FINAL REPORT

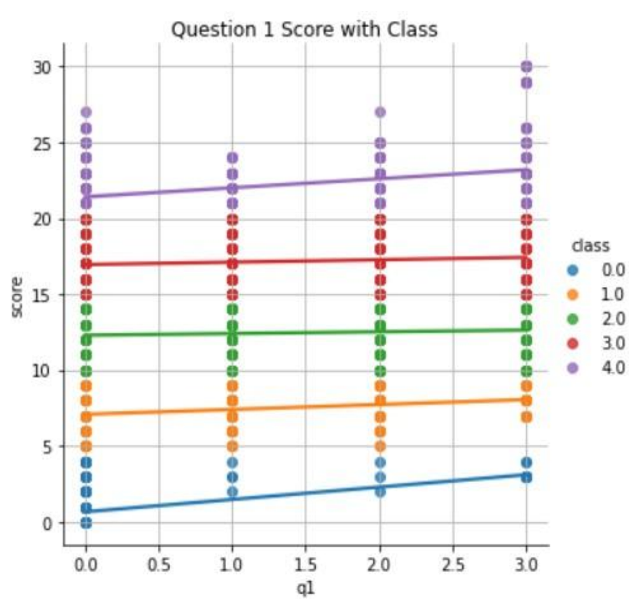
EmoScore – A Mental Health Prediction App Using Machine learning with data visualization and analysis.

Problem Description:

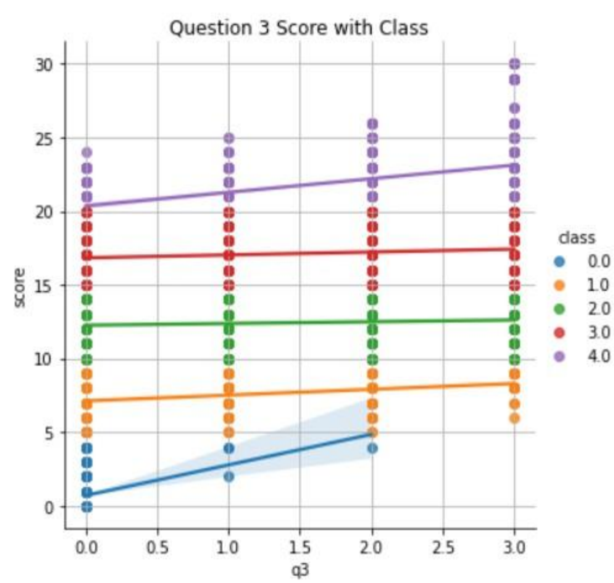
The original project proposed the development of an app named EmoScore, aimed at utilizing mental health data to predict anxiety and other mental health issues. Users would input data related to various factors such as age, gender, occupation, duration of staying indoors, and reported situations during quarantine. The app would then conduct qualitative analysis based on user characteristics and provide visualizations of the analyzed data. Predictions of anxiety and other mental health issues would be categorized as less likely, moderate, or likely. EmoScore aims to provide a user-friendly mobile app that can predict an individual's mental health status based on various factors and visualize the analyzed data to raise self-awareness.

Solution Approach:

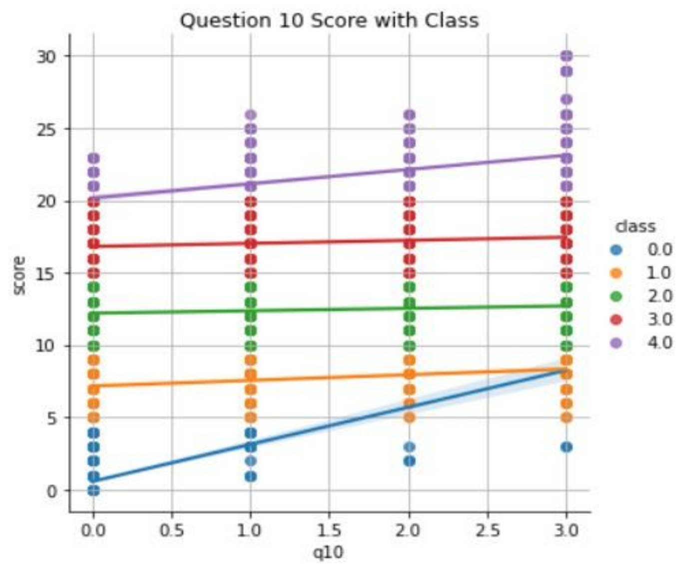
1. **Data Analysis**:
    - Quantitative Analysis: Utilize statistical methods to analyze the frequency of different factors reported by users.
    - Qualitative Analysis: Conduct qualitative analysis based on the characteristics reported by users to assess their mental health status.
    - Predictive Analysis: Develop algorithms to predict anxiety and other mental health issues based on the collected data and user inputs.



