Smoking Detection using Biosignals and Machine Learning

Akhilesh Yadlapalli
200103114898
akya21@student.bth.se

Siddharth Reddy Aleti
20010112-T150
sial21@student.bth.se

I. Introduction

Smoking is a major public health concern that negatively affects nearly every organ of the body and is the leading cause of preventable morbidity and mortality worldwide. The success rate for stopping smoking is still low even with evidence-based treatment. The World Health Organization (WHO) reports that smoking is the greatest global contributor to preventable illness and mortality [4]. Despite the well-known health concerns connected with smoking, many people still do so, which raises serious public health issues. Predictive models for identifying smokers who have a higher possibility of quitting have been presented as a solution to this problem.

Those who want to stop smoking can do it with the support of evidence-based treatments, such as counseling and medication [2]. Nonetheless, less than one third of participants were able to successfully stop smoking, indicating that the success rate for doing so is still low [3]. As a result, scientists have been looking into different ways to spot smokers who have a higher likelihood of quitting.

In leveraging bio-signals like heart rate, blood pressure, and ECG to predict smoking status, machine learning algorithms have showed potential. These bio-signals can be collected via wearable technology, and they can be utilized to build a machine learning model that can determine a person's smoking status.

The usefulness of machine learning algorithms in predicting smoking status using bio-signals has been shown in numerous studies. For instance, Thakur et al. [6]'s study discovered that wearable device data may be used to reliably predict smoking status using machine learning algorithms. Similar to this, Shashikant and Chetankumar [5] discovered that utilizing information about heart rate variability, machine learning algorithms could identify cardiac arrest's in the smoker.

In this project, we are planning to compare three machine learning models, namely Linear Regression, Random Forest, and Support Vector Machines (SVM), to predict the smoking status of individuals based on their demographic and health-related attributes.

II. Methodology

In this section, we get into the details of the process and analysis involved in this work. Initially, the inquiry involved data gathering, cleaning, and feature selection. Subsequently, the featured data was preprocessed to create the necessary format. The next step involved creating training and testing datasets from the data, followed by training the models using algorithms. The accuracy of the models was evaluated using the testing data and studied the results.

A. Data Collection

The dataset used in this study was obtained from "www.kaggle.com", a popular platform for data science competitions and data exploration.[1]

B. Dataset and Attributes

The dataset included a wide range of demographic, anthropometric, and clinical variables, as well as smoking status. The bio-signals used for predicting smoking status included measures of blood pressure (systolic and diastolic), fasting blood sugar, cholesterol levels (total, HDL, and LDL), triglycerides, hemoglobin levels, urine protein, serum creatinine, AST and ALT levels, and dental caries. The smoking variable was binary, with a value of 0 indicating non-smoking and a value of 1 indicating smoking.

C. Data Cleaning and Feature Selection

The dataset was preprocessed by removing missing data and performing feature engineering techniques to extract relevant features for predicting smoking status. The features used in the analysis were age, height, weight, waist circumference, eyesight, hearing, blood pressure (systolic and diastolic), fasting blood sugar, cholesterol (total, HDL, LDL), triglycerides, hemoglobin, urine protein, serum creatinine, AST, ALT, GTP, dental caries, and smoking status.

Pearson correlation-based feature selection was used to choose the most relevant attributes. With this method, the features with the highest correlation coefficients are chosen. It measures the linear correlation between the characteristics and the target variable. For further analysis, only the features with correlation coefficients higher than a predetermined threshold were maintained.

D. Model Training and Evaluation

We used 80% of featured data for training and 20% of the remaining for the testing the model using "train_test_split" module from sklearn. Logistic regression, random forests, and support vector machines machine learning algorithms were considered for predicting smoking status. The algorithms

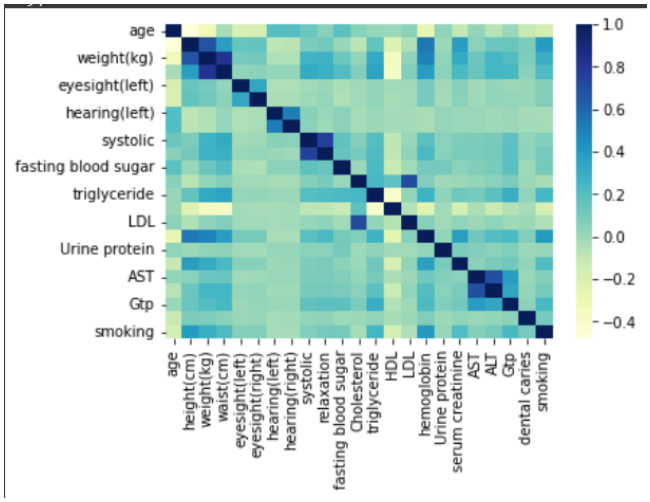

Fig. 1. heatmap of the correlations between all pairs of features

were evaluated based on their performance metrics, such as accuracy, precision, recall, and F1 score.

