AI Assignment 4 report

Preprocessing

We have 20,000 samples with 38 input attributes, 1 target attribute with 34 classes.
2. No null values were found
3. I checked The correlation matrix shows none of the attributes was highly correlated
4. I combined the Job roles into different classes are, those were a total of 9.
5. I scaled the numerical data using Standard Scaler
6. The textual data is encoded using label encoding for columns with 2/3 unique values and one-hot encoding for columns with more unique values.
The output layer uses the 'softmax' function which outputs 9 values, that is the probabilities of the sample belonging to each class.
Categorical cross-entropy is used as the loss function and the model trains on improving accuracy.

Accuracy

1. 70:30
Training accuracy - 39.45 %
Testing accuracy - 15.83 %
2. 60:40
Training accuracy - 42.08 %
Testing accuracy - 15.39 %
3. 90:10
Training accuracy - 35.68 %
Testing accuracy - 14.75 %
4. No shuffling (70:30)
Training accuracy - 38.86 %
Testing accuracy - 15.48 %
5. Only numerical data (70:30)
Training accuracy - 31.39 %
Testing accuracy - 16.60 %