EC 9560: DATA MINING

INDIVIDUAL PROJECT

LAB 01

NAME :	ATPUTHARAVI.R
REG NO:	2020/E/015
DATE :	02 OCTOBER 2024

HEADING:

- Binary prediction of smoker status using bio-signals.

OBJECTIVE:

- The objective is developing an effective machine learning model predicting the smoking status of individuals based on biological signals, and more importantly, to try to explain how health indicators influence smoking behavior. Therefore, this project addresses the issue of enhancing smoking status prediction utilizing different machine learning techniques and feature selection methods.

METHODOLOGY:

1. Data collection
   - ➢ Leverage the available Kaggle dataset containing smoker status prediction based on bio-signals comprising many medical and physical attributes, which will focus on the classification of smoking behavior.

2. Data preprocessing
   - ➢ Handling missing or inconsistent values:
     - o In order to safeguard the accuracy and effectiveness of machine learning models, missing or inconsistent values in datasets should be dealt with in an appropriate manner. It is important to note that there are many possible explanations for the occurrence of absence or inaccessibility of data which may result in some important variable, for instance cholesterol levels, being left out. In this case, imputation strategies are to be employed. When discrete variables like eye sight have some missing data, the missing data can be filled with the most frequent mode value, but for continuous variables like weight or height, the mean or median value of that variable can be used. If it happens that a whole variable is very sparse, it can also be sensible to drop it completely. In the same line, if individual rows have too many missing values, such pieces of data may also be omitted from the whole dataset. This practice helps preserve the quality of the dataset and allows proper training of the model.

   - ➢ Normalization and standardization of continuous variables
     - o In the field of machine learning, especially in case of dealing with features with different ranges, scaling techniques such as normalization and standardization play a key role as a preprocessing step for continuous variables. Continuous variables like age, weight, height, blood pressure, fasting blood glucose, or cholesterol level often exhibit vastly differing scales, for instance, many obese persons weight between 40 to 100 kgs,

and the cholesterol levels may vary from 100 to 300 mg/dl. In machine learning model training especially with stage of performance gradient-based algorithms performance may be a lot higher after the data is scaled. This is due to the fact that features that are not scaled can lead to the learning process being bias by certain features resulting in poor prediction quality and making the model inefficient.

  - Normalization is one mainstream method of scaling data. Normalization is the method to scale any data between 0 and 1. This is particularly useful in cases where it is desired that the data should keep the same proportions rather than being drawn from a normal distribution. The formula for normalization is:

$$x' = \frac{x - \mathrm{xmin}}{xmax - xmin}$$

  - Where, x' is called as a normalized value, which is based upon original value x, and x min is the containing minimum feature value and the x max is the containing maximum feature value. On the contrary, standardization is a different technique which converts the original data with mean equal to zero and a standard deviation of one. This is especially useful when the data is assumed to be Gaussian (normal) as it allows the data to be distributed around zero resulting in more consistency in convergence during training. The standardization formula is given by:

$$z = \frac{x - \mu}{\sigma}$$

  - Here, z is called the transformed variable, where x is the initial variable, μ is the average of the variable feature, and σ is variance. By using these transformations, it can be assured that no single feature dominates the model aiding in speeding up the training process and improving the performance of the model on different machine learning tasks.

- Converting categorical features and target variable into numeric formats
  - Converting features and targets in the categorical variables and their values into numeric formats is extremely important for algorithms used for machine learning, as they usually process only numerical data. For categorical features of attributes such as eyesight and hearing, two encoding methods are usually employed. Label encoding is the process of converting categorical values into numeric labels, for instance eyesight values can be 0, 1, 2 for different eyesight abilities. One-hot encoding instead is applied on categorical variables that are not ranked in a hierarchical order, for example, eyesight classification in the levels of: good, fair and poor, where each category is represented by a creation of columns

of zero and ones: good has three columns as [1 0 0], fair has equidistant columns [0 1 0] and poor has [0 0 1] encodings. Even in the case of the target variable relating to a person's smoking status which is binary by nature, simple encoding may reduce the statuses into 0 for non-smoker and 1 for a smoker. These encoding strategies help provide a necessary representation of categorical data useful in the training and prediction of the models.

