Artificial Neural Network (CT-466)

Complex Computing Problem (C.C.P)

Name: Syed Muhammad Abdullah

Roll no: AI - 22302

Section: ArtificiaI Intelligence (A.I.)

Batch: 2022

Department: Computer Science and Information Technology

Part A — Classification (Supervised Learning)

Objective

To classify whether annual income of an individual exceeds $50K/year based on Census Data.

Methodology

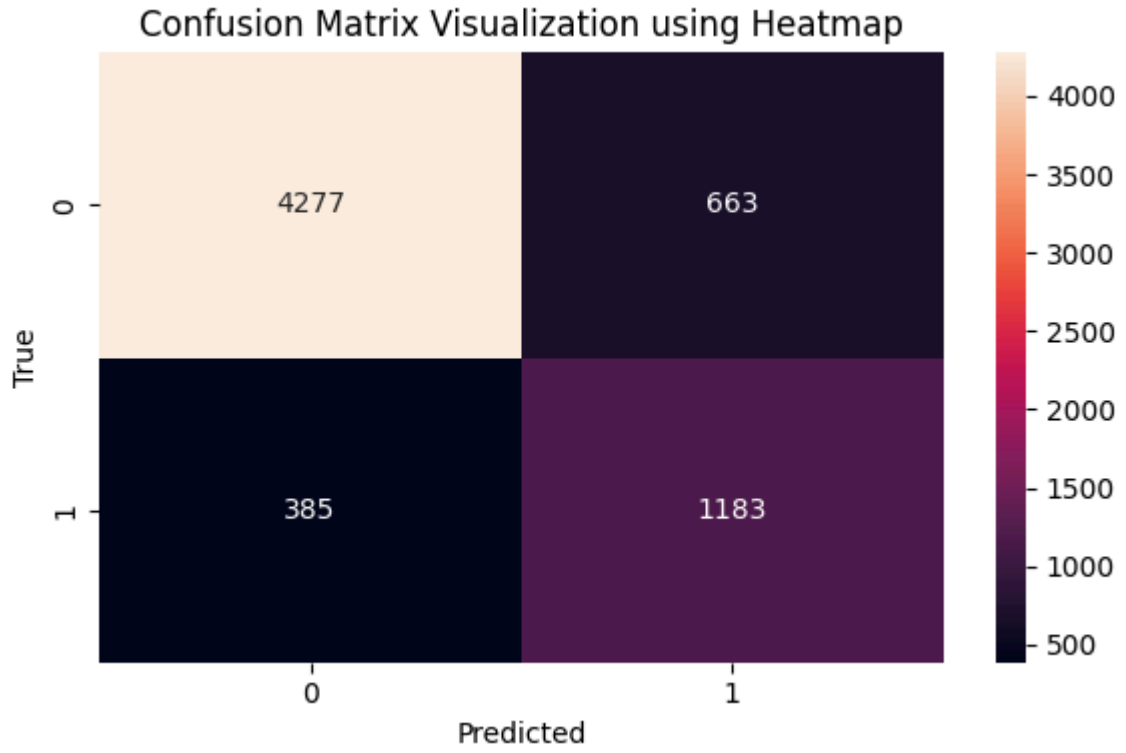
- **Dataset Collection:** The dataset has been collected from Kaggle named Adult Census Income. The shape of the dataset and features present in this dataset are: Shape of the dataset: (32561, 15) Columns in the dataset: ['age', 'workclass', 'fnlwgt', 'education', 'education.num', 'marital.status', 'occupation', 'relationship', 'race', 'sex', 'capital.gain', 'capital.loss', 'hours.per.week', 'native.country','income']
- **Handling Custom Missing Values:** Identified and replaced the placeholder string '?' with NumPy missing value representation (np.nan) across the entire dataset.
- **Duplicate Removal:** Removed all duplicate rows from the dataset.
- **Missing Value Imputation:** Handled missing values in categorical columns (workclass, occupation, native.country) using **Mode imputation** (filling values with the most frequent category).
- **Feature Encoding:** Converted all categorical features into a machine-readable format using **One-Hot Encoding** (pd.get_dummies), dropping the first category to avoid multicollinearity.
- **Feature correlation visualization:** Calculated the correlation matrix between all encoded features and the target variable and plotted it.
- **Data Splitting and Scaling:** The data was split into **training (80%) and testing (20%) sets**, ensuring stratification based on the target variable (income_>50K). Features were standardized using StandardScaler, which was fitted only on the training data and then used to transform both the training and testing sets.
- **Data Preparation for PyTorch:** Scaled feature data and target labels were converted into PyTorch tensors and loaded into a DataLoader with a batch size of **32**.
- **Neural Network Architecture:** A three-layer fully connected network [**(input_features=97) => 48 => 16 => 1**] was constructed, utilizing ReLU activation functions between hidden layers.
- **Loss and Optimization:** Training was performed using the Adam optimizer (learning rate 0.005) and nn.BCEWithLogitsLoss as the criterion, incorporating a positive weight factor of 2.0 to address class imbalance.
- **Training and Saving:** The model was trained for **25 epochs**, and the model state corresponding to the lowest training loss was saved (best model selection).
- **Evaluation and Metrics:** The final model was set to evaluation mode (model.eval()) to generate raw logits on the test set. Probabilities were obtained by applying the Sigmoid function to the logits. Final predictions were made using a classification threshold of 0.5.

