**Introduction:**

This report presents the data mining models along with their respective Evaluation Metrics applied on the OSMI dataset used in the project, providing insights into mental health in the workplace in 2016. The dataset contains 1433 rows and 63 columns.

**Logistic Regression:**

Text

Description automatically generated



The logistic regression model was trained using a grid search with three hyperparameters: penalty ['l1','l2'], C np.logspace(-3,3,7), and solver['newton-cg', 'lbfgs', 'liblinear'] . The confusion matrix shows that the model was able to correctly classify a majority of the instances, with an accuracy of 76.31%. The recall score of 90.69% indicates that the model was able to correctly identify a high percentage of the positive instances, while the precision score of 76.46% indicates that it correctly identified a lower percentage of the positive instances. The F1 score of 82.97% indicates a good balance between precision and recall. The ROC AUC score of 79.17% indicates that the model can distinguish between the two classes moderately well. Overall, the logistic regression model with the chosen hyperparameters is moderately successful in classifying the instances in the OSMI dataset.

Also, the model achieved an accuracy score of 0.7368 on the test set. The confusion matrix shows that the model correctly predicted 23 out of 60 samples in the first class, 66 out of 79 samples in the second class, and 93 out of 108 samples in the third class. The model had the highest recall score of 0.8611 for the second class, indicating that it correctly identified a high proportion of the actual positive samples in that class. The model had the highest precision score of 0.7209 for the first class, indicating that of all the samples predicted to be positive in that class, the majority were true positives. The F1 score was 0.7848, which is the harmonic mean of precision and recall. The ROC AUC score was 0.7753, indicating that the model performed relatively well in distinguishing between positive and negative samples.

Chart

Description automatically generated

Chart

Description automatically generated

**Random Forest:**

Graphical user interface, text, application

Description automatically generated



The Random Forest model was trained using a parameter grid search with n\_estimators ranging from 1 to 100 and max\_depth ranging from 1 to 20. The confusion matrix shows that the model correctly classified 131 instances of the first class, 308 instances of the second class, and 412 instances of the third class. However, there were also misclassifications, as shown in the confusion matrix.

The accuracy score of the model was 0.861, indicating that it correctly classified 86.1% of the instances in the dataset. The recall score, or the ability of the model to correctly identify true positives, was 0.983, indicating that the model correctly identified 98.3% of instances of all three classes. The precision score, or the ability of the model to correctly identify positive instances, was 0.798, indicating that the model correctly identified 79.8% of positive instances.

The F1 score, which is the harmonic mean of precision and recall, was 0.881, indicating that the model achieved a balance between precision and recall. Finally, the ROC AUC score was 0.872, indicating that the model had a good ability to distinguish between the three classes.

Also, the model achieved an accuracy of 0.77 on the test data. It correctly identified 96% of the positive cases (high recall), but its precision was relatively low at 72%. The F1 score was 0.83, indicating a reasonably good balance between precision and recall. The ROC AUC score was 0.80, which indicates that the model is better than random at distinguishing between positive and negative cases.

Chart

Description automatically generated with low confidence

Chart

Description automatically generated

**Support Vector Classifier:**

Graphical user interface, application, Word

Description automatically generated



The SVC model was trained with a grid-search of different kernel options: poly, rbf, linear, and sigmoid. The model achieved an accuracy score of 0.80, which means that 80% of the samples were correctly classified. The recall score of 0.95 indicates that the model was able to correctly identify 95% of the positive samples (True Positives) out of all positive samples. The precision score of 0.76 indicates that out of all the samples predicted as positive, only 76% were actually positive. The F1 score of 0.85 is the harmonic mean of precision and recall and provides a balance between them. The ROC AUC score of 0.82 indicates that the model has moderate discrimination ability between the positive and negative samples. Overall, the model achieved decent performance, but it may be improved by further hyperparameter tuning or trying different models.

