Replicating the Study: Predicting Mental Health Disorders using Machine Learning for Employees in Technical and Non-Technical Companies

By Belaynesh Mossie (EP7203363)

Introduction:

The project involves reproducing the findings from the 2020 IEEE paper titled Predicting Mental health disorders using Machine Learning for employees in technical and non-technical companies by Rahul Katarya and Saurav Maan. The paper explored the application of machine learning algorithms to predict mental health disorders in working professionals, specifically using a dataset from the publicly available Mental Illness (OSMI) survey. The authors compared six different machine learning models, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forest, to classify mental health disorders based on workplace-related factors and they also examined the most contributor of mental health disorder among the selected features. The objective of this reproduction project is to replicate the methodology used in the original paper to validate the results and gain insights into the practical implementation of machine learning for predicting mental health disorders in the workplace. The reproduction includes training machine learning models on the same dataset and evaluating their performance in terms of precision, recall, and accuracy and find the best predictor of mental health disorder by doing feature importance analysis.

Original paper reference: Predicting Mental health disorders using Machine Learning for employees in technical and non-technical companies | IEEE Conference Publication | IEEE Xplore

Methodology

To ensure a proper comparison with the original study, I applied the same methodology used by the authors of the original Paper. The same algorithms were implemented and I used same features they selected for their analysis.

While the original study does not specify their data preprocessing steps, I took a conservative approach to prepare the dataset for analysis.

Data Cleaning: Any ambiguous or uncertain entries such as those labeled as "possibly" or "Don’t know" were removed. Missing values in feature columns were addressed using imputation to ensure dataset completeness and suitableness for modeling. After cleaning, I retained 251 valid entries for model training and evaluation.

Machine Learning Models: The same machine learning algorithms are then implemented as the original paper, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, and Naive Bayes and all the rest. Additionally, I included XG Boosting, a more advanced machine learning model to test if it offered improved predictive performance.

Evaluation metrics: Using K-Fold Cross Validation, I aimed to assess each model's performance with the primary metrics being Accuracy, Precision, Recall, and F1 Score as used in the original paper.

In summary, I followed the original methodology closely, addressing the handling of missing and imbalance data and introducing XGBoost to further evaluate the predictive power of the chosen features.

Results & Discussion

In this section, I present the results obtained from applying machine learning models similar to those utilized in the original study by the authors, along with improvements I performed. The performance of each model is evaluated, and the findings are discussed in comparison with the original paper to identify differences and their potential causes.

Results

1. K-Nearest Neighbors

The K-Nearest Neighbors algorithm achieved a mean accuracy of 84.8%, with precision at 82% and recall at 95%. The high recall suggests that KNN is very effective at identifying true positives, meaning it successfully detects a large proportion of actual cases of mental health disorders. However, its precision of 82% implies that it also incorrectly classifies some negative cases as positive, leading to false positives. This can be expected due to the class imbalance in the dataset, where the positive class is more prevalent. Even though class weights were applied, the majority class still has influence on the model's performance. The F1 score, which balances precision and recall, stood at 88, indicating that KNN achieves a solid tradeoff between these metrics.

2. Support Vector Machine

Similarly, the Support Vector Machine model produced the same accuracy of 84.8% as KNN, with precision also at 82% and recall a bit lower 94.5%. SVM, like KNN, excels at identifying true positives but is a bit more conservative when it comes to classifying negative cases as positive. Its F1 score was also 88, reflecting a comparable performance to KNN in balancing precision and recall. The positive class, being dominant, could have influenced the model's ability to classify more cases as positive to capture the majority class, even though class weights were adjusted to counter this.

3. Logistic Regression

The logistic regression yielded an accuracy of 84.45%, which is very close to KNN and SVM. Its precision was slightly lower, at 80.8%, but it demonstrated a high recall of 96%, suggesting that it was particularly strong in detecting all the true positives, even at the cost of some false positives. Logistic Regression's low precision and high recall may be influenced by the larger number of positive cases in the dataset, as the model is encouraged to predict more positive instances. The F1 score for Logistic Regression was 87, reflecting its tradeoff between precision and recall, with a small lean toward better recall.