III. Machine Learning Algorithms

The following algorithms are selected fro predicting the smoking status using the bio-signals:

A. Logistic Regression

For binary classification issues like these, logistic regression is a viable option. Using the provided dataset, we can create a logistic regression model utilizing the bio-signals features as input and the "smoking" column as the dependent variable (i.e., the column we want to predict). Each input will have a probability value between 0 and 1, representing the likelihood that the input belongs to the positive class, according to the logistic regression model's output (i.e., the person is a smoker). The input can then be classified as positive (a smoker) or negative (a non-smoker) based on the anticipated likelihood using a threshold value, such as 0.5. This trained model can be evaluated its performance on the test data using evaluation metrics such as accuracy.

B. Support Vector Machines (SVMs)

We can train an SVM model on the given dataset, using the bio-signals features as input and the "smoking" column as the target variable. SVMs try to find a hyperplane that separates the positive and negative classes with the largest possible margin. We can use the trained SVM model to predict the smoker status of new input by checking which side of the hyperplane the input falls on.

SVM is a widely used classification algorithm that works by finding the hyperplane that maximally separates the two classes in the input data. We experimented with different kernels such as linear, polynomial, and radial basis function (RBF) to find the best-performing SVM model. We also used the grid search technique to optimize the hyperparameters of the SVM model.

C. Random Forest

Random forest is an ensemble learning method that combines multiple decision trees to make a final prediction. We can train a random forest model on the given dataset, using the bio-signals features as input and the "smoking" column as the target variable. The random forest model will output a binary prediction for each input, based on the majority vote of the decision trees. The final prediction will be either positive (smoker) or negative (non-smoker) based on the majority vote.

We trained a random forest classifier on the training data and evaluated its performance on the test data using accuracy as the evaluation metric. We also used the feature importance scores provided by the random forest to identify the most important bio-signals in predicting smoking status.

IV. Result and Analysis

Results and Analysis:

In this study, three machine learning algorithms were used to predict smoking status based on the given dataset - Logistic Regression, Random Forest Classifier, and Support Vector Machine (SVM). The accuracy scores of each algorithm were obtained and analyzed.

The Logistic Regression model achieved an accuracy of 72.4%. The confusion matrix shows that out of 7797 samples, 4013 were correctly classified as non-smokers while 962 were falsely classified as non-smokers, and 1191 were correctly classified as smokers while 1631 were falsely classified as non-smokers. The precision, recall, and f1-score were 0.77, 0.81, and 0.79 respectively for non-smokers, and 0.63, 0.58, and 0.60 respectively for smokers.

```
LogisticRegression
Accuracy: 0.723868154418366
Confusion matrix:
 [[4013  962]
 [1191 1631]]
Classification report:
              precision    recall  f1-score   support

           0       0.77      0.81      0.79      4975
           1       0.63      0.58      0.60      2822

    accuracy                           0.72      7797
   macro avg       0.70      0.69      0.70      7797
weighted avg       0.72      0.72      0.72      7797
```

Fig. 2. Results of Logistic Regression

The Random Forest Classifier model achieved an accuracy of 80.1% The confusion matrix shows that out of 7797 samples, 4169 were correctly classified as non-smokers while 806 were falsely classified as smokers, and 742 were correctly classified as smokers while 2080 were falsely classified as non-smokers. The precision, recall, and f1-score were 0.85, 0.84, and 0.84 respectively for non-smokers, and 0.72, 0.74, and 0.73 respectively for smokers.