Fig(a)



Fig(b)

Fig(c) 

2. **Data Visualization**:

   a. Generate visualizations to present the analyzed data in an understandable format. This could include graphs, charts, and other visual aids to represent patterns and trends.

   b. Provide visualizations that show the distribution of factors such as age groups, gender, occupation, duration of staying indoors, reported situations during quarantine, etc.

IMPLEMENTATION:

Algorithms Used:

Since we are dealing with multi-class classification problem, we used several machine learning algorithms:

1. **Naive Bayes** which is a probabilistic classifier based on the Bayes' theorem with an assumption of independence between features. In the context of EmoScore, Naive Bayes can be used to model the probability of a user belonging to each mental health risk category (less likely, moderate, or likely) based on the feature values.

2. **K-Nearest Neighbor (KNN)** KNN is a non-parametric algorithm that classifies new instances based on their similarity to the k closest training examples in the feature space. In the case of EmoScore, KNN can be used to predict the mental health risk category of a user by finding the k nearest neighbors in the training data and assigning the majority class among those neighbors.

3. **Support Vector Machine (SVM)** SVMs are powerful supervised learning models that construct hyperplanes in a high-dimensional space to separate different classes. In the context of EmoScore, SVMs can be used to find the optimal hyperplane that maximizes the margin between the mental health risk categories, allowing for accurate classification of new instances.

	model	accuracy	time
0	GaussianNB()	0.872214	0.000949
1	GaussianNB()	0.872214	0.000910
2	KNeighborsClassifier()	0.897959	0.001273
3	KNeighborsClassifier()	0.898273	0.001372
4	SVC()	0.957614	0.001108
5	SVC()	0.956672	0.001092
6	DecisionTreeClassifier()	0.802826	0.000931
7	DecisionTreeClassifier()	0.806593	0.000774
8	(DecisionTreeClassifier(max_features='sqrt', r...	0.894819	0.008478
9	(DecisionTreeClassifier(max_features='sqrt', r...	0.901727	0.008595
10	<keras.engine.functional.Functional object at ...	0.963579	0.043399

Fig(d): Accuracy and Run time for all models using Normalized and Unnormalized data.

4. **Decision trees** are tree-like models that learn decision rules from the training data by recursively partitioning the feature space. In the EmoScore project, decision trees can be used to create a hierarchical model that makes decisions based on the user's features, ultimately leading to the prediction of the mental health risk category.

5. **Random Forest** is an ensemble learning method that combines multiple decision trees trained on different subsets of the data and features. In the context of EmoScore, Random Forest can be used to reduce the risk of overfitting and improve the generalization performance of the mental health risk prediction.

6. **Neural Network** are a class of machine learning models inspired by the structure and function of biological neural networks. In the EmoScore project, neural networks can be employed to learn the non-linear relationships between the user's features and the mental health risk categories.

Technologies Used:

For this project, we utilized a publicly available dataset obtained from an online source. The dataset contains information related to mental health factors, including demographic data, self-reported mental health conditions, and various features that could potentially influence mental health status.

Data Preprocessing

Before implementing the dataset in our machine learning models, we performed several data preprocessing steps using Python libraries such as Pandas and NumPy which were utilized for data preprocessing tasks, such as handling missing values, encoding categorical variables, and scaling numerical features.

Data Splitting

After preprocessing, we split the dataset into training and testing subsets. The training set was used to train and optimize our machine learning models, while the testing set was held out to evaluate the models' performance on unseen data, ensuring an unbiased assessment of their generalization capabilities.

Here's a brief overview of how the various technologies were utilized in this project and the corresponding results:

React Native was used to develop the user interface and core functionality of the EmoScore mobile app. It allowed for the creation of a cross-platform application with a native-like user experience on both iOS and Android devices. The app's user interface, including screens for data collection, visualization, and prediction display, was built using React Native components and libraries.