4. Decision Tree

The decision tree algorithm got a mean accuracy of 82%, with an 81% precision and 91% recall. While this model performed reasonably well, the accuracy and precision were lower compared to other classifiers. This may reflect its tendency to overfit the data, despite I uses the optimal parameter tuning using grid search.

5. Random Forest

Random Forest and Naive Bayes both achieved an accuracy of 83%. They performed similarly in terms of precision (81.7%) and recall (92.5%). The F1 scores for both models were 86.6, indicating an effective tradeoff between precision and recall while still capturing a high proportion of true positives. This might arise from low dimension of data and models similarity with such low dimensional data. The class imbalance still influences the models, potentially leading to lower precision due to their sensitivity to the positive class

7. XG Boost

Finally, XG Boost performed similarly to the other top scored models, with an accuracy of 84%, precision at 80%, and recall at 95%. Its F1 score of 87 reflects good overall performance, with strong recall and a good precision. As with the other models, the class imbalance may have an influence in the lower precision, since the model likely predicted more positives to balance its overall performance.

Generally, the results indicate that KNN and SVM performed the best in terms of accuracy and recall, showcasing their ability to detect mental health disorders effectively. While Decision Trees

showed slightly lower performance metrics, ensemble methods like Random Forest and XG Boost demonstrated competitive results, confirming their utility in predictive modeling.

The dataset used in this replication consists of 104 negative instances and 147 positive instances, creating a class imbalance where the positive class is the majority. This imbalance has the potential to affect model performance, especially in terms of precision. Despite the use of class weights in all models to address this, the distribution of the positive class means that models are often more sensitive to detecting positive instances, which can lead to higher recall at the expense of precision. In essence, while the models were designed to avoid misclassifying the minority class, the class imbalance might still have caused the models to predict positives more frequently. The following confusion matrix of some models can show the distribution of false positives and how the models bias to positive classification.

