Face Recognition & Loan Prediction Project Documentation

---

1) Source Code & Datasets

- **Source Code:** Uploaded to GitHub “ https://github.com/Salmaazoz22/fcai-ml-2025-project ”
- **Datasets:**
    - CK+ Extended Facial Expression Dataset "https://www.kaggle.com/datasets/davilsena/ckdataset "
    - Loan Approval Dataset "https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset "

---

2) Project Description

Dataset 1 — Facial Emotion Recognition (Classification)

A) General Information

- **Dataset Name:** CK+ Extended Facial Expression Dataset
- **Original Source:** Adapted from CK+, comparison with FER2013
- **Total Images:** 920 images
- **Image Size:** 48 × 48 pixels (grayscale)
- **Preprocessing:** Face-cropped using Haar cascade, noise adapted
- **Columns:** emotion (label), pixels (flattened), Usage (train/test)
- **Classes:** Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral, Contempt
- **Selected Classes for Project:** Disgust (1), Happiness (3), Surprise (5), Neutral (6)
- **Balanced Dataset Size:** 236 images
- **Data Split:** Training 80% (188), Testing 20% (48)

B) Implementation Details

Model 1 — Logistic Regression (Supervised)

- **Features:** Raw pixel values (2304 features)
- **Scaling:** StandardScaler
- **Cross-Validation:** 5-fold for hyperparameter tuning
- **Hyperparameters:** Multinomial LR, SAGA solver, ElasticNet regularization, C=1, L1 ratio=0.2, max_iter=2000

Model 2 — K-Means + HOG (Unsupervised)

- **Feature Extraction:** HOG (Orientations=9, Pixels per cell=8×8, Cells per block=2×2, L2-Hys normalization)
- **Extracted Features:** 900 per image

- **K-Means:** 8 clusters, k-means++ init, max_iter=1000, n_init=50
- **Label Mapping:** Majority voting from cluster to true label

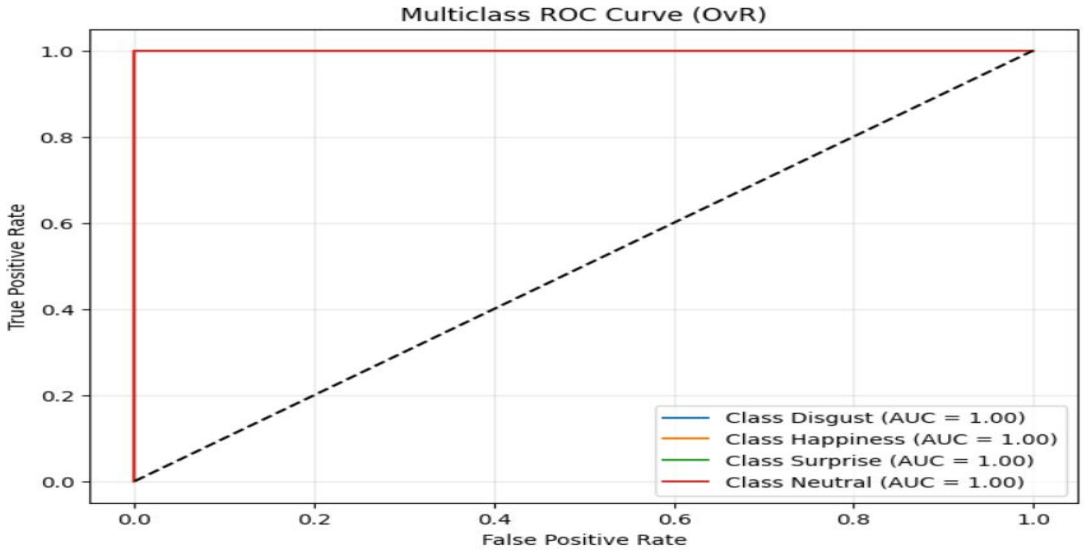
C) Results (Testing Data)

Model	Accuracy	Loss / Inertia	Confusion Matrix	ROC / AUC
Logistic Regression	97.92%	Log-loss decreased steadily	High diagonal, minimal misclassification	High AUC for all classes
K-Means + HOG	71.19%	-----	Moderate alignment with true labels	Approximate ROC from cluster distances

