Bio-Signal Analysis for Smoking Prediction

INFO7390 Fall 2023

SHRUTI TAMBE - 002762916

Outline

- Dataset: https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking
- Goal: The goal is to determine the presence or absence of smoking (smoker/non-smoker) through bio-signals.
- Result: After thorough analysis, Random Forest Classifier exhibited the capability to predict smoking status based on bio-analysis with an accuracy rate of 84%

Tools used

- Python
- Google Colab
- Libraries: Scikit-Learn, matplotlib, seaborn,Pandas, NumPy
- Grid Search and Hyperparameter Tuning
- Data Visualization(EDA)
- Machine learning: Random forest classifier, Logistic Regression, SVM

Dataset: 55k rows, 26 features

- Biometric features - age, gender, weight, height, hemoglobin levels etc
- Understand the relationship between biometric features and smoking status (smoker/non-smoker)

Shape of the dataset

```
[3] smoke.shape

(55692, 27)
```

```
smoke = pd.read_csv("smoking.csv")
smoke.head()
```

	ID	gender	age	height(cm)	weight(kg)	waist(cm)	eyesight(left)	eyesight(right)	hearing(left)	hearing(right)	...	hemoglobin	Urine protein	serum creatinine	AST	ALT	Gtp	oral	denta carie
0	0	F	40	155	60	81.3	1.2	1.0	1.0	1.0	...	12.9	1.0	0.7	18.0	19.0	27.0	Y	
1	1	F	40	160	60	81.0	0.8	0.6	1.0	1.0	...	12.7	1.0	0.6	22.0	19.0	18.0	Y	
2	2	M	55	170	60	80.0	0.8	0.8	1.0	1.0	...	15.8	1.0	1.0	21.0	16.0	22.0	Y	
3	3	M	40	165	70	88.0	1.5	1.5	1.0	1.0	...	14.7	1.0	1.0	19.0	26.0	18.0	Y	
4	4	F	40	155	60	86.0	1.0	1.0	1.0	1.0	...	12.5	1.0	0.6	16.0	14.0	22.0	Y	

5 rows × 27 columns

Methodology : Data Pre-processing

Dropping 'id','oral' column and target column 'smoking'. oral has only one value, so it is not relevant.

```
[4] df = smoke.drop(columns=['ID','oral','smoking'], axis=1)
    df.head()
```

	gender	age	height(cm)	weight(kg)	waist(cm)	eyesight(left)	eyesight(right)	hearing(left)	hearing(right)	systolic	...	HDL	LDL	hemoglobin	Urine protein	serum creatinine	AST	ALT
0	F	40	155	60	81.3	1.2	1.0	1.0	1.0	114.0	...	73.0	126.0	12.9	1.0	0.7	18.0	19.0
1	F	40	160	60	81.0	0.8	0.6	1.0	1.0	119.0	...	42.0	127.0	12.7	1.0	0.6	22.0	19.0
2	M	55	170	60	80.0	0.8	0.8	1.0	1.0	138.0	...	55.0	151.0	15.8	1.0	1.0	21.0	16.0
3	M	40	165	70	88.0	1.5	1.5	1.0	1.0	100.0	...	45.0	226.0	14.7	1.0	1.0	19.0	26.0
4	F	40	155	60	86.0	1.0	1.0	1.0	1.0	120.0	...	62.0	107.0	12.5	1.0	0.6	16.0	14.0