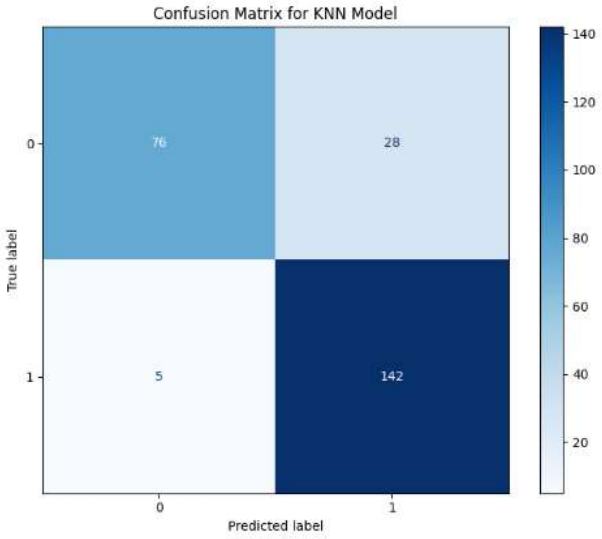

Figure 1. Confusion matrix for SVM

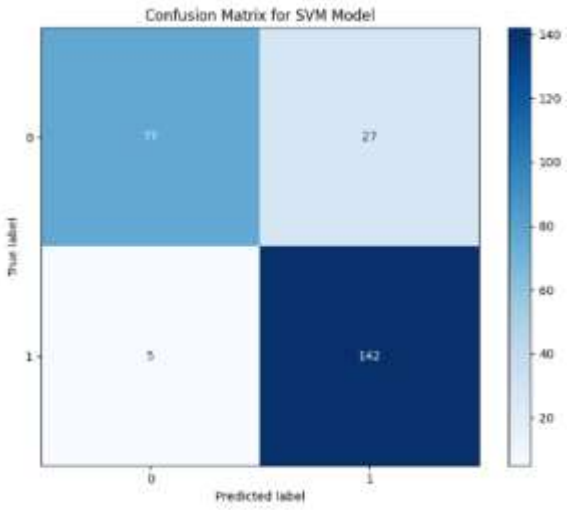

Figure 2. Confusion matrix for SVM

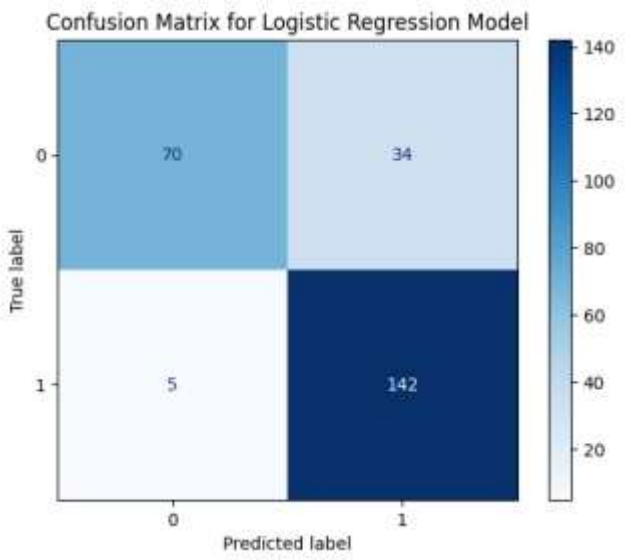

Figure 3. Confusion matrix for SVM

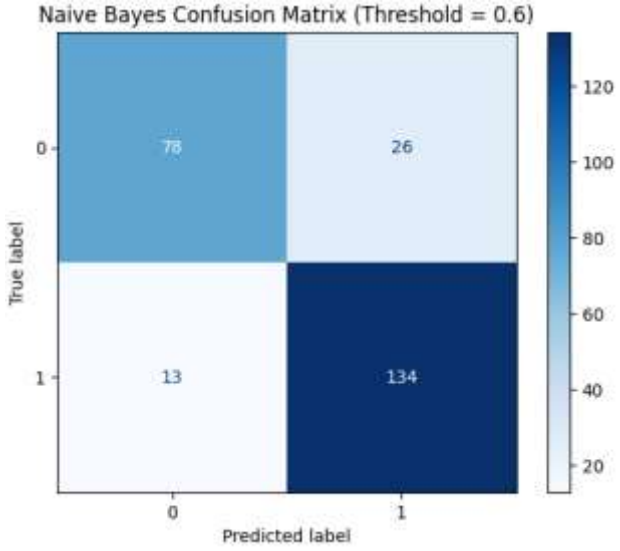

Figure 4. Confusion matrix for SVM

Overall, The KNN and Support Vector Machine algorithms exhibited the best overall performance, both achieving the highest accuracy with high recall and balanced F1 scores, making them the top candidates for detecting mental health disorders effectively.

Feature Importance

In addition to evaluating the performance of the models, feature importance analysis discovered that past history of mental health disorders was the most significant predictor, followed by whether the individual had discussed their mental health with an employer, and then family history**.** This aligns with original paper, showing that a person's previous mental health history, and family history are crucial factors in predicting mental health disorders.

Also, the result reveals that machine learning algorithms has a potential for predicting mental health disorders in employees across various sectors. The ongoing challenge remains in obtaining a balanced performance across precision and recall, thereby minimizing both false positives and false negatives in clinical applications

Discussion

In the original study, the Decision Tree and Logistic Regression algorithms showed the highest accuracy. Both models emphasized the need to trade-off between these precision and recall metrics, which are critical in the context of mental health disorder detection.

In the replication, K-Nearest Neighbors and Support Vector Machine both achieved 84.8% accuracy, matching the performance of Decision Tree from the original paper. These models also exhibited strong recall values of 95% (KNN) and 94.5% (SVM), suggesting that these algorithms were very effective in classifying true positives. However, their precision was slightly lower compared to other metrics, indicating that they might have misclassified more negative cases as positive. Logistic Regression also performed well with an accuracy of 84.45%**,** showing the highest recall of 96%**,** but the precision was slightly lower at 80.8%**.**

In both the original paper and my replication precision was consistently low across all the models it might arise from the data imbalance and being the precision the minimum of all the metrics in it all results were highly improved the replication.

