﻿**DSAA 5002 - Data Mining and Knowledge Discovery in Data Science**

**Final Exam Report – Q5 ﻿Smoke Status Recognition**

**50015940 Jiaxiang Gao**

1. **Data Preprocessing:**
   1. The 'id' column is removed from the training dataset as it's not relevant for prediction.
   2. Missing values in various columns are handled:

* For 'height(cm)' and 'waist(cm)', missing values are filled with the mean of the respective columns.
* For 'eyesight(left)' and 'eyesight(right)', missing values in one eye are replaced with values from the other eye.
* For 'hearing(right)', missing values are replaced with values from 'hearing(left)'.
* Missing values in 'Urine protein' are filled with the mean.
  1. Duplicate rows in the training dataset are identified and removed.

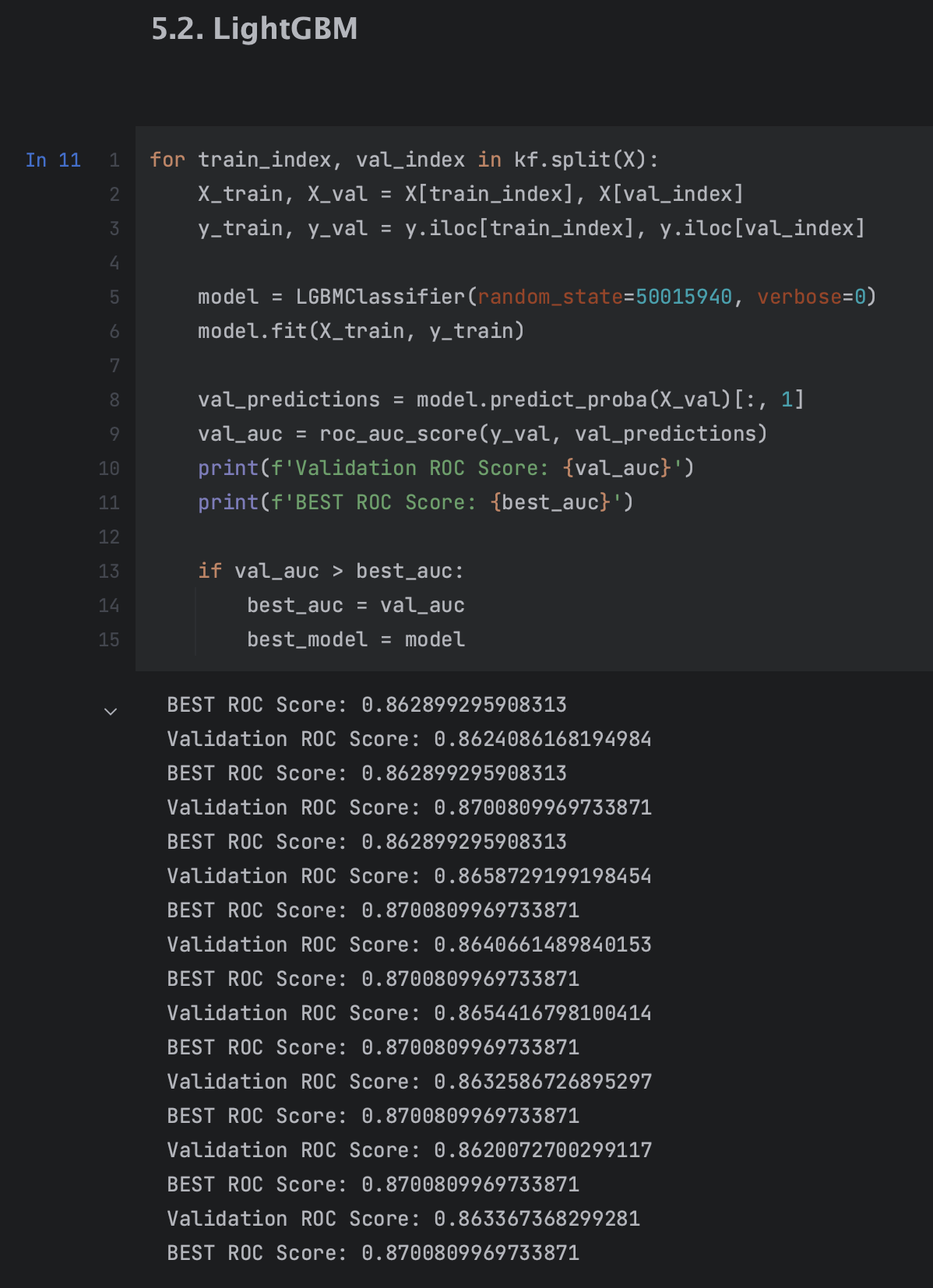
1. **Model Training and Validation:**
   1. The data is scaled using StandardScaler to normalize feature values.
   2. The K-Fold cross-validation approach (with 10 splits) is applied to validate the model's performance.
   3. Two types of models are trained and validated:
      1. LGBMClassifier: A Light Gradient Boosting Machine classifier.
      2. CatBoostClassifier: A classifier from the CatBoost framework.
   4. In each fold of the cross-validation:

* The model is trained on the training subset.
* The model's performance is evaluated on the validation subset using the ROC AUC score.
* The best model is updated if the current model's ROC AUC score is higher than the previously recorded best score.

1. **Final Model Prediction:**

The model with the highest ROC AUC score from the cross-validation process is selected as the best model.

This best model is used to predict the 'smoking' variable on the test dataset.

文本

描述已自动生成