The preprocessed dataset was then utilized to train and evaluate various machine learning models, including algorithms such as logistic regression, random forests, gradient boosting, and neural networks. These models were implemented using libraries like scikit-learn and TensorFlow.

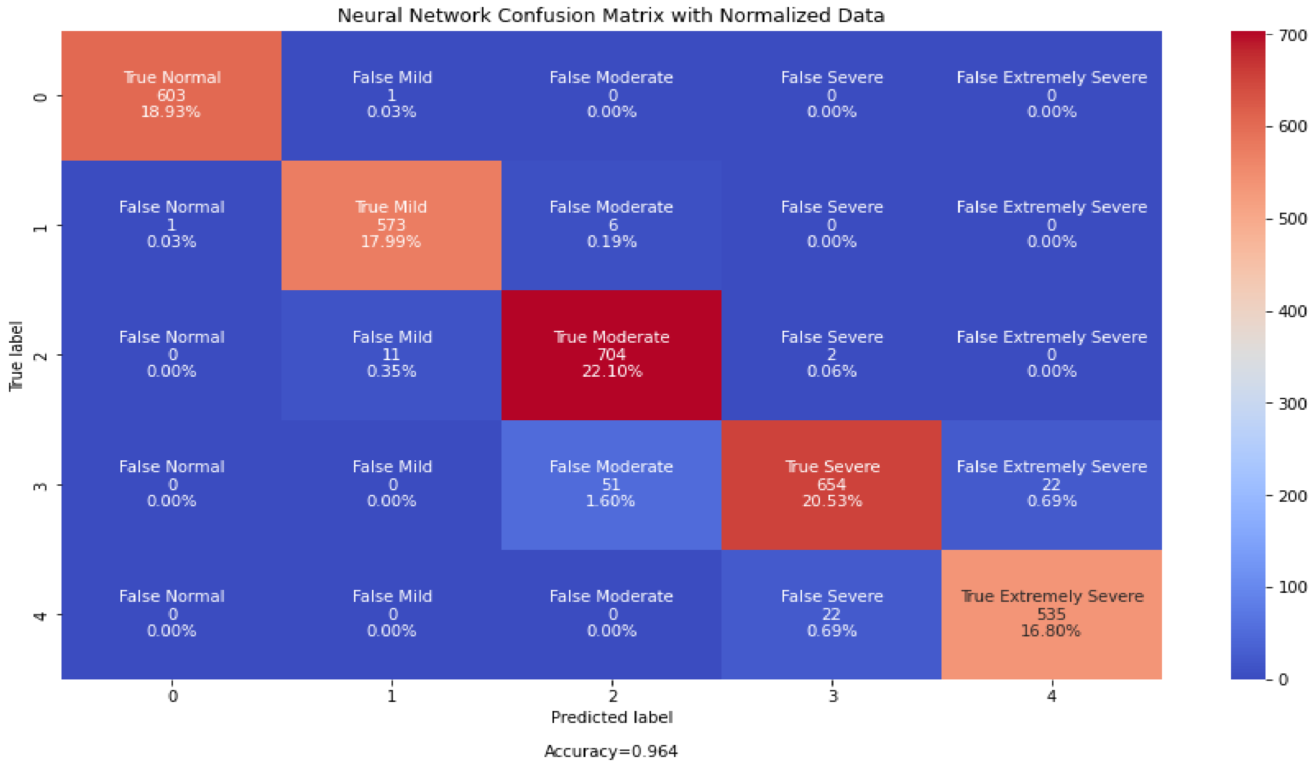
During the model development process, we employed techniques like cross-validation and hyperparameter tuning to optimize the models' performance and prevent overfitting. The trained models were then integrated into the EmoScore mobile app using TensorFlow Lite for on-device inference, allowing users to receive personalized mental health risk predictions based on their input data.

By leveraging a publicly available dataset and implementing it through rigorous data preprocessing, feature engineering, and machine learning techniques, we aimed to develop accurate and reliable models for predicting mental health risk categories in the EmoScore app.

Model Evaluation Results:

For evaluating the performance of the machine learning models developed for the EmoScore app, we employed various techniques and metrics, including confusion matrices and TensorFlow model evaluation metrics.