Table1: Original study results

S.no	Algorithms	Accuracy (%)	Precision	Recall	F1
1.	KNN	74%	76	82	79
2.	SVM	76%	75	88	81
3,	Logistic Regression	84%	82	94	87
4.	Decision tree	84%	83	92	87
5.	Random Forest	77%	81	80	81
6.	Naïve bayes	79%	78	90	83

Table2: New results

No	Algorithm	Accuracy	Precision	Recall	F1
1	KNN	84.8%	82	95	88
2	SVM	84.8%	82	94.5	88
3	Logistic regression	84.45%	80.8	96	87
4	Decision Tree	82%	81	91	85.6
5	Random Forest	83%	81.7	92.5	86.6
6	Naive Bayes	83%	81.7	92.5	86.6
7	XG Boost	84%	80	95	87

In my reproduction of the study, I also applied cross-validation to ensure the robustness of the results, especially considering the limited amount of data available. Cross-validation helps in addressing overfitting and provides a more generalized estimate of model performance across

different subsets of the data. This was particularly important in my replication due to the small sample size, as it allowed for more reliable performance metrics.

Additionally, I applied grid search to optimize the hyperparameters of the models. This method allowed me to find the best-performing parameters for each algorithm, which likely contributed to the improvements seen across most of the models. The class balancing technique, including class weights, was also used to address the class imbalance in the dataset. By compensating for the unequal number of positive and negative instances, this method helped to reduce the bias toward predicting the majority class, and improving precision and recall for most models.

In conclusion, almost all models especially KNN and SVM, showed significant improvements with the use of cross-validation, grid search for hyperparameter tuning, and class balancing, except the Decision Tree algorithm a bit lower performance which the accuracy is compromised to to tradeoff between precision and recall.

