Homework 3 - Behavioral Data Science
Kelly Modi, Dev Dwivedi
Github Repository Link

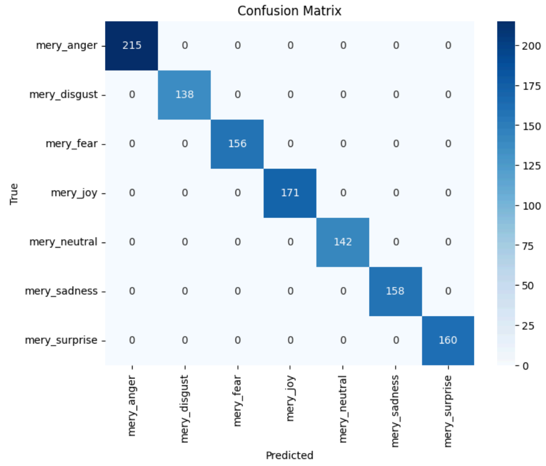
Problem 1 (5 pts):

1. False - 1x1 filters only operate at the same spatial location compared to 3x3 filters which spread to neighboring pixels.
2. True
3. False - Decision trees are not *always* more interpretable because a distilled model with a simple structure can be more easier to understand than a complex decision tree.
4. False - If the teacher model is overfit to the data and the student model has more simplicity and generalizes better, then there's a chance it can outperform the teacher model.
5. True
6. False - A latent variable model marginalizes out the latent variable (x), not the data (z).
7. True
8. False - PCA is deterministic with no probabilistic model (therefore doesn't assume Gaussian).
9. False - PCA relies on eigendecomposition of the empirical covariance matrix, not data matrix.
10. False – PCA and FA used for different purposes as PCA is used for modeling psychological constructs (e.g. personality) while FA is used for dimensionality reduction and feature extraction.

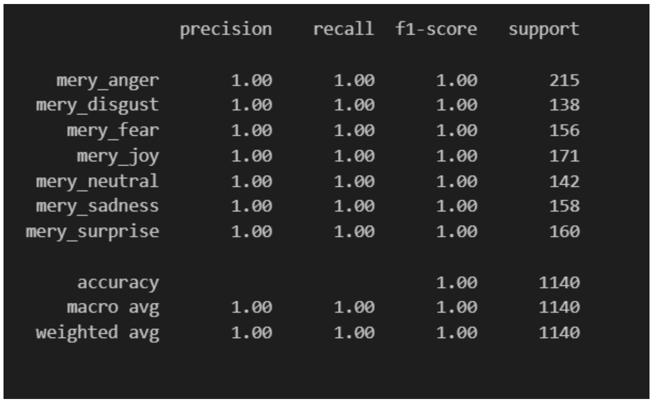
Problem 2 (8 pts):

Code attached in GitHub repository link.

Confusion Matrix:



Analysis Results:

