Machine Learning Assignment (Loan Repayment Dataset) by Kamya Sarda

- This study applies logistic regression to predict agricultural loan repayment status using demographic and financial data from farmers in Karnataka, India. The dataset contained 1,218 records and 12 variables, including indicators such as income per acre, training on sericulture, crop insurance, chawki practices, district code, and pest impact. The target variable was loan repayment (1 = repaid, 0 = default). The objective was to conduct exploratory data analysis (EDA), prepare features, train the model, evaluate its performance, and derive insights for agricultural finance decision-making.
- EDA revealed that income per acre had a wide spread, ranging from –₹79,350 to ₹155,000, with a mean of around ₹36,000 and a median of ₹25,000, indicating high variability and the presence of extreme outliers. Approximately 17% of farmers reported negative incomes, suggesting agricultural losses. Binary features such as crop insurance, sericulture training, and pest impact were highly skewed, with many farmers not participating in these programs. Missing values were negligible (<0.25%) and were treated using median imputation for numeric variables and zero imputation for binary indicators, along with missing-value flags. Outlier detection using the IQR method confirmed extreme values in income per acre, which were winsorized to reduce distortion. Correlation analysis revealed near-perfect correlation between chawki_bivol and chawki_combo_other (≈0.99), so both were not included together to avoid multicollinearity.
- Feature extraction involves dropping non-predictive identifiers such as person_id, gender, and district, while retaining relevant predictors. DecisionTree-based Gini feature selection identified income_per_acre, female, training_on_sericulture, crop_insured, chawki_bivol, chawki_combo_other, and affected_by_pest as the most important features. The dataset was split 70:30 into training and test sets, and predictors were standardized prior to model training. Logistic regression was then applied, producing interpretable coefficients and probability-based outputs.
- The model achieved an accuracy of 67.2% and an AUC of 0.656, reflecting moderate predictive ability. The confusion matrix indicated that the model correctly predicted 245 repaid loans and only 1 default, while misclassifying 117 defaults as repaid and 3 repayments as defaults. Precision and recall metrics highlighted this imbalance: for class 1

(repaid), precision was 0.68 and recall 0.99, while for class 0 (default), precision dropped to 0.25 and recall to 0.01. This shows that the model is highly sensitive to identifying repayments but fails to detect non-repayments. Among the coefficients, income per acre had a positive effect ($\beta = +0.486$), confirming that higher farm income increases repayment probability. Other features such as training, insurance, and chawki practices showed weaker but generally positive influence.

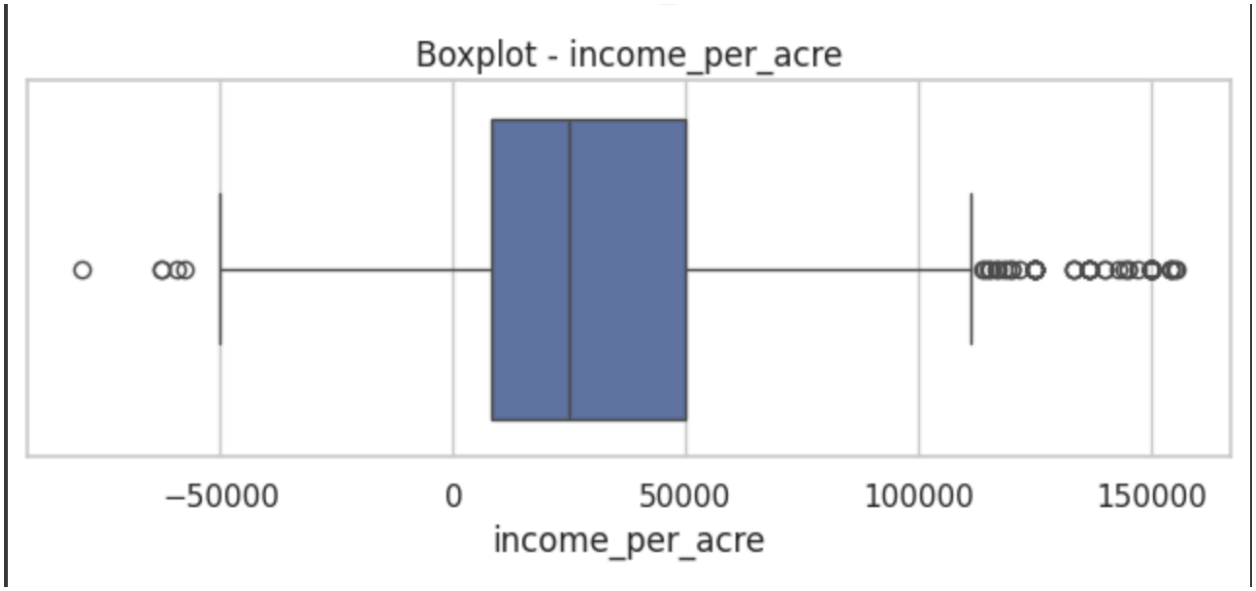
- The insights suggest that income stability is the strongest driver of repayment behavior, while low adoption of crop insurance and pest infestations remain significant risks. The model's bias toward predicting repayment highlights the challenge of class imbalance, where most observations correspond to successful repayments. From a business perspective, this implies that while the model can identify creditworthy farmers, it underestimates default risk, which could expose lenders to financial losses if used in isolation.
- Based on these findings, several recommendations emerge. First, class imbalance should be addressed in future modeling using resampling techniques such as SMOTE or by adjusting the probability threshold below 0.5 to improve default detection. Second, lenders could require or incentivize crop insurance for higher loan amounts, reducing default risk. Third, expanding sericulture training programs and pest management support could improve repayment behavior across districts. District-level profiling is also advised, as repayment rates vary regionally. Finally, future data collection should include more borrower-specific attributes (e.g., landholding size, household income, credit history) to strengthen risk assessment.
- In conclusion, the logistic regression model provides a useful baseline for predicting loan repayment, achieving 67.2% accuracy with strong recall for repaid loans but weak performance in identifying defaults. While the model highlights income per acre as the key repayment driver and offers actionable business insights, it is not yet sufficient as a standalone risk management tool. With improvements in class balancing, additional features, and possibly ensemble methods, the model can be enhanced to support more reliable and data-driven lending decisions in the agricultural sector.