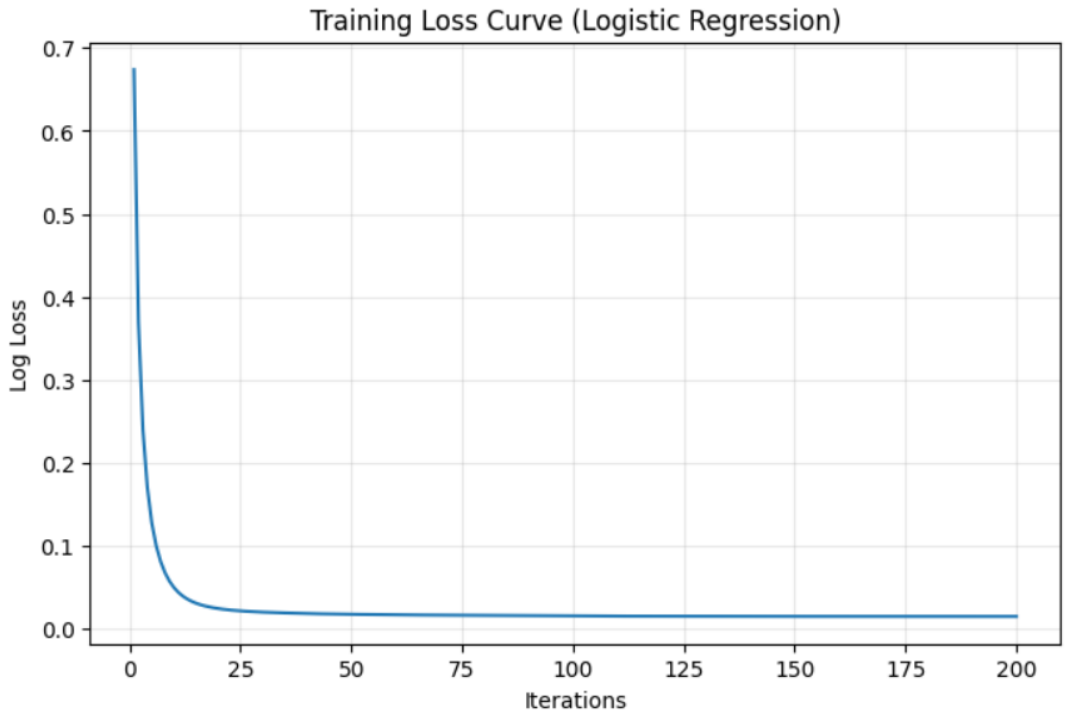
- K-Means is unsupervised and does not use true labels during training. Therefore, ROC curves, loss curves, and confusion matrices are not directly applicable. Instead, we evaluate clustering quality using accuracy after mapping clusters to labels and inertia trends.