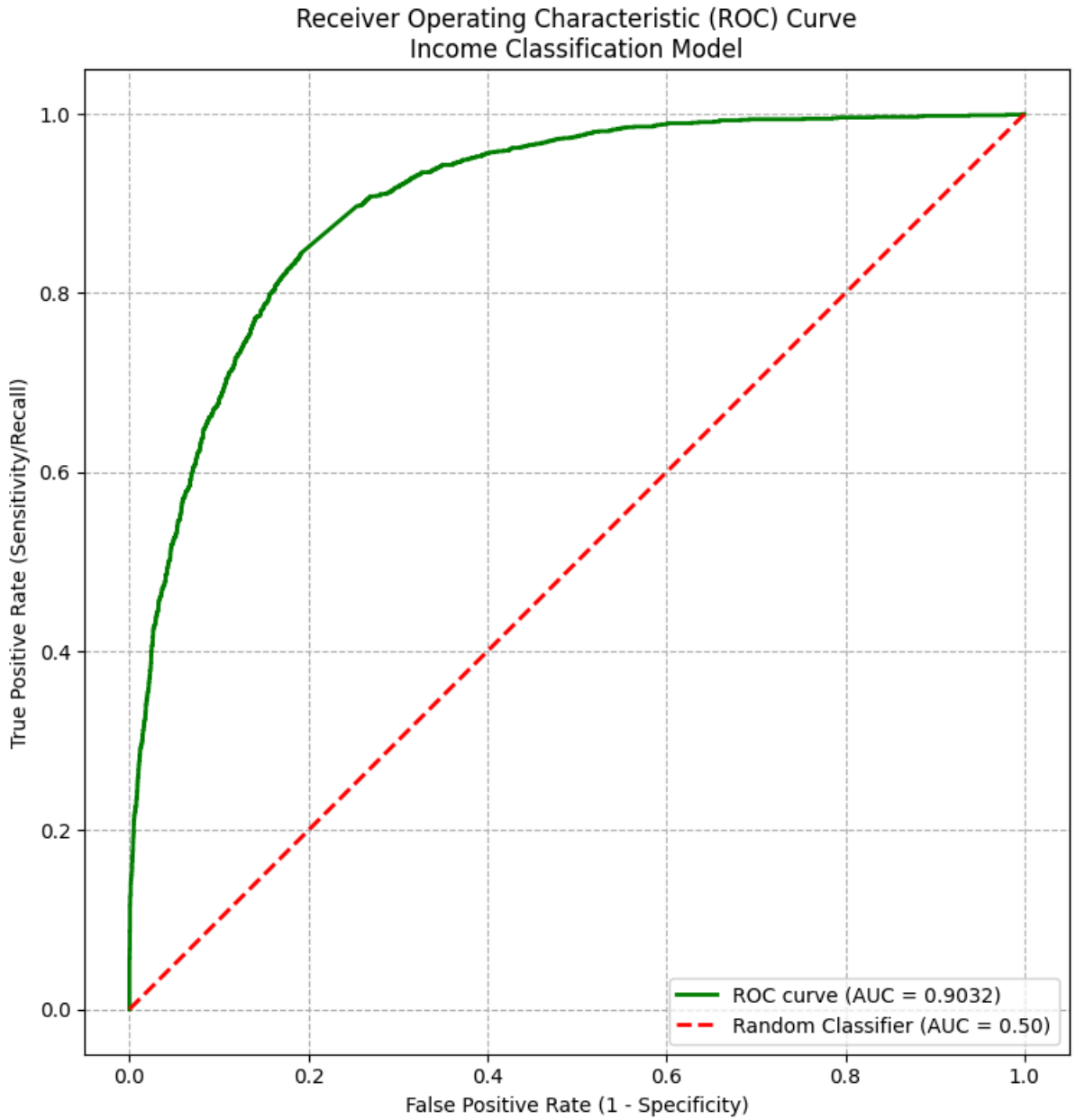
Results and Discussion

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.87	0.89	4940
1	0.64	0.75	0.69	1568
accuracy			0.84	6508
macro avg	0.78	0.81	0.79	6508
weighted avg	0.85	0.84	0.84	6508

