📊 Detailed Financial ML Case Study Report Summary

# 1. Dataset Highlights & Red Flags

• 34,638 people are spending more than their income — a strong sign of financial stress and overspending.

• 67,440 individuals have extremely low disposable income (< ₹1,000/month), indicating vulnerability.

• There is a significant gap between desired savings (~₹7,514) and average disposable income (~₹5,980).

# 2. Savings Behavior Insights

• Most people want to save more than they realistically can.

• This gap suggests unrealistic savings goals and poor financial planning — opportunity for smarter savings prediction models.

# 3. Statistical Skewness in Financial Behavior

• Columns with high positive skewness (>1): Healthcare, Entertainment, Miscellaneous, Education, Eating Out savings.

• Suggests that majority save very little in these, while a few save a lot — right-skewed distribution.

# 4. Strong Feature Correlations

• Healthcare ↔ Potential\_Savings\_Healthcare: 0.81

• Entertainment ↔ Potential\_Savings\_Entertainment: 0.78

• Desired\_Savings ↔ Desired\_Savings\_Percentage: 0.79

• Education savings and expenses are also closely linked — indicates redundancy.

# 5. Zero/Anomaly Indicators

• Thousands of users have zero values in key columns like Dependents, Healthcare, Rent, Entertainment — possible data entry issues or actual conditions (e.g., students).

# 6. Inference Testing Results

• ANOVA: Desired\_Savings\_Percentage varies significantly by Occupation (p < 0.05). Occupation is a good predictor.

• Confidence Interval confirms stability of savings percentage estimation.

• Correlation Test: Income and Education Savings show strong positive correlation with Desired Savings %. Expenses negatively impact it.

# 7. Best ML Models for This Dataset

✓ Linear Regression — baseline, easy to interpret

✓ Ridge/Lasso — handle feature overlap and multicollinearity

✓ Random Forest — great for tabular data, captures nonlinear patterns

✓ Gradient Boosting / XGBoost — high accuracy, real-world ready

✓ Optional: SVR or MLP if dataset scaled and cleaned for advanced learning

# 8. Summary Recommendation

• Start with Linear Regression ✅

• Try Random Forest, then Gradient Boosting

• Choose model with best R² and lowest RMSE on test set