5 rows × 24 columns

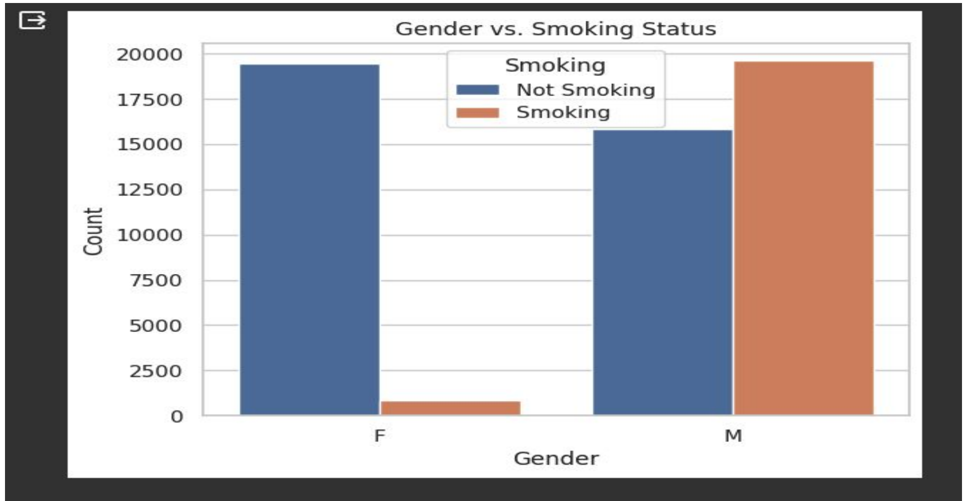
Renaming column names:

```
[5] df.rename(columns = {"height(cm)" : "height_cm", "weight(kg)" : "weight_kg",
                         "waist(cm)" : "waist_cm", "eyesight(left)" : "eyesight_left",
                         "eyesight(right)" : "eyesight_right", "hearing(left)" : "hearing_left",
                         "hearing(right)" : "hearing_right", "fasting blood sugar" : "fasting_blood_sugar",
                         "Cholesterol" : "cholesterol", "HDL" : "hdl", "LDL" : "ldl",
                         "Urine protein" : "urine_protein", "serum creatinine" : "serum_creatinine",
                         "AST" : "ast", "ALT" : "alt", "Gtp" : "gtp", "dental caries" : "dental_caries"},
          inplace = True)
```

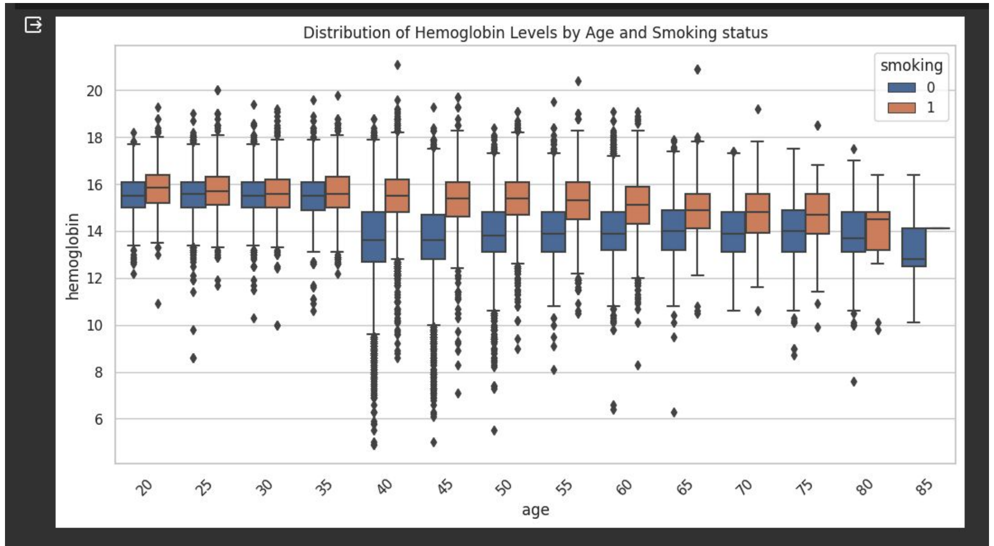
Methodology :