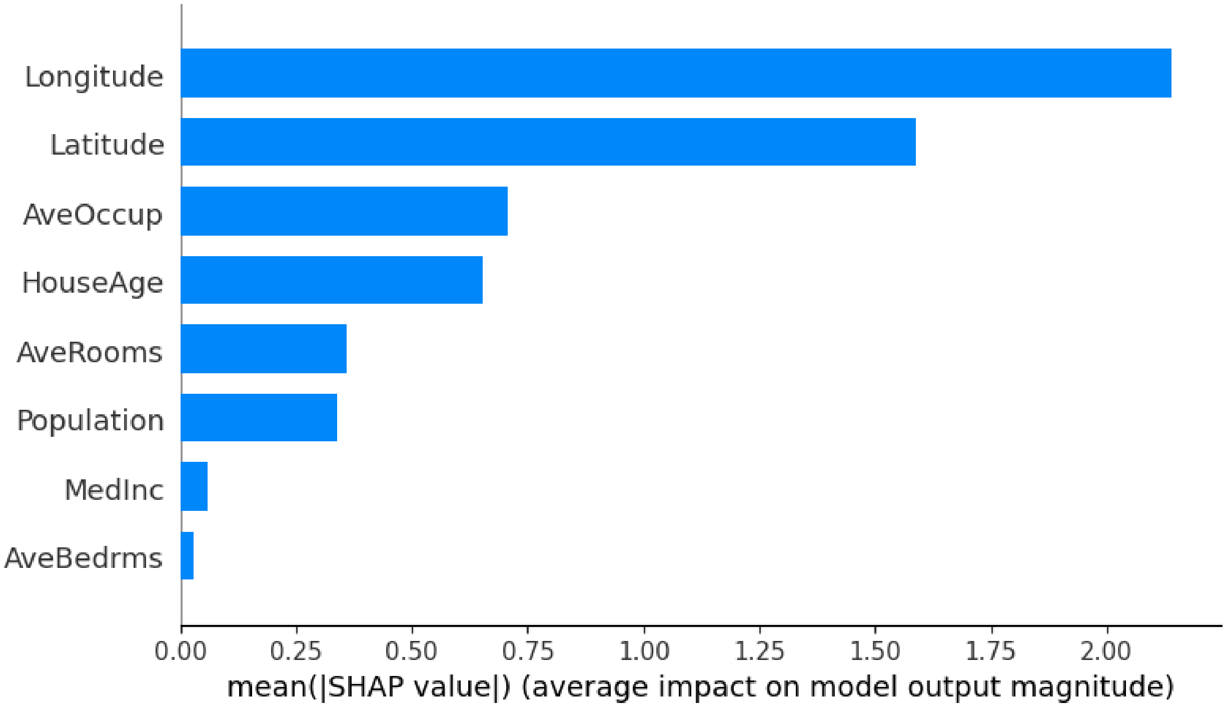


```
               precision    recall  f1-score   support

   mery_anger       1.00      1.00      1.00       215
 mery_disgust       1.00      1.00      1.00       138
    mery_fear       1.00      1.00      1.00       156
     mery_joy       1.00      1.00      1.00       171
 mery_neutral       1.00      1.00      1.00       142
 mery_sadness       1.00      1.00      1.00       158
mery_surprise       1.00      1.00      1.00       160

     accuracy                           1.00      1140
    macro avg       1.00      1.00      1.00      1140
 weighted avg       1.00      1.00      1.00      1140
```

Problem 3 (8 pts):

Permutation Importance vs. Shapley Ratios Table		
Aspect	***Permutation Importance***	***Shapley Ratios***
Definition	Used to determine a feature's importance by disrupting its relationship with the outcome	Used to inform us about how to fairly distribute predictions among features in a model by accounting for the number of possible feature orderings to normalize marginal contribution
Formula	$I^{(j)} = \frac{1}{N}\sum_{n=1}^{N} L\big(f(\mathbf{x}_n), y_n\big) - L\big(f(\tilde{\mathbf{x}}_n^{(j)}), y_n\big)$	$\phi_j = \sum_{S\subseteq F\setminus\{j\}} \frac{\lvert S\rvert!(\lvert F\rvert - \lvert S\rvert - 1)!}{\lvert F\rvert!}[v(S\cup\{j\}) - v(S)]$
Interpretation	Larger I_j indicates a greater reliance on feature j	Φ_j is the fair contribution of feature j to a single prediction
Feature Interaction	Ignores feature interactions	Accounts for all feature interactions through subsets/coalitions
Computation Complexity	Simple and fast	Exact computation often infeasible (NP-hard) so approximation of Shapley values is done using Monte Carlo sampling
Correlation Problem	Sensitive to feature correlations	More robust to correlations because marginal contributions are averaged across all subsets/coalitions
Relation to Perturbation	Example of perturbation method	Rooted in game theory

DeepExplainer applied on California data set (Code attached in GitHub repository link):



3 most important features: 1) longitude, 3) latitude, and 3) average occupants

Problem 4 (10 pts):

a) Derivation of MSE loss ~~function~~ regression from Gaussian Likelihood:

Assumption: $y_n \sim \mathcal{N}(f(x_n;\theta), \sigma^2)$

PDF: $p(y_n | x_n; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_n - f(x_n;\theta))^2}{2\sigma^2} \right)$

Total Likelihood: $p(D|\theta) = \prod_{n=1}^{N} p(y_n | x_n; \theta)$

Log-likelihood: $\log p(D|\theta) = \sum_{n=1}^{N} \log p(y_n|x_n;\theta) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - f(x_n;\theta))^2$

