing the data for modelling, I employed Standard Scaling to ensure uniformity in feature scales, a crucial step when dealing with algorithms sensitive to varying magnitudes [27]. The subsequent Train-Test Split segregated the dataset into subsets for training and evaluation, a fundamental practice to assess the model's generalization to new, unseen data. The incorporation of Cross-Validation, coupled with Random Seeding, fortified the robustness of my methodology. Cross-Validation, by repeatedly splitting the dataset into different train-test folds, provided a comprehensive evaluation of the model's performance. The inclusion of Random Seeding ensured the reproducibility of

results, adding a layer of reliability to the experimentation process. Upon executing the methodology, the achieved accuracy of 93.18% underscored the effectiveness of the chosen approach. This metric serves as a benchmark against which future model enhancements and sophisticated algorithms can be measured. Importantly, this baseline accuracy serves as a reference point, allowing me to gauge the incremental improvements that subsequent iterations of my research may bring. This journey through establishing a baseline accuracy not only lays the foundation for my research but also showcases the thoughtful considerations and strategic choices made to navigate the nuances of an unbalanced dataset [28]. In this part I detail a comprehensive feature selection strategy to optimize the detection of mental disorders such as stress, depression, and suicidal tendencies in the Indian Army. The methodology is a fusion of filter, wrapper, and embedded methods, structured to refine the predictive capability of the analytical models employed.

The feature selection strategy began with a full set of 301 potential features. The initial phase employed filter methods, starting with the removal of 23 constant value features, reducing the set to 278. This step was followed by the elimination of 59 quasi-constant features, leading to 219 features. Subsequently, 57 duplicate features were identified and removed, yielding a set of 162 unique features. The second phase involved correlation treatment, where I applied various correlation coefficients, including Pearson, Spearman, and Kandell, to address issues of multicollinearity and information gain, which further refined the feature set to 53 based on their predictive relevance [29].

In the third phase, statistical treatment through ANOVA tests and Fisher Scores was used to determine the statistical significance of each feature, which led to a selection of 147 features that exhibited a strong relationship with the target variables. The final phase was the application of supervised feature selection techniques, encompassing embedded methods such as regularization and feature importance methods to ascertain the contribution of each feature to model performance. This resulted in a narrowed down set of features deemed most significant. From the wrapper methods perspective, I employed Forward Feature Selection, which further reduced the feature space from 301 to 42. These 42 features were initially deemed insufficient by a group of experts; however, the utility of these features was to be empirically validated against the final model accuracy [30].

The models tested included Linear Regression, LibLinear, L-BFGS, Newton CG, and SAG algorithms under the wrapper methods, while Random Forest Classifier (RFC), Extra Tree Classifier, Gradient Boosting, ADA Boost, and Decision Tree were applied within the exhaustive feature selection framework. Each of these models was assessed for their performance, with a subset showing a significant reduction in feature space without compromising the accuracy. The RFC, Extra Tree Classifier, Gradient Boosting, ADA Boost, and Recursive Feature Elimination (RFE) methods all converged on a final selection of 59 features that balanced model complexity and predictive accuracy [31].

In the latest part, I address the critical task of reducing feature space complexity to improve the detection of mental disorders such as stress, depression, and suicidal tendencies within the Indian Army [32]. The feature selection process began with an extensive set of 301 features, which was meticulously narrowed down to 42 features through various statistical and machine learning techniques. A key element of this reduction was the identification of a subset of 22 features within the 42, which sparked a debate among experts regarding its adequacy for maintaining model accuracy. Despite reservations about the limited number of features, these 22 were selected based on their strong predictive potential and relevance to the

psychological profiles of interest [33]. The consensus among the expert panel was that a minimalistic approach might oversimplify the complexity of mental health diagnostics. Nonetheless, it was decided that the utility of these 22 features should not be dismissed without empirical validation. Hence, the next step in my research was to rigorously test the 22-feature model against the final accuracy measurements to ascertain whether this reduced feature set could still reliably predict mental health disorders. This investigation is significant because it directly impacts the operability of mental health diagnostics in field conditions [34]. A model with fewer features is not only computationally more efficient but also easier to deploy and interpret by military health practitioners. If the 22-feature subset proves to be as accurate as the broader set, it could revolutionize the way mental health support is provided in the Indian Army, enabling quicker, more efficient, and potentially life-saving interventions for service personnel at risk of stress, depression, or suicidal behavior [35].

The paper culminates with a summary of the methodological rigor applied in feature selection and the implications of using a highly reduced feature set in practical, real-world scenarios. It underscores the delicate balance between model simplicity and diagnostic accuracy in the context of mental health surveillance in the military. This set can now be utilized to build a predictive model that is both accurate and computationally efficient, offering significant contributions to the field of military mental health [36]. The feature selection strategy employed here is not only methodologically sound but also sensitive to the operational constraints of field deployment, making it a viable approach for real-world applications.