Exploratory Data Analysis

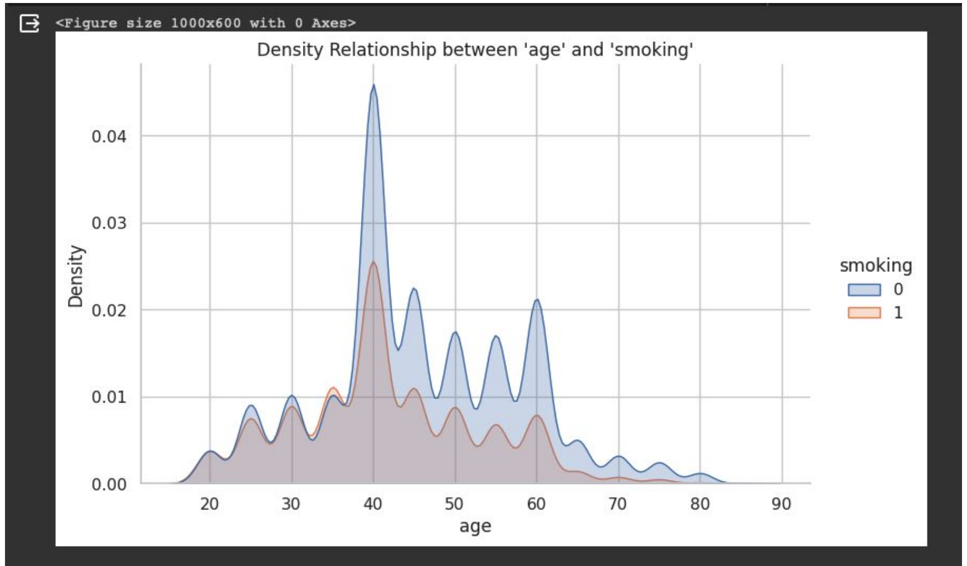
- Smokers exhibit significantly elevated hemoglobin levels
- Smoking is more prevalent among younger individuals
- The male population has a higher representation among smokers compared to females



Observation: Males smokers are significantly more



Observation: Hemoglobin levels are notably higher in individuals who smoke. This increase among smokers may be a compensatory response to address a potential reduction in oxygen-carrying capacity.



Observation: People smoke less as their age increases.

Methodology: Feature Engineering

- Encoding categorical features
- Creating new features
- Feature Importance Analysis
- Scaling the dataset

Creating new features to reduce the number of columns and combining features that are very highly correlated to each other.

```
df_encoded['average_eyesight'] = (df_encoded['eyesight_left'] + df_encoded['eyesight_right']) / 2
```

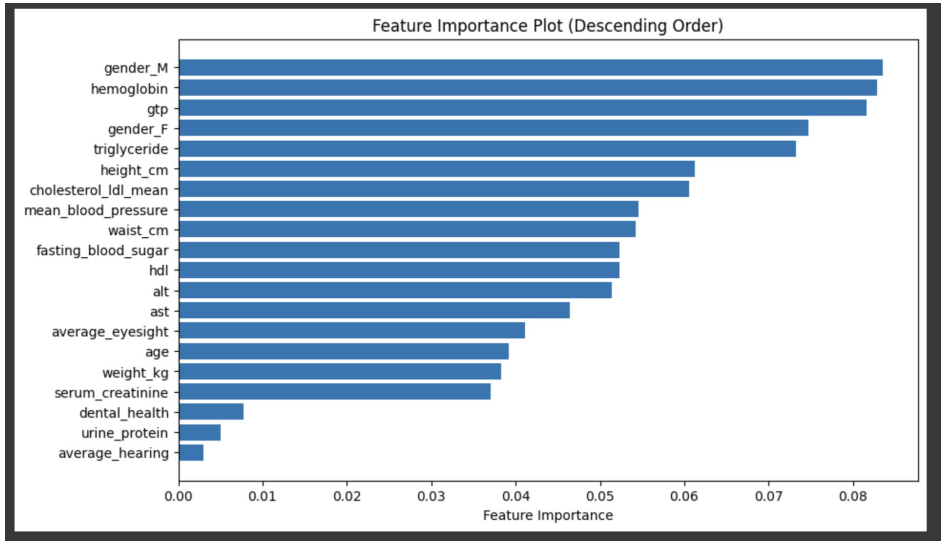
```
df_encoded['average_hearing'] = (df_encoded['hearing_left'] + df_encoded['hearing_right']) / 2
```

```
[16] df_encoded['dental_health'] = df_encoded['dental_caries'] + df_encoded['tartar_N']+ df_encoded['tartar_Y']
```

```
[17] df_encoded['cholesterol_ldl_mean'] = (df_encoded['cholesterol'] + df_encoded['ldl']) / 2
```

```
[18] df_encoded['mean_blood_pressure'] = (df_encoded['relaxation'] + df_encoded['systolic']) / 2
```

```
[19] df_new = df_encoded.drop(columns=['eyesight_left','eyesight_right','hearing_left','hearing_right','dental_caries','tartar_N','tartar_Y','cholesterol
```