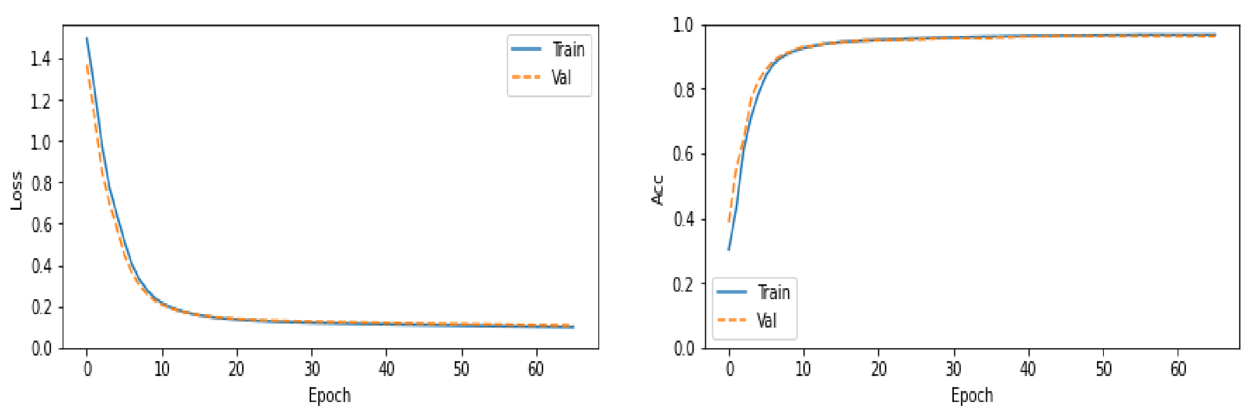
Confusion Matrix is a tabular representation that summarizes the performance of a classification model by comparing the predicted labels with the actual labels from the ground truth data. In the context of the EmoScore project, where we aimed to classify users into several mental health risk categories (Normal, Mild, Moderate, Severe, Extremely Severe), the confusion matrix helped us understand how well the model performed in predicting each category. By analyzing the confusion matrix, we could identify potential misclassifications and areas for improvement.



Fig(e): Confusion Matrix

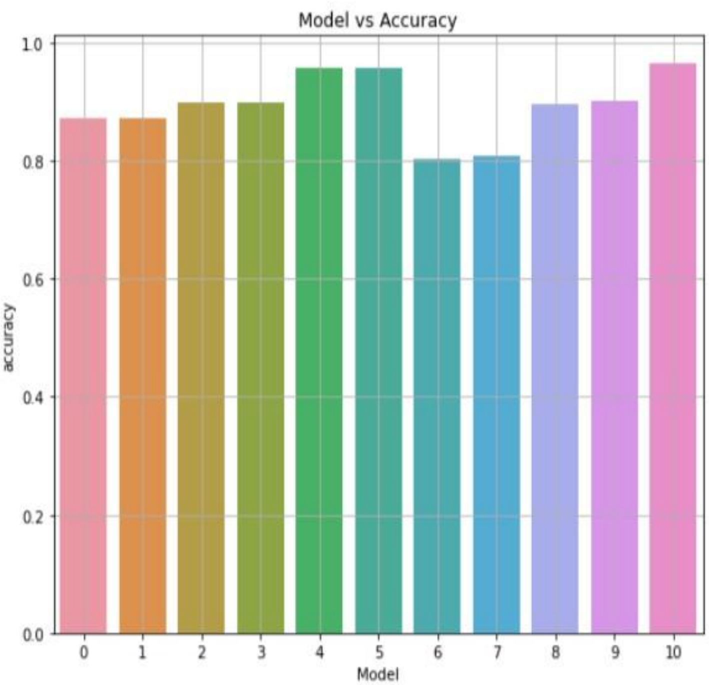
TensorFlow provides a range of evaluation metrics and techniques that we utilized to assess the performance of our machine learning models. Some of the key metrics used in the EmoScore project include:

