Summary (Income Qualification)

Objective: This project aimed to improve income qualification predictions for families in Latin America using data from Costa Rica. Many social programs struggle to allocate aid to the most vulnerable due to the lack of accurate income records. A traditional method, **Proxy Means Test (PMT),** relies on observable household attributes but has accuracy limitations. The goal was to explore new approaches using machine learning to improve these predictions.

Problem Statement:

The project involved predicting income levels based on household characteristics. The dataset comprised various features related to housing conditions, family members, education levels, and asset ownership. The primary challenge was to classify families into four income categories: extreme poverty, moderate poverty, vulnerable, and non-vulnerable households.

Solution Approach:

1. **Data Preprocessing and Cleaning:** The dataset was cleaned to address missing values in critical columns like rent payment, number of tablets, and education levels. Unnecessary columns were dropped, and essential categorical variables were converted to numerical values. Households without heads were identified, and the income level of each family member was set based on area-specific rent thresholds.

2. **Model Building with Random Forest:**
    - A Random Forest Classifier was chosen to model the dataset. The data was split into training and testing sets, and the classifier was trained on the features.

    - The initial model achieved a 93.5% accuracy on the testing set, indicating reliable performance in classifying income levels.

3. **Cross-Validation and Feature Importance:**

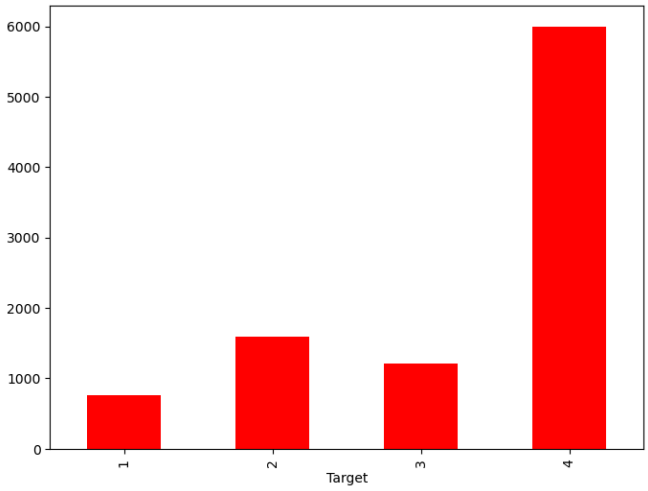
- Cross-validation was conducted, achieving an average accuracy of 92.9%. Important features influencing income classification included variables such as the number of rooms, ceiling material, family size, and education level.

4. **Model Optimization with Key Features:**
    - The model was optimized by focusing on key features with high importance, achieving an accuracy of 91.8%. This revealed that using all relevant features yields higher accuracy.

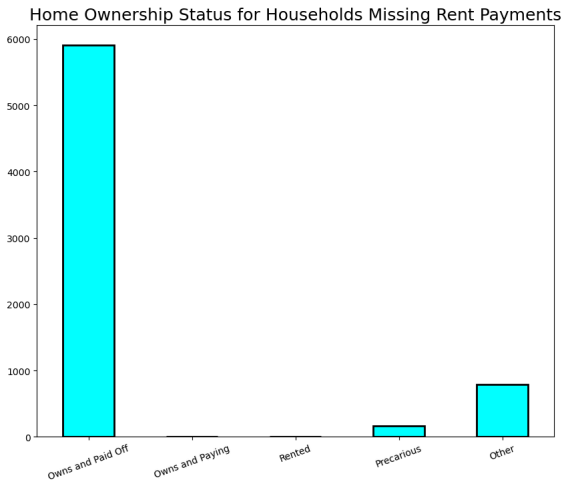
Conclusion

The project demonstrated that machine learning models could enhance traditional methods like PMT, achieving reliable accuracy in income qualification predictions. However, comprehensive data usage remains critical to maintaining prediction accuracy. Machine learning techniques like Random Forests provide a promising alternative to traditional methods for income qualification. While focusing on essential features can be effective, leveraging a comprehensive dataset ensures better accuracy in income-level predictions.

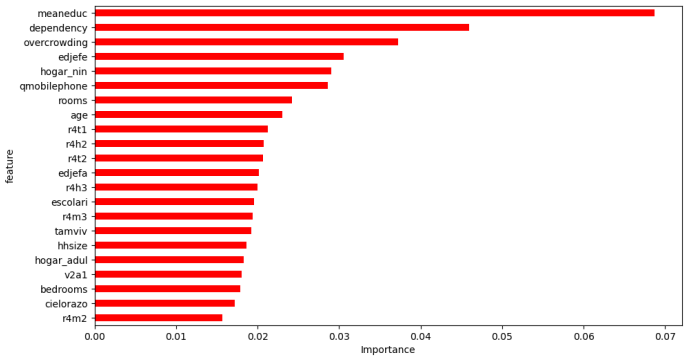
Visualization of Data (Output)



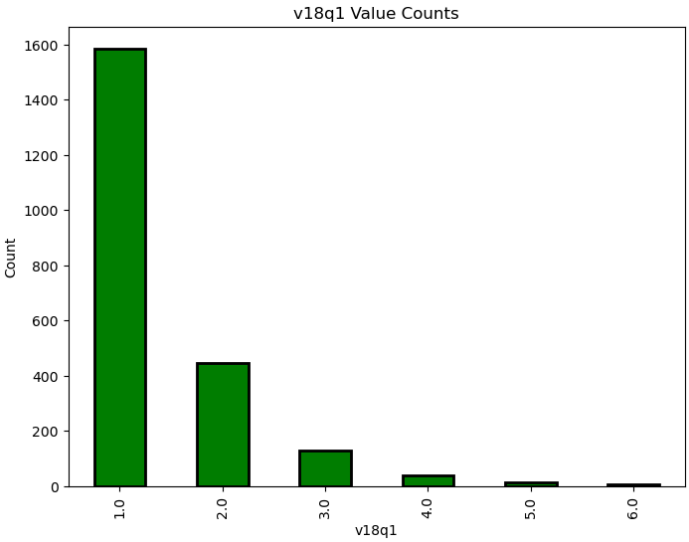
Biases Target



Home Ownership Status



Random Forest with Cross Validation



Value Counts