Encoding categorical features

```
[12] categorical_cols = ['gender','tartar']

     df_encoded = pd.get_dummies(df, columns=categorical_cols)
     df_encoded.head()
```

	age	height_cm	weight_kg	waist_cm	eyesight_left	eyesight_right	hearing_left	hearing_right	systolic	relaxation	...	urine_protein	serum_creat
0	40	155	60	81.3	1.2	1.0	1.0	1.0	114.0	73.0	...	1.0	
1	40	160	60	81.0	0.8	0.6	1.0	1.0	119.0	70.0	...	1.0	
2	55	170	60	80.0	0.8	0.8	1.0	1.0	138.0	86.0	...	1.0	
3	40	165	70	88.0	1.5	1.5	1.0	1.0	100.0	60.0	...	1.0	
4	40	155	60	86.0	1.0	1.0	1.0	1.0	120.0	74.0	...	1.0	

5 rows × 26 columns

Methodology: Model Selection and Development

1. Baseline Models- Logistic Regression, SVM, Random Forest Classifier
2. Fit the model with Grid Search and Cross Validation Techniques
3. RFC gave the highest accuracy rate of 83%
4. Hyperparameter tuned Random Forest to increase the accuracy
5. Achieved 84% accuracy over the baseline model

Results

RFC gave superior accuracy over the base-models, 84%

```
Logistic Regression Model:
Accuracy: 0.75
              precision    recall  f1-score   support

           0       0.80      0.79      0.80      6861
           1       0.66      0.68      0.67      4155

    accuracy                           0.75     11016
   macro avg       0.73      0.73      0.73     11016
weighted avg       0.75      0.75      0.75     11016

SVM Model:
Accuracy: 0.76
              precision    recall  f1-score   support

           0       0.82      0.79      0.81      6861
           1       0.68      0.72      0.70      4155

    accuracy                           0.76     11016
   macro avg       0.75      0.75      0.75     11016
weighted avg       0.77      0.76      0.76     11016

Random Forest Model:
Accuracy: 0.83
              precision    recall  f1-score   support

           0       0.87      0.86      0.87      6861
           1       0.77      0.79      0.78      4155

    accuracy                           0.83     11016
   macro avg       0.82      0.82      0.82     11016
weighted avg       0.83      0.83      0.83     11016
```

```
Random Forest Classifier (Hyperparameter Tuned):
Best Hyperparameters: {'max_depth': 30, 'n_estimators': 200}
Accuracy: 0.84
              precision    recall  f1-score   support

           0       0.88      0.86      0.87      6861
           1       0.77      0.80      0.79      4155

    accuracy                           0.84     11016
   macro avg       0.83      0.83      0.83     11016
weighted avg       0.84      0.84      0.84     11016
```

Conclusion

- Achieved 84% accuracy to predict whether a person is a smoker or non-smoker based on bi-signals
- Future work: track changes in smoking behaviors and health indicators over time