D) Visualizations

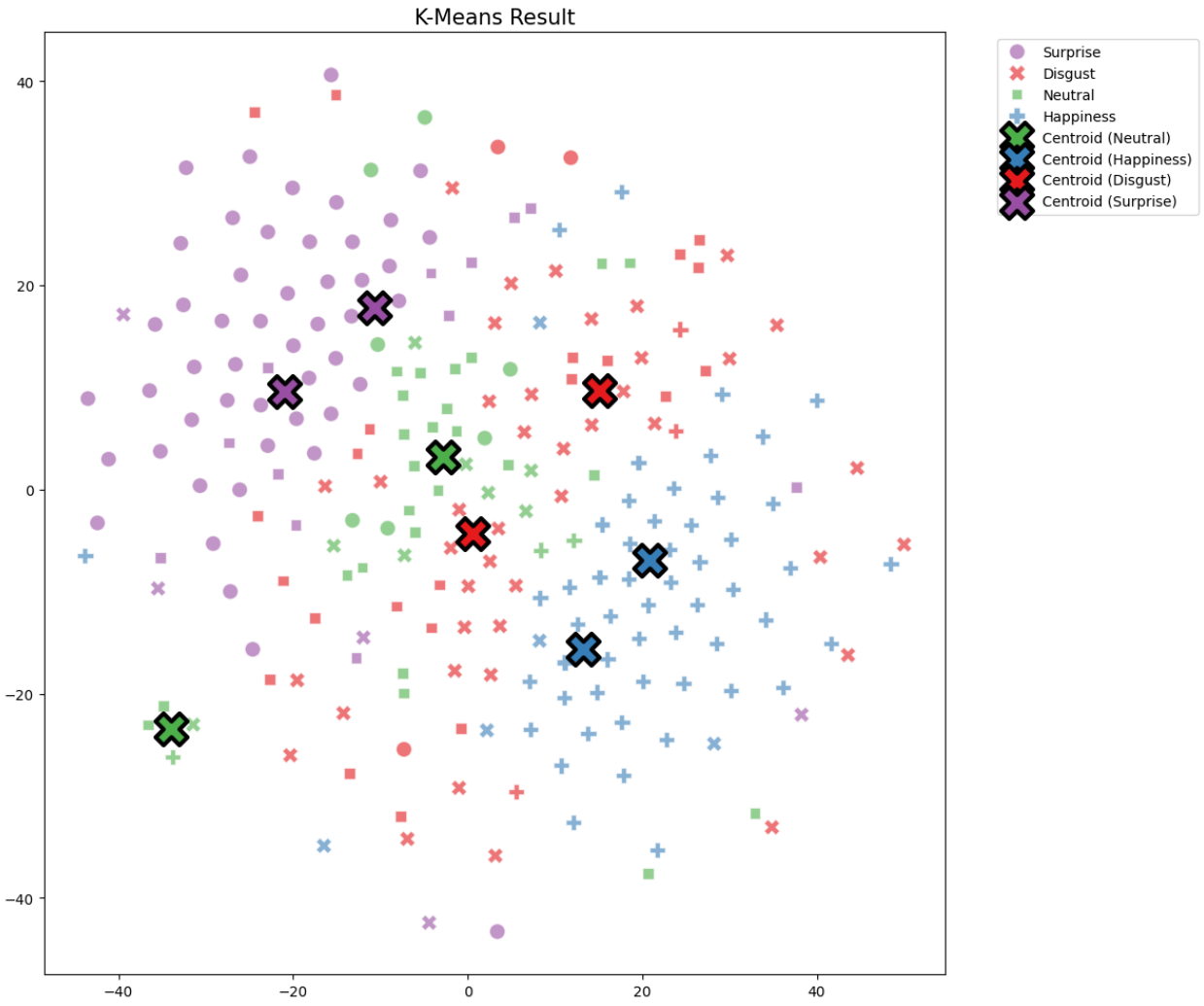
- Multi class ROC Curve



- Loss Curve for Linear Regression



- Scatter plot with centroids for Kmeans



Comparative Summary

- Supervised learning (Logistic Regression) significantly outperforms unsupervised K-Means.
- HOG features allow reasonable clustering despite lower accuracy.
- Visualizations (t-SNE, confusion matrix, ROC curves) support qualitative analysis.

Dataset 2 — Loan Amount Prediction (Regression)

A) General Information

- **Dataset Name:** Loan Approval Dataset
- **Original Records:** 4,269
- **Filtered Records (Approved Loans):** 2,656
- **Features:** 11 numeric and categorical features (['no_of_dependents', 'education', 'self_employed', 'income_annum', 'loan_amount', 'loan_term', 'cibil_score', 'residential_assets_value', 'commercial_assets_value', 'luxury_assets_value', 'bank_asset_value', 'loan_status'])
- **Target:** loan_amount
- **Data Split:** Training 1,699, Validation 425, Testing 532

B) Implementation Details

Preprocessing

- Dropped: loan_id, loan_status (after filtering)
- Categorical Encoding: education, self_employed (LabelEncoder)
- Feature Engineering: total_assets = residential + commercial + luxury + bank
- Outlier Handling: IQR clipping and Z-score validation
- Scaling: StandardScaler

Models

1. **Linear Regression**
    - Hyperparameters: default
    - Cross-Validation: 5-fold
2. **KNN Regressor**
    - Hyperparameters: k=5
    - Cross-Validation: 5-fold

C) Results

Linear Regression (Test Set)

