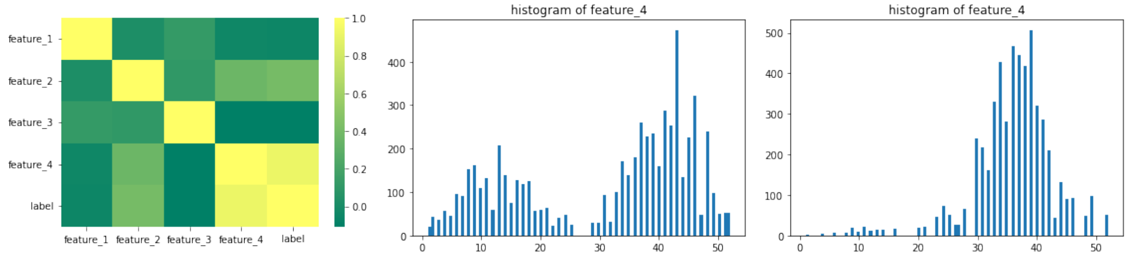
MINI-PROJECT REPORT

Author: Huihan YANG； Student ID: 120090438

Github repo-link: https://github.com/foxintohumanbeing/cuhksz_dda4210_miniproject.git

- Clarify: We mainly focus on the data from training dataset and augmented dataset. The discussion of tesing dataset used by Kaggle will be in the last section.

1. Data Analysis



Left:Correlation Heatmap of Training Data.Center:Feature4 Histogrm of Training Data.Right: Feature4 Histogrm of Augmented Data

- The heatmap of training data shows that feature 4 is the most siginificant factor in classification.
- The histogram of feature4 from train data and augmented data shows that they have similar distribution pattern (mainly distributed around some values).

2. Binary Classification Algorithm

2.1 Algorithm: Majority vote of nine small classifiers

2.1.1 Small Classifiers and Their Weights:

Classifier_index	**1**	**2**	**3**	**4**	**5**
Classifier_name	Random Forest1	Adaboosting	SVC	Naive Bayes	Random Forest2
Weight	4	1	2	1	1

2.1.2 Majority Vote

Denote the output label given by classifier j for a sample i is l_{ij}, corresponding weight is w_{ij} and the final prediction is l_i.

$$l_i = l_{ia}, \qquad where\ a = \arg\max_{a\in\{1,2,3,4,5\}} \Sigma_{j=1}^{5} w_{ij}\mathbf{1}_{l_{ij}=l_{ia}}, \qquad \mathbf{1}_{l_{ij}=l_{ia}} = \begin{cases}1, l_{ij} = l_{ia}\\ 0, l_{ij} \neq l_{ia}\end{cases} \tag{1}$$

2.2 Key Components and Rational

1. Random Forest1 - Parameter: num_{tree}=100, $num_{fea_{max}}$ = 2
   - **Key Algorithm**: Build many trees with limited number of features. Take average or majority to reduce variance.
   - **Role and Rationale**: **Major Classifier**
     - As noted before, feature4 is the most important factor. RF implicitly select important factors and

amplify their influences.
    - One random forest could reach the accuracy around 98%.
    - However, it fails to reduce tree structure's bias. To fix it, we use other classifier to modify.

2. Adaboosting - Parameter: Nan
    - **Key Algorithm**: Sequentially build one tree with slow learning rate using adaptive gradient
    - **Role and Rationale**: Reduces tree structure‘s bias.
3. SVC -Parameter: Kernel = 'rbf', C = 8
    - **Key Algorithm**: Project data on higher dimension to seperate with penalty.
    - **Role and Rationale**: Find more information of feature's correlation in high dimension. Reduce the influences of outliers using penalty term.
4. Naive Bayes - Parameter: Prior = Gaussian
    - **Key Algorithm**: Maximize feature's likelihood with features' prior assumption as Gaussian.
    - **Role and Rationale**: Helps to deal with unbalanced and limited size data.
5. Random Forest2: - Parameter: num_{tree}=100, $num_{fea_{max}}$ = 3
    - **Key Algorithm**: same as RF1
    - **Role and Rationale**: Increase number of maximum number of features. Dig out more information missed by RF1.

3. Results

- We tried different ensamble models. For each different model, we run for 30 times to see the average accuracy and variance.

	RF	Boosting	SVM	Gaussian Bayes	Ensemble RF	SVM-boosting	Vote-boosting
Mean	0.9922	0.9913	0.9882	0.9877	0.9926	0.9891	0.9933
Variance	0.0345	6.04E-06	0.0728	0.0870	0.0343	0.0583	0.0153

- To avoid the problem of underflow, we multiply the accuracy with 100 when calculating variance.
- It is worth to note that Vote-boosting has highest mean accuracy and lowest accuracy variance.
- This results highlight its stability and generalization.

4. Algorithm used on Kaggle