Negative log-likelihood $= NLL = \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - f(x_n;\theta))^2 + C$; C = constant

$$\downarrow$$

$$\boxed{MSE(\theta) = \frac{1}{N}\sum_{n=1}^{N}(y_n - f(x_n;\theta))^2}$$

↑ dropped constant

b) Derivation of BCE loss classification from Bernoulli Likelihood:

Assumption: $y_n \sim \text{Bernoulli}(f(x_n;\theta))$

PMF: $p(y_n|x_n;\theta) = f(x_n;\theta)^{y_n}(1 - f(x_n;\theta))^{1-y_n}$

Total Likelihood: $p(D|\theta) = \prod_{n=1}^{N} p(y_n|x_n;\theta)$

Log-likelihood: $\log p(D|\theta) = \sum_{n=1}^{N}\left[y_n \log f(x_n;\theta) + (1-y_n)\log(1 - f(x_n;\theta))\right]$

Negative Log-likelihood: $NLL = -\sum_{n=1}^{N}\left[y_n \log f(x_n;\theta) + (1-y_n)\log(1-f(x_n;\theta))\right]$

$$\boxed{BCE(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log f(x_n;\theta) + (1-y_n)\log(1-f(x_n;\theta))\right]}$$

Bonus Question:

c) Heteroskedastic Loss Bonus

Assume: $y_n \sim N(f(x_n;\theta), \sigma^2(x_n;\theta))$

$$p(y_n|x_n;\theta) = \frac{1}{\sqrt{2\pi\sigma^2(x_n;\theta)}} \exp\left(-\frac{(y_n - f(x_n;\theta))^2}{2\sigma^2(x_n;\theta)}\right)$$

* Intermediate steps →

$$\text{NLL} = \sum_{n=1}^{N}\left[\frac{(y_n - f(x_n;\theta))^2}{2\sigma^2(x_n;\theta)} + \frac{1}{2}\log\sigma^2(x_n;\theta)\right]$$

$$\boxed{\mathcal{L}(\theta) = \sum_{n=1}^{N}\left[\frac{(y_n - f(x_n;\theta))^2}{2\sigma^2(x_n;\theta)} + \frac{1}{2}\log\sigma^2(x_n;\theta)\right]}$$

$$-\log p(y_n|x_n;\theta) = \frac{1}{2}\log(2\pi) + \frac{1}{2}\log\sigma^2(x_n;\theta) + \left(\frac{(y_n - f(x_n;\theta))^2}{2\sigma^2(x_n;\theta)}\right)$$

↑

dropped constant

Problem 5 (12 pts):

Code attached in GitHub repository link.

Evaluating PCA and Factor Analysis:

Do items belonging to the same subscale load highly on a single factor while showing minimal loadings on others?

Factor analysis demonstrated a variation in these results. For example, items from subscale C load coherently on one factor while subscale D demonstrated a split pattern as most items load moderately on Factor 2, but D3, D6, and D8 have low loadings across all factors. Similarly, subscale I shows moderate loadings for I1-I4 but weak loadings for I5, I6, and I10. Subscale L is largely cohesive except for L6 and in subscale P, only P1, P2, and P5 load strongly while several other items are loading weakly. Thus, while some subscales demonstrate acceptable cohesion, others include items that do not load clearly onto a single factor.

Which factor explains the most variance?

The PCA results reveal that Principal Component 1 explains the most variance (approximately 22.78%). In the factor analysis, Factor 1 is dominant, as many items (C, I, L, and parts of P) from the selected personality variables (C, D, I, L, and P) load strongly on this factor. This suggests that Factor 1 captures a broad, underlying dimension across the selected items.

Which items should be removed due to low loadings?

Based on the analysis, items with consistently low loadings across all factors are candidates for removal. Specifically, within subscale D, items D3, D6, and D8 show weak loadings. In subscale I, items I5, I6, and I10 show weak loadings. In subscale L, item L6 shows weak loading. In subscale P, items P3, P4, P6, P7, P9, and P10 show weak loadings. Removing these items could refine the model by enhancing the clarity and internal consistency of each subscale.