- R^2: 0.8725
- MAE: 16.64%
- Overfitting: Low
- Stability (CV Std R^2): High

KNN Regressor (Test Set)

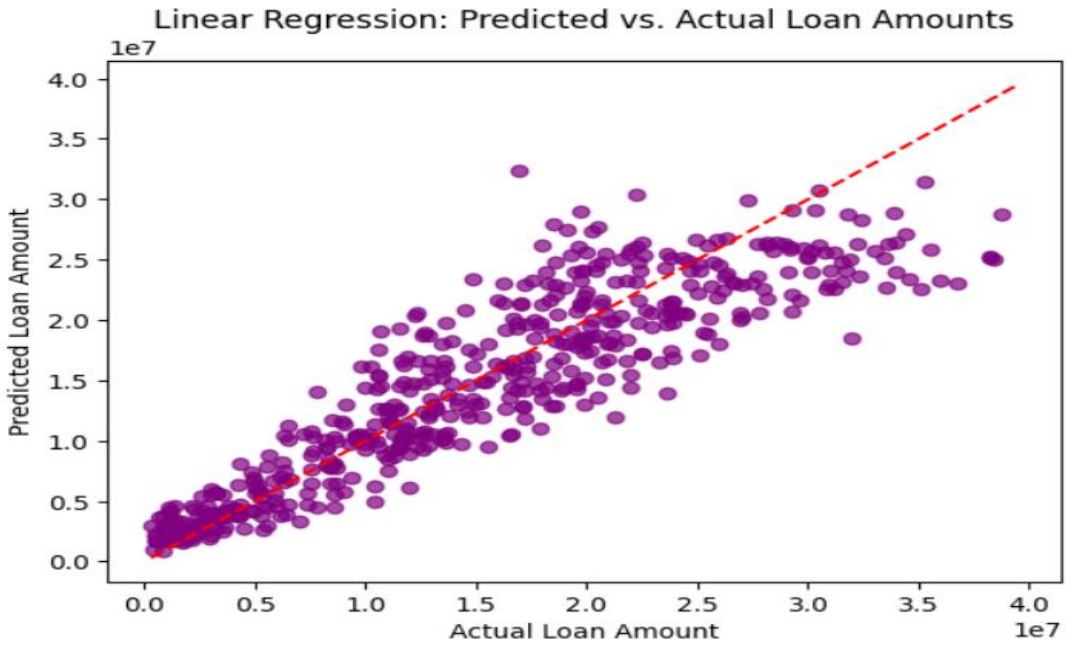
- R^2: 0.8106
- MAE: 20.13%
- Overfitting: Minimal
- Stability: Moderate