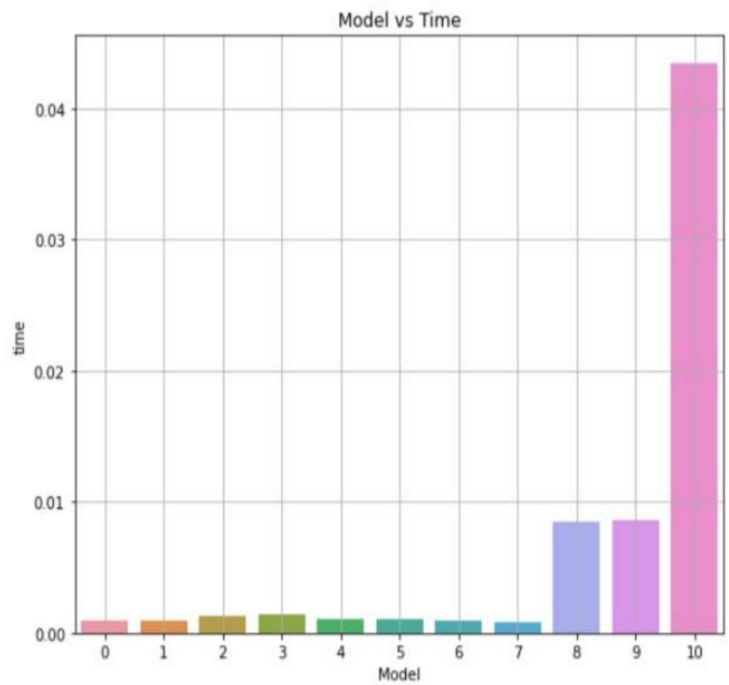
1. **Accuracy**: This metric measures the overall correctness of the model's predictions by calculating the ratio of correctly classified instances to the total number of instances.
2. **Precision**: Precision measures the proportion of true positive instances among the instances predicted as positive by the model. It is particularly useful when the cost of false positives is high.
3. **F1-score**: The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance.



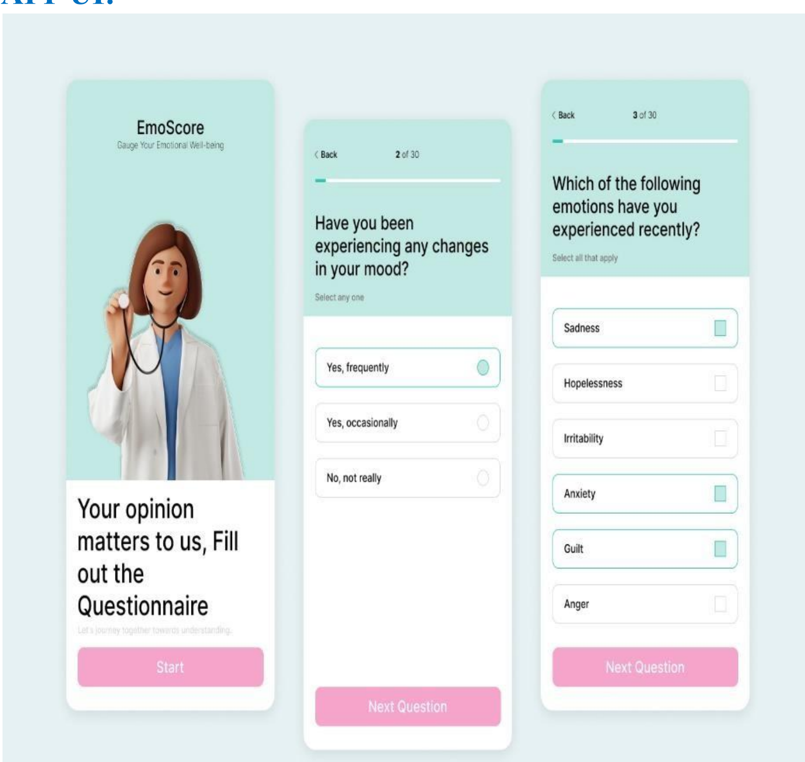
Fig(f)

4. **Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC)**: The ROC curve is a plot of the true positive rate against the false positive rate at various classification thresholds.





APP UI:



REFERENCES:

1. https://www.mayoclinic.org/diseases-conditions/depression/diagnosis-treatment/drc20356013.

2. World Health Organization. (2022). Mental health. https://www.who.int/health-topics/mentalhealth

3. Torous, J., Kiang, M. V., Lorme, J., & Onnela, J. P. (2019). New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR mental health, 6(3), e11734. https://doi.org/10.2196/11734

4. Cho, C. H., Lee, T., Kim, M. G., In, H. P., Kim, L., & Lee, H. J. (2019). Mood prediction of patients with mood disorders by machine learning using passive digital markers from daily life. Scientific reports, 9(1), 1-11. https://doi.org/10.1038/s41598-019-44097-2

5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

6. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467. https://arxiv.org/abs/1603.04467