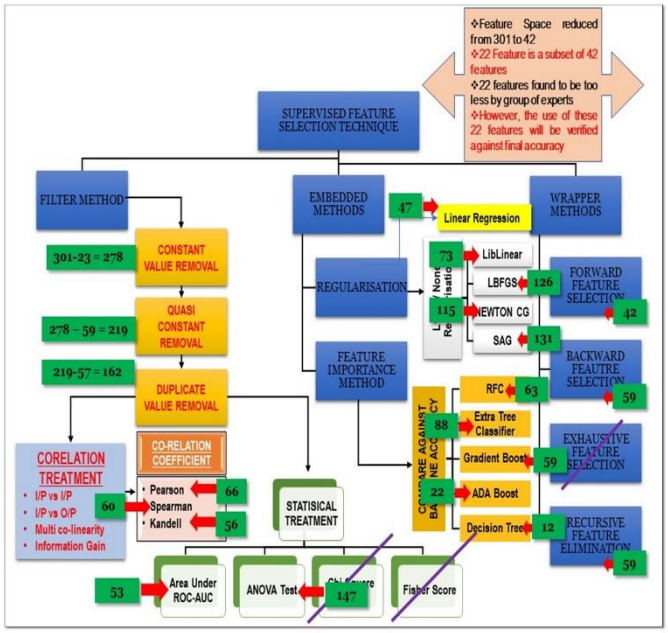

Fig. 3. Stage of Employment of Feature Engineering in Automation

IV.Conclusion

In my paper, I present a novel approach to the detection of mental disorders within the Indian Army by employing a data-driven methodology to reduce the complexity of disorder prediction. The table depicted in Figure 35 showcases the substantial reduction in the number of attributes used for various assessments related to mental health conditions. Initially, a broad array of attributes was considered across multiple domains, including demographic, organizational, environmental, and various psychological assessments. Through systematic feature reduction techniques, I was able to streamline the attribute space significantly. Notably, the assessment areas of individual depression and suicide ideation/suicidal assessment were reduced from 49 to 3 and 51 to 2 attributes, respectively. This substantial reduction indicates that a small number of highly predictive attributes can be utilized effectively for the detection of complex mental health issues, such as depression and suicidality, which are of particular concern in high-stress environments like the military. The final deduction summary of my research emphasizes the importance of an adaptive model that can operate with a reduced attribute space while still

providing high diagnostic accuracy. This is crucial in a military context where timely and efficient screening can lead to early intervention and potentially save lives. It also underscores the flexibility required in clinical settings, suggesting that while the model provides a robust starting point for assessment, clinicians should still exercise their judgment, focusing on the clinical and psychosocial needs of the individual. This approach allows for a personalized assessment, combining the strengths of machine learning with the nuanced understanding of a human clinician. The ability to distil vast datasets into a manageable number of key indicators without losing predictive power is a significant advancement in the field of mental health diagnostics. This strategy paves the way for developing a more agile and focused assessment tool for the Indian Army, facilitating a proactive stance in the management of mental health disorders.

V. REFERENCES

[1] Kosinski, M., Stillwell, D., and Graepel, T. Private traits and attributes are predictable from digital records of human behavior. Proc. Natl. Acad. Sci. U.S.A. 110, 5802–5805, (2013).

[2] Monaro, M., Galante, C., Spolaor, R., Li, Q. Q., Gamberini, L., Conti, M., et al. Covert lie detection using keyboard dynamics. Scientific Reports 8 (1976).

[3] Vieira, S., Pinaya, H., and Mechelli, A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. Neurosci. Biobehav. Rev. 74(Part A), 58–75, (2017).

[4] Obermeyer, Z., and Emanuel, E. J. Predicting the future: big data, machine learning, and clinical medicine. N. Engl. J. Med. 375, 1216–1219, (2016).

[5] Pace, G., Orrù, G., Merylin, M., Francesca, G., Roberta, V., Boone, K. B., Malingering detection of cognitive impairment with the B test is boosted using machine learning. Front. Psychol. 10:1650 (2019).

[6] Navarin, N., and Costa, F. An efficient graph kernel method for noncoding RNA functional prediction. Bioinformatics 33, 2642–2650, (2017).

[7] Seidenberg, M. S. Connectionist models of word reading. Curr. Dir. Psychol. Sci. 14, 238–242(2005).

[8] Pashler, H., and Wagenmakers, E. J. Editors' introduction to the special section on reliability in psychological science: a crisis of confidence? Perspect. Psychol. Sci. 7, 528–530(2012).

[9] Breiman, L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat. Sci. 16, 199–231, (2001).

[10] Ioannidis, J. P., Tarone, R., and McLaughlin, J. K. The false-positive to false-negative ratio in epidemiologic studies. Epidemiology 24, 450–456. (2011).