Model Comparison

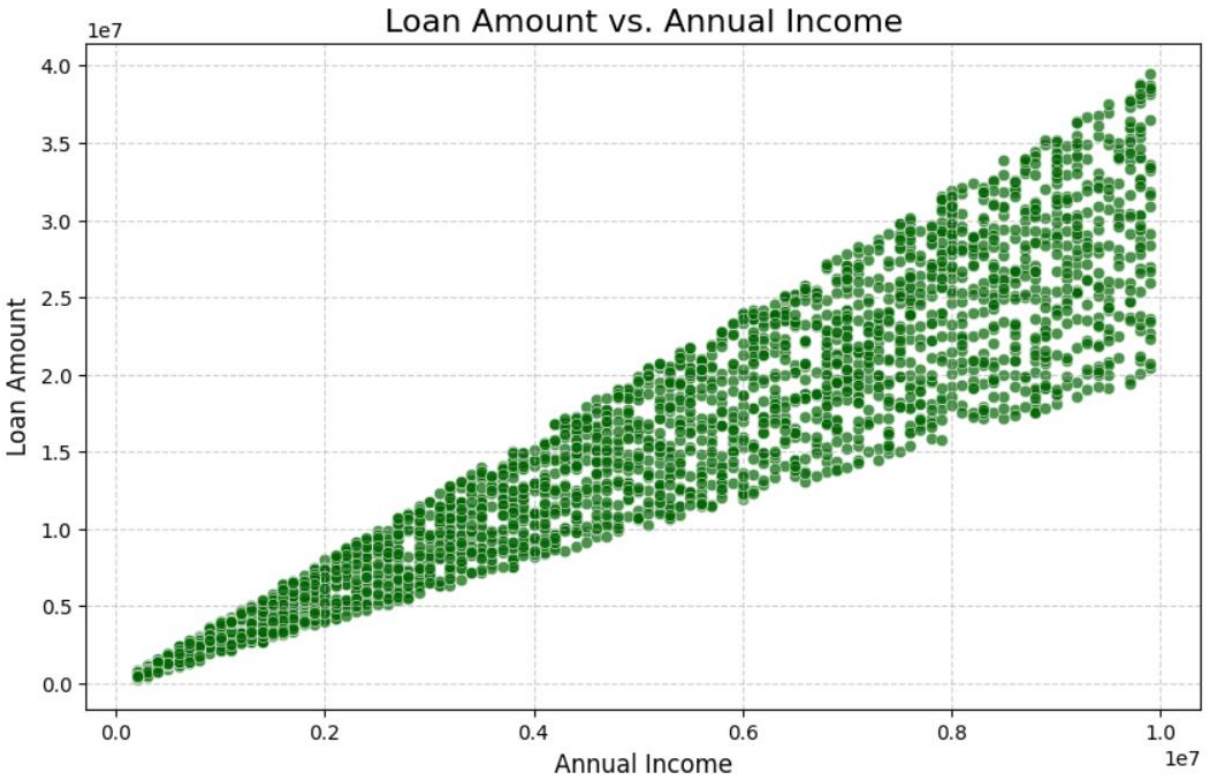
- Linear Regression outperforms KNN
- Lower MAE and higher R^2 indicate better generalization