Confusion Matrix:[[4277 663] [385 1183]]





The trained model successfully classifies the majority class (Income <50K) with High Precision (0.92) and Recall (0.87). It struggled with the minority class initially but the choice of BCEWithLogitsLoss with a pos_weight successfully achieved a favorable shift in the **trade-off** between **Precision** and **Recall** for the Minority class (Income >50K):

Good Recall (0.75): The model correctly identified 75% of all genuinely high-income individuals.

Mediocre Precision (0.64): The consequence of prioritizing Recall is the observed mediocre Precision. A Precision of 0.64 means that 36% of all instances predicted as high-income were actually low-income (False Positives). The model became more sensitive, leading it to classify more borderline cases as positive, which drives up the False Positive count and drives down Precision.

ROC AUC Validation: The high **AUC score of 0.9032** independently validates the model's overall quality and ability to separate the classes. The ROC curve lying far above the random baseline confirms that the model's assigned probabilities are highly accurate and reliable, regardless of the final threshold chosen.

Conclusion

The final model achieved an overall accuracy of 84%, and its performance metrics were successfully balanced using class weighting. The roc-auc value of 0.9032 confirms the model is a robust and effective classifier of high income, successfully mitigating the challenges posed by the data imbalance.

Part B — Clustering (Unsupervised Learning)

Objective

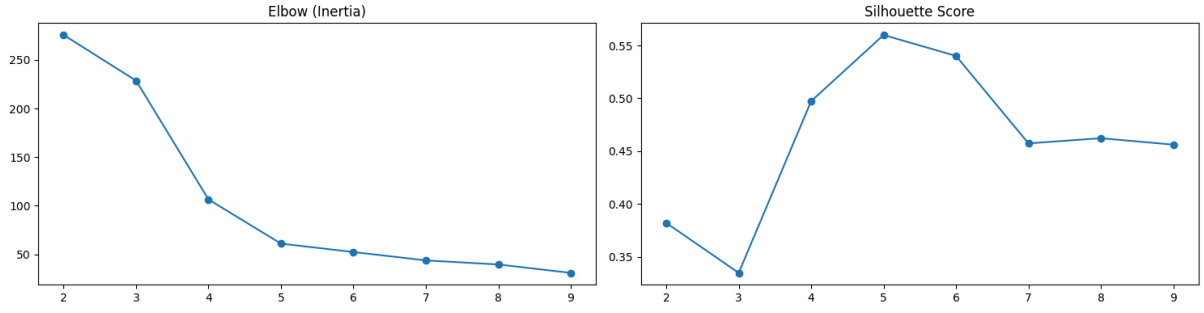
To apply K-Means clustering to identify distinct customer segments in a Mall based on the given features.