Data Source: Jayalaxmi Agrotech Agricultural Finance Dataset

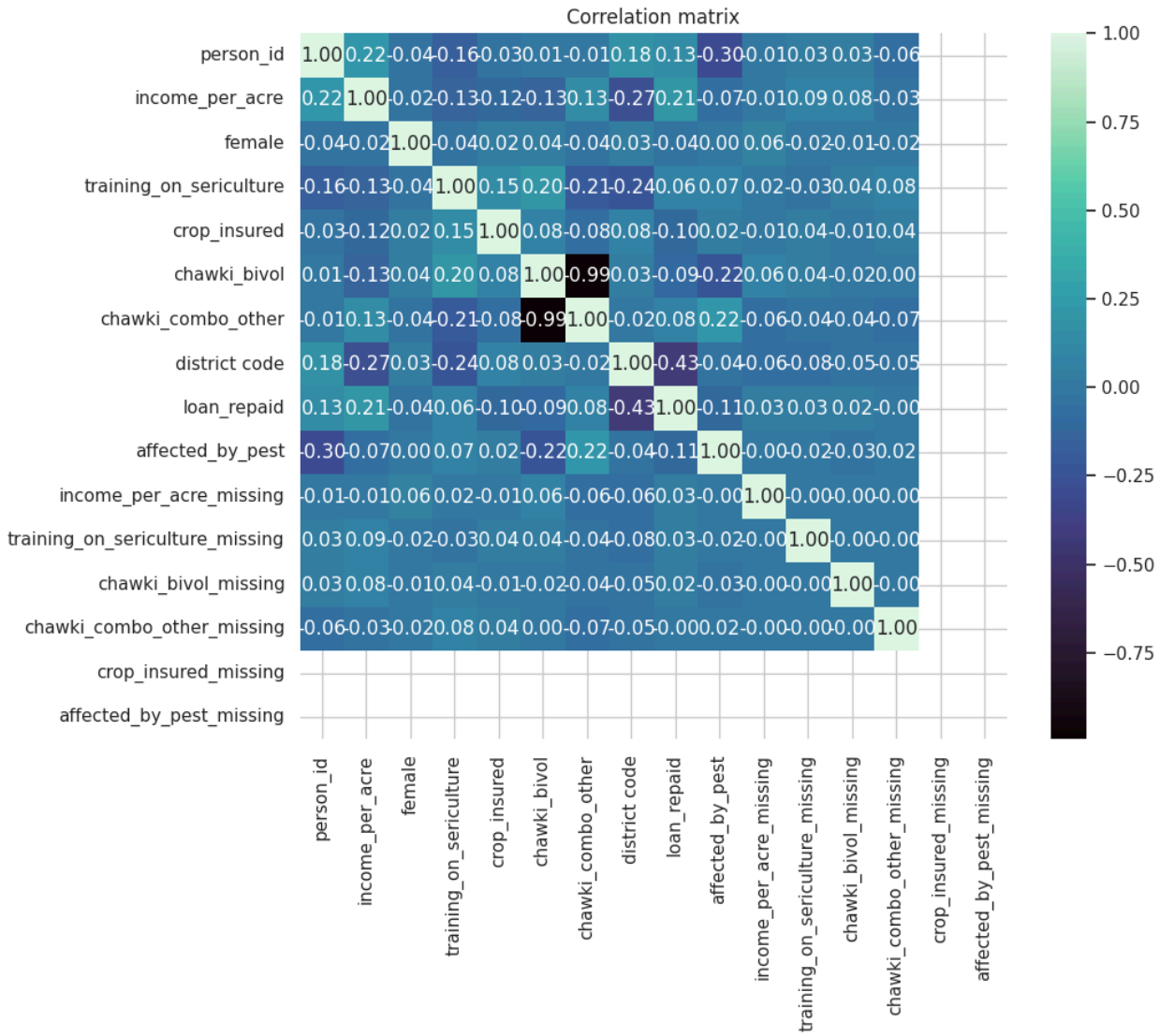
Appendix: Supporting Figures (EDA and Model Results)

1. **Histogram – Income per Acre Missing** (Explains missing value handling/indicators)

2. **Boxplot – Income per Acre** (Shows distribution and outliers in farmer income)



3. **Correlation Matrix/Heatmap** (Shows multicollinearity check and feature relationships)



4. **Confusion Matrix + Classification Report Output** (screenshot/text table) (Helps explain precision, recall, f1, and accuracy)

```
Logistic Regression Coefficients:
income_per_acre: 0.4856
Intercept: 0.7805

Accuracy: 0.6721
Confusion Matrix:
 [[  1 117]
 [  3 245]]

Classification Report:
              precision    recall  f1-score   support

           0       0.25      0.01      0.02       118
           1       0.68      0.99      0.80       248

    accuracy                           0.67       366
   macro avg       0.46      0.50      0.41       366
weighted avg       0.54      0.67      0.55       366

AUC: 0.6562
```

5. **ROC Curve (AUC = 0.656)** (Model performance evaluation)