The SVM model achieved an accuracy of 75.6%. The confusion matrix shows that out of 7797 samples, 3985 were correctly classified as non-smokers while 990 were falsely classified as smokers, and 914 were correctly classified as

```
RandomForestClassifier
Accuracy: 0.8014621008080031
Confusion matrix:
 [[4169  806]
 [ 742 2080]]
Classification report:
              precision    recall  f1-score   support

           0       0.85      0.84      0.84      4975
           1       0.72      0.74      0.73      2822

    accuracy                           0.80      7797
   macro avg       0.78      0.79      0.79      7797
weighted avg       0.80      0.80      0.80      7797
```

Fig. 3. Results of Random Forest Classifier

smokers while 1908 were falsely classified as non-smokers. The precision, recall, and f1-score were 0.81, 0.80, and 0.81 respectively for non-smokers, and 0.66, 0.68, and 0.67 respectively for smokers.

```
Support Vector Machines Accuracy: 0.7588816211363345
Support Vector Machines Confusion Matrix:
 [[4000  975]
 [ 905 1917]]
Support Vector Machines Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.80      0.81      4975
           1       0.66      0.68      0.67      2822

    accuracy                           0.76      7797
   macro avg       0.74      0.74      0.74      7797
weighted avg       0.76      0.76      0.76      7797
```

Fig. 4. Results of SVM model

Overall, the Random Forest Classifier model performed the best with the highest accuracy score of 80.1%. The precision, recall, and f1-score for both smokers and non-smokers were also relatively high compared to the other two models. Therefore, the Random Forest Classifier model can be considered as the most effective algorithm to predict the smoking status of an individual based on the given dataset.

V. Conclusion

In this study, we explored the use of machine learning algorithms for predicting smoking status based on bio-signals and demographic information. We used Logistic Regression, Random Forest, and Support Vector Machine (SVM) algorithms for prediction, and our results indicate that Random Forest performed the best, achieving an accuracy of 80.1

Our feature selection process using Pearson correlation revealed that certain bio-signals such as age, height, weight, waist circumference, and hemoglobin were highly correlated with smoking status as shown in fig.4 . This highlights the importance of these factors in determining smoking behavior and can be useful for developing targeted interventions to reduce smoking prevalence.

```
['smoking', 'hemoglobin', 'height(cm)', 'weight(kg)', 'triglyceride',
 'Gtp', 'waist(cm)', 'serum creatinine', 'HDL', 'age', 'dental caries',
 'relaxation', 'fasting blood sugar', 'ALT', 'systolic'],
dtype='object')
```

Fig. 5. top 15 features

While our study achieved promising results, there are limitations to our methodology. Firstly, our dataset was limited to a single population and may not generalize to other populations. Additionally, our study did not explore the causal relationship between bio-signals and smoking behavior, and further research is needed to better understand the mechanisms behind these relationships.

Overall, our study demonstrates the potential of machine learning for predicting smoking status based on bio-signals and provides valuable insights into the importance of certain factors in determining smoking behavior. Further research in this area has the potential to inform public health interventions and reduce smoking prevalence.

VI. Contribution

• I, Akhilesh Yadlapalli have performed the data collection, features selection and data pre-processing steps for the project and wrote the same sections in the report as well along with the introduction.

• I, Siddharth Reddy Aleti have implemented the models and analyzed the results thoroughly. I have written them in the report along with conclusions.

References

[1] kaggle dataset url. `https://www.kaggle.com/datasets/gau` Accessed: 20/01/2023.

[2] Anil Batra. Treatment of tobacco dependence. *Deutsches Ärzteblatt International*, 108(33):555, 2011.

[3] Sheldon Cohen, Edward Lichtenstein, James O Prochaska, Joseph S Rossi, Ellen R Gritz, Clifford R Carr, C Tracy Orleans, Victor J Schoenbach, Lois Biener, David Abrams, et al. Debunking myths about self-quitting: Evidence from 10 prospective studies of persons who attempt to quit smoking by themselves. *American Psychologist*, 44(11): 1355, 1989.

[4] World Health Organization. *Global health risks: mortality and burden of disease attributable to selected major risks*. World Health Organization, 2009.

[5] R Shashikant and P Chetankumar. Predictive model of cardiac arrest in smokers using machine learning technique based on heart rate variability parameter. *Applied Computing and Informatics*, (ahead-of-print), 2020.

[6] Saurabh Singh Thakur, Pradeep Poddar, and Ram Babu Roy. Real-time prediction of smoking activity using machine learning based multi-class classification model. *Multimedia Tools and Applications*, 81(10):14529–14551, 2022.