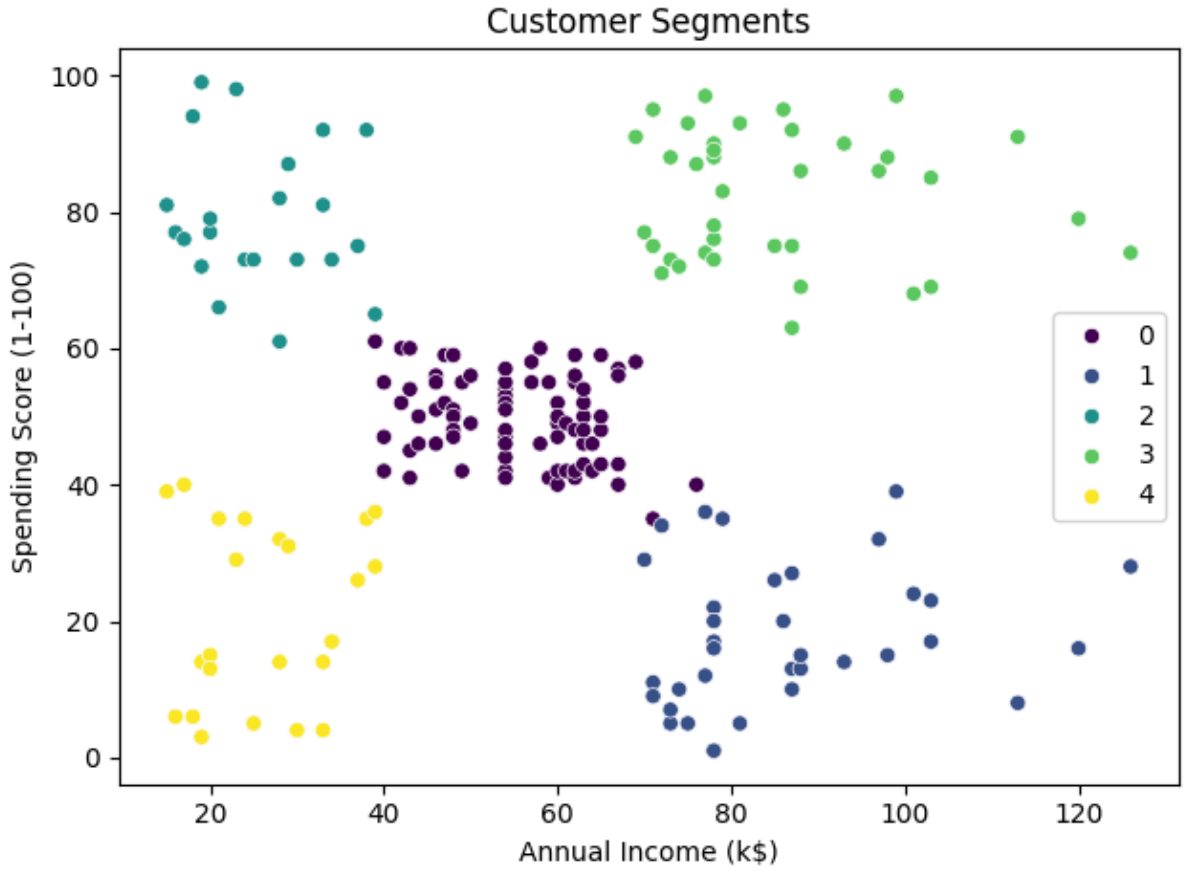
Methodology

- The dataset has been collected from Kaggle named Mall Customer Segmentation Data. The info.about the data is as follows: Shape of the dataset:(200,5)
  Columns in the dataset: ['CustomerID', 'Gender', 'Age', 'Annual Income (k$)', 'Spending Score (1-100)'].
- Performed checks for missing values (none found) and duplicate rows (none found). Descriptive statistics were generated to understand the data distribution.
- The non-predictive CustomerID column was dropped. The categorical Gender column was converted to a numerical format using One-Hot Encoding.
- Outliers in all numerical features (Age, Annual Income (k$), Spending Score (1-100)) were detected using the **Interquartile Range (IQR) method** (1.5 * IQR). Outlier rows were subsequently removed from the dataset.
- The final feature set was prepared for K-Means clustering by initializing the **StandardScaler** (a necessary step for distance-based algorithms like K-Means).
- Initial attempts using all features yielded poor clustering results. So the focus became two key behavioral features—**Annual Income** and **Spending Score**—as these are theoretically the strongest drivers for defining distinct market segments.
- A secondary analysis incorporated the **Age** feature to assess if the addition of a key demographic variable would enhance cluster separation beyond purely monetary factors.
- The optimal number of clusters (K) was determined by evaluating K values from 2 to 10 using two primary metrics: Elbow Method (Inertia) and Silhouette Score.
- K-Means clustering was executed twice:
    1. On the two-feature set (Annual Income, Spending Score) and visualized with a 2D scatter plot.
    2. On the three-feature set (Age, Annual Income, Spending Score) and visualized with a 3D scatter plot.

Results and Discussion

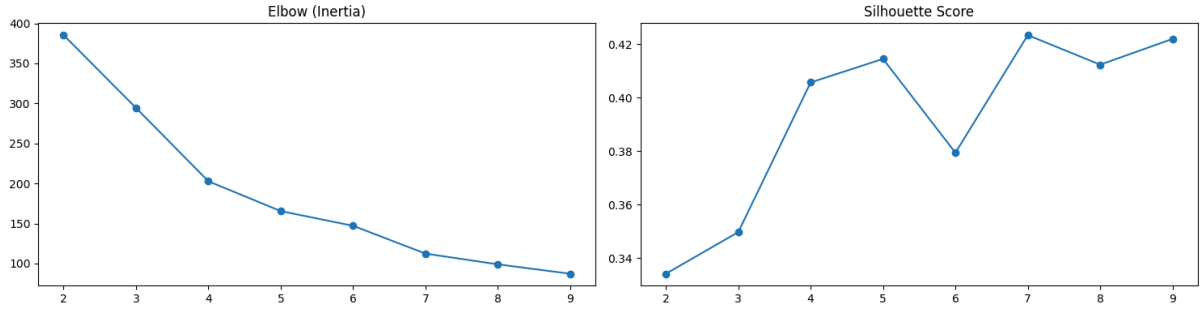
MODEL-1: K-Means on Annual Income and Spending Score Features:The **Silhouette Score** plot clearly indicates a peak at K=5 (Score approx 0.58). While the Elbow curve also suggests K=5.