[11] Zhang, J. M., Harman, M., Ma, L., and Liu, Y. Machine learning testing: survey, landscapes and horizons. arXiv [Pre-print]. (2019).

[12] Stef van Buuren, Karin Groothuis- Oudshoorn, "MICE: Multivariate Imputation by Chained Equations in R". Journal of Statistical Software 45: 1-67, (2011).

[13] Roderick J, A Little and Donald B Rubin "Statistical Analysis with Missing Data". John Wiley & Sons, Inc., New York, NY, USA, (1986).

[14] Domański, P.D. 'Study on Statistical Outlier Detection and Labelling'. Int. J. Autom. Computing. 17, 788–811, (2020).

[15] Jishan S.T., Rashu R.I., Mahmood A., Billah F., Rahman R.M. "Application of Optimum Binning Technique in Data Mining Approaches to Predict Students' Final Grade in a Course". Computational Intelligence in Information Systems. Vol 331. Springer, Cham, (2015).

[16] Jajuga, Krzysztof, and Marek Walesiak. "Standardisation of data set under different measurement scales." In Classifica-tion and information processing at the turn of the millennium, pp. 105-112. Springer, Berlin, Heidelberg, (2000).

[17] Reddy, G. Thippa, et al. "Analysis of dimensionality reduction techniques on big data." IEEE Access 8, (2020).

[18] Mladenić, Dunja. "Feature selection for dimensionality reduction." International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection". Springer, Berlin, Heidelberg, (2005).

[19] Pan, Sinno Jialin, James T. Kwok, and Qiang Yang. "Transfer learning via dimensionality reduction." AAAI. Vol. 8. (2008).

[20] Peluffo, Diego H., John A. Lee, and Michel Verleysen. "Recent methods for dimensionality reduction: A brief compara-tive analysis." ESANN, (2014).

[21] Khalid, Samina, Tehmina Khalil, and Shamila Nasreen. "A survey of feature selection and feature extraction techniques in machine learning." 2014 Science and Information Conference. IEEE, (2014).

[22] Ajzen, I. 'The Theory of Planned Behaviour. Organizational Behaviour and Human Decision Processes', 50, 179-211. (1991).

[23] Clark, L. A., & Watson, D. Constructing validity: Basic issues in objective scale development. Psychological Assess-ment, 7, 309–319, (1995).

[24] Kyriazos, T. A., & Stalikas, A. Applied Psychometrics: The Steps of Scale Development and Standardization Process. Psychology, 9, 2531-2560, (2018).

[25] Fabrigar, L. R., & Ebel-Lam, A. Questionnaires. In N. J. Salkind (Ed.), Encyclopedia of Measurement and Statistics. Thousand Oaks, CA: Sage, pp. 808-812 (2007).

[26] Dorans, N. J. Scores, Scales, and Score Linking. The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development, V.II, pp. 573-606, (2018).

[27] Chadha, N. K. Applied Psychometry. New Delhi, IN: Sage Publications. (2009).

[28] Price, L. R., Psychometric Methods: Theory into Practice. New York: The Guilford Press. (2017).

[29] Dorans, N. J. "Scores, Scales, and Score Linking. The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development", V.II (pp. 573-606), (2018).

[30] DeVellis, R. F. 'Scale Development: Theory and Applications' (4th ed.). Thousand Oaks, CA: Sage. (2017).

[31] Jenkins, G. D., & Taber, T. D. 'A Monte Carlo Study of Factors Affecting Three Indices of Composite Scale Reliability'. Journal of Applied Psychology, 62, 392-398. (1977).

[32] Streiner, D. L., Norman, G. R., & Cairney, J. 'Health Measurement Scales: A Practical Guide to Their Development and Use' (5th ed.). Oxford, UK: Oxford University, (2015).

[33] Dimitrov, D. M. "Statistical Methods for Validation of Assessment Scale Data in Counselling and Related Fields". Alexandria, VA: American Counselling Association. (2012).

[34] Morrison, K. M., & Embretson, S. 'Item Generation. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), The Wiley Hand-book of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development", V.I (pp. 46-96), (2018).

[35] Demaio, T., & Landreth, A. "Do Different Cognitive Interview Methods Produce Different Results", Questionnaire Development and Testing Methods. Hoboken, NJ: Wiley. (2004).

[36] Raykov, T. "Scale Construction and Development Using Structural Equation Modelling". R. H. Hoyle (Ed.), Handbook of Structural Equation Modeling (pp. 472-492). New York: Guilford Press. (2012).

Cite this article as :

Sudipto Roy, Jigyasu Dubey, "Optimizing Mental Health Detection in Indian Armed Forces Personnel through Feature Engineering Driven Dataset Reduction, Addressing Suicide, Depression, and Stress", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 10, Issue 2, pp. 70-81, March-April-2024.