The models were trained on a dataset and then tested on a separate testing dataset to evaluate their performance. The logistic regression model achieved an accuracy of 73.68% with a recall score of 86.11% and a precision score of 72.09%. The random forest model achieved an accuracy of 77.33% with a recall score of 96.30% and a precision score of 72.22%. The SVM model achieved an accuracy of 75.30% with a recall score of 91.67% and a precision score of 71.74%. Overall, the random forest model performed the best among the three models. However, all models have fairly similar F1 scores and ROC AUC scores, which indicates that they all performed reasonably well at predicting the target variable.

Text, letter

Description automatically generated

Text

Description automatically generated

**Gradient Boosting Classifier:**



The results of the Gradient Boosting Classifier model trained with the given parameter grid "learning\_rate": [0.01, 0.05, 0.1], "max\_depth": [3,5,8], "n\_estimators": [10] show an accuracy score of 0.763 and an AUC-ROC score of 0.792, indicating a moderately good performance of the model. The recall score of 0.907 indicates that the model was able to correctly identify a high proportion of positive cases, while the precision score of 0.765 indicates that a substantial number of negative cases were classified as positive by the model. The F1 score of 0.830 indicates that the model has a balance between precision and recall. Further parameter tuning and model optimization can be done to improve the performance of the model.

The Gradient Boosting Classifier achieved an accuracy of 73.68% on the test set, correctly classifying 23 samples of class 0, 66 samples of class 1, and 93 samples of class 2. The recall score was highest for class 1, indicating that the model was able to correctly identify most of the samples of this class. The precision score was highest for class 0, suggesting that the model was most precise in predicting this class. The F1 score, which balances precision and recall, was highest for class 1, indicating that the model was best at predicting this class overall. The ROC AUC score was 0.775, suggesting that the model performs moderately well at distinguishing between positive and negative classes.

Chart

Description automatically generated

Chart

Description automatically generated

**Neural Networks Classifier:**

Graphical user interface, text, application, Word

Description automatically generated



The given code performs hyperparameter tuning using GridSearchCV of 'hidden\_layer\_sizes': [(10,), (100,), (10,10,), (50,50,), (100,100,)], 'activation': ['identity', 'logistic', 'tanh', 'relu'], 'solver': ['lbfgs', 'sgd', 'adam'], 'alpha': [0.0001, 0.001, 0.01, 0.1], 'learning\_rate': ['constant', 'invscaling', 'adaptive']

for the Neural Networks model. The parameters include the number of hidden layers and neurons, activation functions, solvers, regularization parameter (alpha), and learning rate. The model was trained on the dataset and evaluated based on various metrics such as accuracy, recall, precision, F1 score, and ROC AUC score.

The confusion matrix indicates the number of true positive, true negative, false positive, and false negative values. The accuracy score measures the overall performance of the model in terms of the proportion of correctly classified instances. The recall score measures the proportion of correctly classified positive instances, while the precision score measures the proportion of correctly classified positive instances among all predicted positive instances. The F1 score is the harmonic mean of recall and precision, which provides a balanced evaluation metric. The ROC AUC score is a measure of how well the model can distinguish between positive and negative classes.

In this case, the neural network model with the given hyperparameters resulted in an accuracy score of 0.76, a recall score of 0.91, a precision score of 0.76, an F1 score of 0.83, and an ROC AUC score of 0.79. It is important to note that these scores may be improved by further hyperparameter tuning or using different models altogether.

Chart

Description automatically generated

Chart

Description automatically generated

**Performance Evaluation Summary:**

Overall, the performance of the models varied with different evaluation metrics. The Random Forest and SVC models showed the highest accuracy scores of 0.77 and 0.75, respectively, with good precision, recall, F1, and AUC scores. The Gradient Boosting Classifier and Logistic Regression models had slightly lower accuracy scores of 0.74 and 0.73, respectively, with comparable precision, recall, and F1 scores but lower AUC scores. The Neural Network model had an accuracy score of 0.76 with good recall and F1 scores but lower precision and AUC scores.

Considering all the evaluation metrics, the Random Forest and SVC models performed the best among the tested models. However, since Random Forest had a slightly better AUC score and showed better performance in predicting the rarest class (class 3), it is selected as the best model for this classification task.