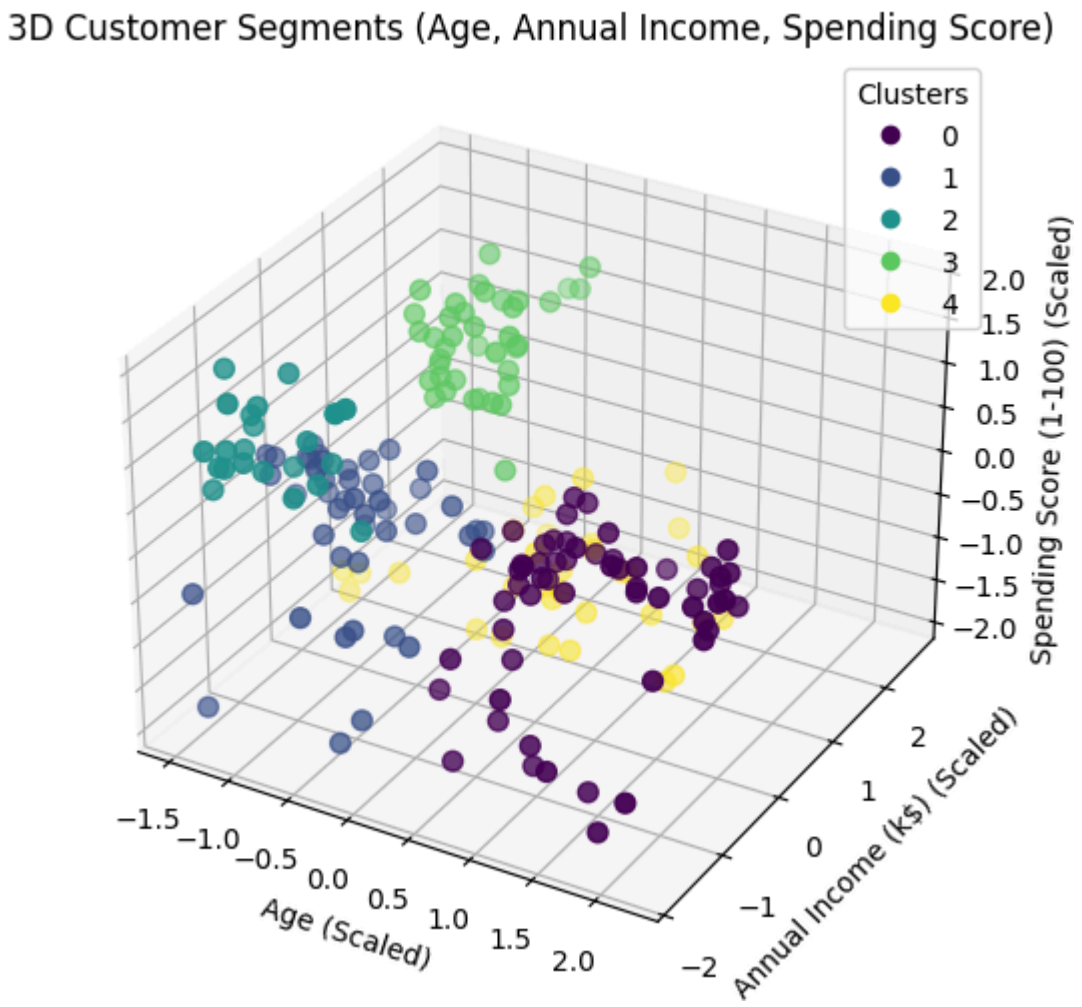




The 2D visualization for the K=5 clusters using Annual Income and Spending Score reveals five highly distinct and commercially actionable segments.

<u>MODEL-2</u> K-Means on Age, Annual Income, and Spending Score features: The Silhouette Score plot indicates a peak at K=7 (Score approx 0.42) but the score is lower than Model-1.





The 3D visualization demonstrated that incorporating the **Age** variable did not improve cluster quality as it is clearly observable that there is lack of clear separation among clusters and the Silhouette Score is also lower than that of Model-1's.

Conclusion

The initial feature selection is justified: Annual Income and Spending Score are the primary drivers of segmentation and provide the most robust model for direct business application.