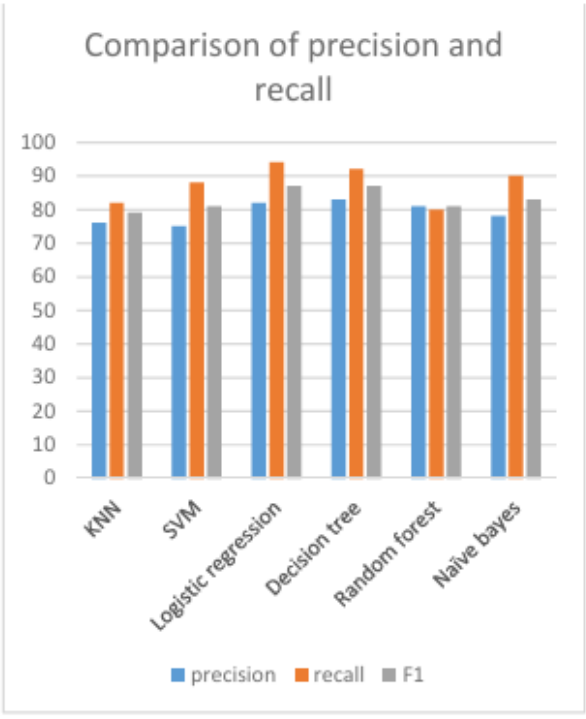

Figure 3 Fig.1Performance comparison of models for the original paper

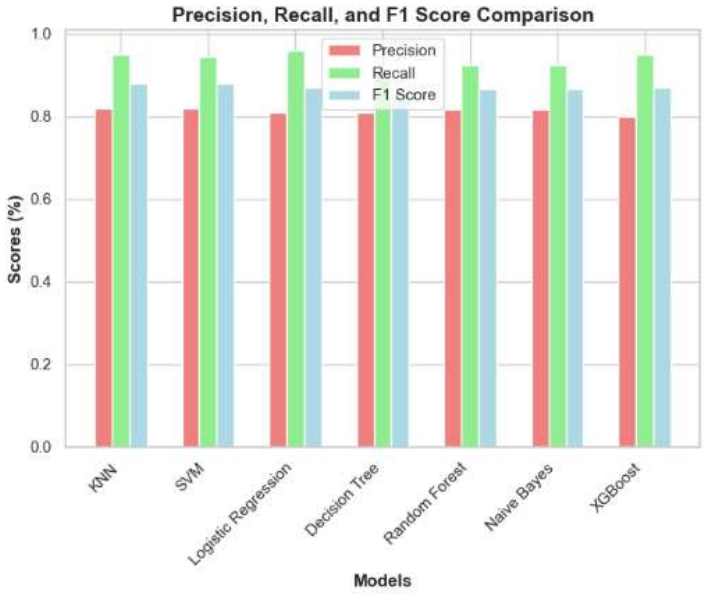

Figure 3 Fig.1 Performance comparison for new result

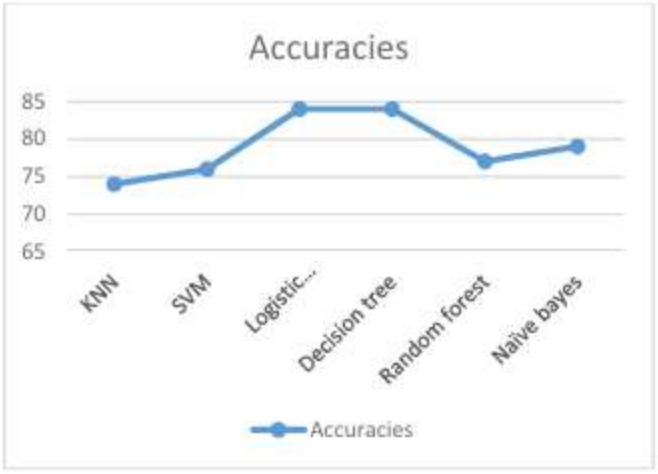

Figure 3Accuracy comparison original paper

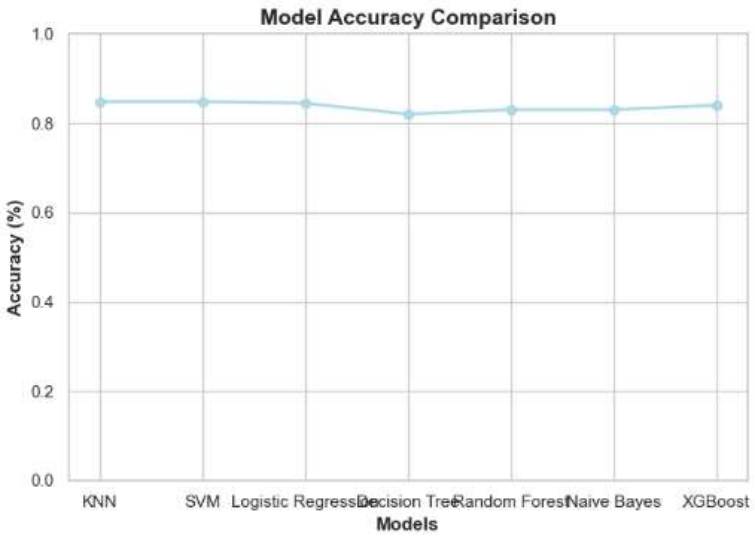

Figure 2Accuracy comparison New

Feature importance comparison

In the original study, family history and mental health history were identified as the most important features for predicting mental health disorders. My reproduction of the study resulted similar results, with past history of mental health disorders being as the most significant contributor. This was followed by whether the individual had discussed their mental health with an employer**,** and then by family history.

The discussing mental health with an employer being as the second most important feature. suggests that in addition to personal and family health histories, the social and professional context of the individual, such as whether they have openly discussed mental health, also plays a vital role in predicting mental health disorders.

Conclusion

This reproduction successfully validated the original study's findings while introducing additional analyses and provided new insights into the effectiveness of various machine learning models for predicting mental health disorders. Improvements were observed in the models' performance due to the application of cross-validation**,** grid search for hyperparameter tuning, and class balancing to address the dataset's class imbalance. While, the Decision Tree algorithm underperformed slightly compared to others, it is still a useful model, especially when seen in light of its trade-offs between precision and recall.

The feature importance analysis further reinforced the key role of past mental health history and family background in predicting mental health disorders, while also emphasizing the influence of workplace dynamics, such as discussions with employers, in shaping mental health outcomes.

In conclusion, the reproduction not only confirmed the original study's findings but also demonstrated an improvement and shown the potential of ensemble methods, such as Xg Boosting, for further improvement in prediction accuracy.