Visualization

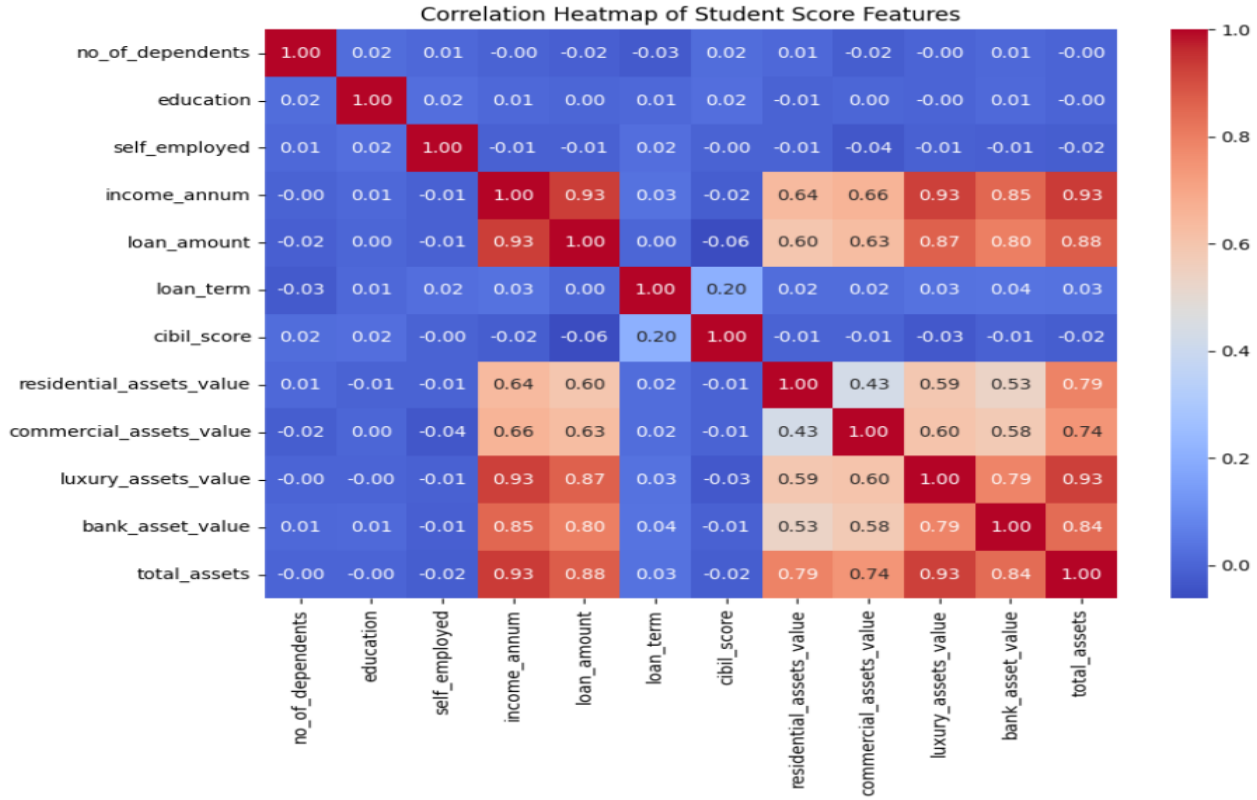
- Predicted vs Actual scatter plots



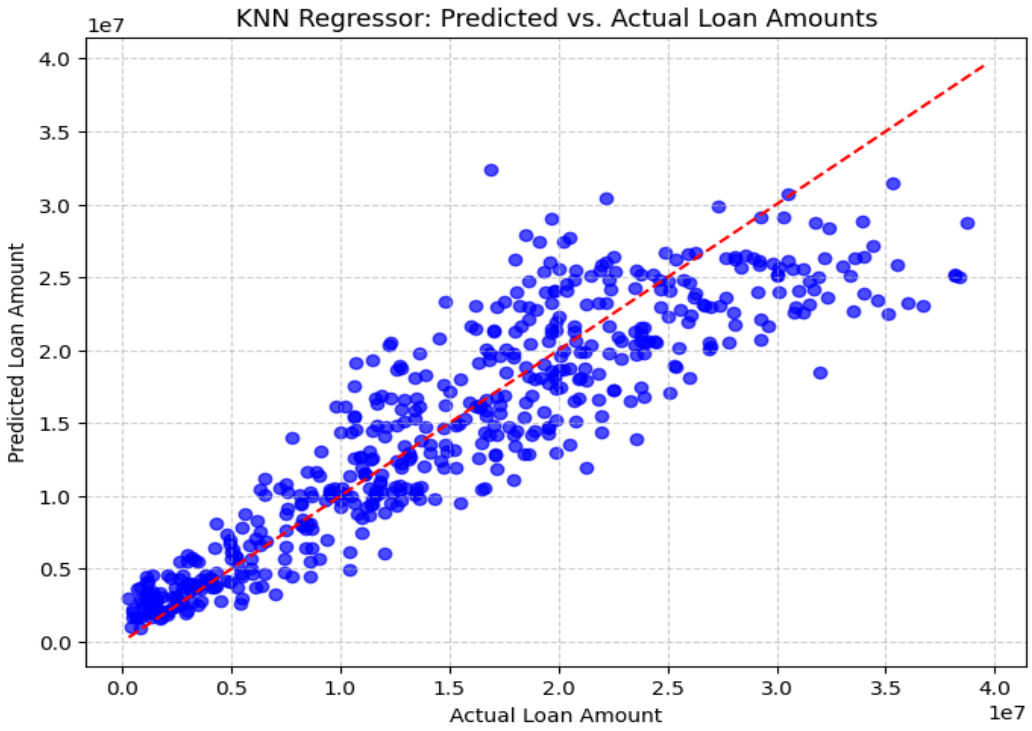
- Scatter plot between 2 of the most corelated features



- Heatmap to show the correlation between the features



- KNN Regressor: Predicted vs. Actual Loan Amount



4) Final Comparison & Discussion

Dataset	Task	Models	Best Performance
CK+ Images	Classification	Logistic Regression	Accuracy 97.92%
CK+ Images	Classification	K-Means + HOG	Accuracy 71.19%
Loan Dataset	Regression	Linear Regression	R^2 0.8725, MAE 16.64%
Loan Dataset	Regression	KNN Regressor	R^2 0.8106, MAE 20.13%

Key Takeaways

- Classification requires careful feature selection and benefits from supervised models.
- Regression benefits from feature engineering (total_assets) and scaling.
- Cross-validation ensures model stability and prevents overfitting.
- Visualizations support both qualitative and quantitative evaluation.