3. Exploratory data analysis (EDA)
   - The preliminary Analysis of Data is an integral process in data analysis and focuses on the description of the essential features of the data set, usually in a visual way. Its purpose is to provide a clear insight into the data that will assist in the following analyses and modeling choices. One of the steps in EDA is Distribution Analysis, which concentrates on studying the spread, center, and shape of the dataset in-depth. The distribution of the data is demonstrated using tools including histograms, density plots, box plots, which portray different normality, skewness, and outliers. This first phase of exploration in turn allows the data analyst to calculate the necessary statistical techniques and model to implement depending on the type of data available.

   - Another important part of EDA is looking at the features and target variables and how they are related, since this helps in observing the features that are predictive. Pair plots, scatter plots added with box plots help to show the interaction of particular features and their relation to the smoking status. Moreover, correlation analysis is also an important factor in feature selection because the gaps between the target variable and the features should be minimized. This can be done by using the Pearson correlation coefficient and its heat maps which help in exposure of the possible target variables. With all the work of EDA, an understanding of the data is acquired, especially its network and qualities, which controls the process of data cleansing, selecting the indicators of relevance and building the algorithm. At last, Exploratory Data Analysis or EDA is the foundation stone of any productive machine learning modeling activity as it explores the patterns, relationships, and oddities present in the data prior to its fitting.

4. Feature selection
   - Feature selection warrants importance in the machine learning process in that its aim is to determine the most relevant variables that enhance the derived model. Common methods used to rank and evaluate features based on the relevance include Recursive Feature Elimination (RFE), Lasso Regression, and tree based feature importance. RFE involves elimination of the least significant features in the

course of the model building so as to keep only those that elevate the performance of the model, while Lasso Regression inserts a tax on the inclusion of insignificant predictors, rendering their estimates as zero. Moreover, Random Forest or Gradient Boosting tree methods provide feature importance by looking at the contribution of the feature in decreasing the impurity of the given model. In the end after these procedures has been executed, useful features can also be analyzed concerning the model predictiveness, which further facilitates the construction of a better and a less complex model thus minimizing the effects of overfitting, as well as the computational costs.

5. Modeling

   - During the modeling process, many machine learning algorithms including but not limited to Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machines are implemented to forecast the target variable. Each model is then subjected to a hyperparameter search using cross validation to improve its performance and ensure that it can generalize well to fresh out of sample data. To improve the intelligibility of these machine learners, explanatory techniques such as SHAP or LIME are used to visualize the importance of individual features within the model and their contribution to the model's predictions. This procedure of applying advanced models combined with the frameworks for explaining them allows creating precise and reasonable models, which may support the process of decision-making – and that is all along the line of the objectives of this study.

6. Model evaluation
   - Regression techniques, utilized in the evaluation phase, commonly referred to as performance metrics, are used to evaluate the model effectiveness in predicting outcomes. Accuracy determines the global correctness of the model by taking the total number of correctly predicted instances and dividing it by the total number of instances – correctness ratio. Precision defines the model’s contingency due to true positive cases only and provides the quality of the positive predictions made by the model. Recall or sensitivity explains how well the model can find all the relevant data and examines one aspect of relevance i.e., the proportion of true positives in all the positive cases. The F1-score measures precision and recall and brings them together in one figure, which is especially important for models dealing with unbalanced classes. The performance of the model is also evaluated in a specific scoring system called ROC-AUC (Receiver Operating Characteristic - Area Under Curve) scores the ability of the model to separate the positive and negative classes at different thresholds such as 0.2 and 0.3 as well as in between such thresholds providing the overall model performance in different classification instances. All these metrics hence contribute together to understand the strengths and weaknesses of the model in detail.

DATA DESCRIPTION WITH A LINK TO DATA IN DATA REPOSITORY:

- The dataset consists of bio-signals and medical measurements for each individual, with the target variable being their smoking status. The feature include:

  age: 5-years gap
  height(cm)
  weight(kg)
  waist(cm) : Waist circumference length
  eyesight(left)
  eyesight(right)
  hearing(left)
  hearing(right)
  systolic : Blood pressure
  relaxation : Blood pressure
  fasting blood sugar
  Cholesterol : total
  triglyceride
  HDL : cholesterol type
  LDL : cholesterol type
  hemoglobin
  Urine protein
  serum creatinine
  AST : glutamic oxaloacetic transaminase type
  ALT : glutamic oxaloacetic transaminase type
  Gtp : γ-GTP
  dental caries
  smoking status (Target variable)

- Dataset from Kaggle: https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction-using-biosignals/data