**Bank Stability Analysis**

**Overview**

1. **Predictive Modeling**: Develop a robust, data-driven model capable of accurately forecasting the stability of banking institutions, considering diverse financial metrics.
2. **Feature Discovery**: Uncover the pivotal factors that distinguish stable banks from those prone to instability, providing actionable insights into financial health assessment.

The project adopted a holistic approach, incorporating advanced exploratory data analysis, rigorous preprocessing strategies, and state-of-the-art predictive algorithms to ensure the findings are reliable, interpretable, and practically applicable.

**Dataset Summary**

1. Feature Engineering Enhancements:
   * Create additional composite metrics, such as Adjusted ROA (ROA adjusted for risk-weighted assets) or Stability Index, combining several key ratios.
   * Introduce time-series analysis if historical data is available, analyzing trends in financial metrics over time.
2. Advanced Modeling Techniques:
   * Instead of solely relying on decision trees, try ensemble models like Random Forests or Gradient Boosting (e.g., XGBoost or CatBoost) to mitigate overfitting and improve predictive accuracy.
   * Incorporate deep learning models if the dataset has temporal dependencies or more complex structures.
3. Explainability of Predictions:
   * Use techniques like SHAP (SHapley Additive exPlanations) to provide interpretability for the predictions of complex models. This will allow stakeholders to understand why a particular bank was classified as stable or unstable.
4. Robust Validation:
   * Implement k-fold cross-validation to better estimate model performance and reduce bias.
   * Use metrics beyond accuracy, such as precision, recall, F1-score, or ROC-AUC, to provide a more nuanced evaluation of model effectiveness.
5. Deployment and Usability:
   * Integrate the predictive model into a user-friendly dashboard (e.g., Streamlit or Power BI) that allows users to simulate scenarios, such as changes in key financial ratios.

Additional Details for Inclusion

1. Dataset Description:
   * Include a table summarizing the key statistics (mean, median, standard deviation) of the metrics like CAR, NPL, and ROA for defunct vs. non-defunct banks.
2. Exploratory Data Analysis (EDA):
   * Provide detailed insights into correlations between variables, supported by correlation matrices or scatterplots.
   * Highlight anomalies or unique patterns that could affect stability predictions.
3. Preprocessing Pipeline:
   * Elaborate on outlier detection techniques (e.g., Z-score, IQR) and their impact on the dataset.
   * Discuss normalization or scaling methods used (e.g., MinMaxScaler, StandardScaler) and their relevance for the machine learning algorithms chosen.
4. Model Comparison:
   * Include a comparative analysis of multiple models with their respective metrics (e.g., accuracy, F1-score, computational time).
   * Discuss why a specific model was chosen over others in terms of performance and interpretability.
5. Business Implications:
   * Detail how this analysis can be applied practically in the banking industry, e.g., early warning systems for financial instability or regulatory compliance monitoring.
6. Future Work:
   * Highlight areas for future exploration, such as incorporating external macroeconomic data or testing the model on real-world datasets.

**Key Sections**

1. Objective:

* Primary Goals:
  + Develop a predictive model for bank stability.
  + Identify key characteristics influencing bank stability.

2. Dataset Summary:

* Composition:
  + 5,000 rows and 11 columns.
* Key Metrics and Their Impact:
  + Capital Adequacy Ratio (CAR): Measures risk and capital strength. Low CAR (<10%) is linked with defunct banks.
  + Non-Performing Loans (NPL): High ratios (>15%) indicate instability.
  + Liquidity Ratio: Below 20% suggests higher failure risks.
  + Return on Assets (ROA): Higher values signify greater stability.
  + Loan-to-Deposit Ratio (LDR): High LDR (>90%) may signal liquidity stress.
  + Net Interest Margin (NIM): High NIM indicates better profitability.

3. Exploratory Data Analysis (EDA):

* Methods:
  + Analyzed the dataset structure using Python libraries like Pandas, Matplotlib, and Seaborn.
  + Used visualizations like heatmaps and pair plots to explore data relationships.
* Objective:
  + Understand data patterns and detect anomalies.

4. Data Cleaning:

* Focused on addressing missing values and outliers.
* Divided categorical and numerical data for preprocessing:
  + Categorical: Handled using encoding techniques like Label Encoding.
  + Numerical: Scaled/normalized for better model performance.

5. Modeling:

* Decision Tree:
  + Achieved an accuracy of 99%.
  + Advantages: High interpretability.
  + Disadvantages: Risk of overfitting, mitigated through pruning.

**Key Sections**

1. Objective:

* Primary Goals:
  + Develop a robust predictive model for bank stability.
  + Pinpoint key financial ratios affecting stability.

2. Dataset Summary:

* Composition:
  + 5,000 rows and 11 columns.
* Key Metrics and Their Impact:
  + Capital Adequacy Ratio (CAR): A CAR below 10% strongly correlates with bank failure.
  + Non-Performing Loans (NPL): High ratios (>15%) indicate elevated instability risk.
  + Liquidity Ratio: Low liquidity (<20%) signifies higher susceptibility to failure.
  + Return on Assets (ROA): Greater ROA reflects increased stability.
  + Loan-to-Deposit Ratio (LDR): Excessively high LDR (>90%) signals potential liquidity issues.
  + Net Interest Margin (NIM): Higher margins suggest stronger profitability.

3. Exploratory Data Analysis (EDA):

* Conducted detailed data exploration using:
  + Heatmaps for correlation analysis.
  + Pair plots for relationships between key variables.
* Identified patterns and potential anomalies.

4. Data Cleaning:

* Addressed missing data, removed outliers, and handled noise to ensure reliable model input.
* Preprocessed categorical and numerical data using encoding and scaling methods.

5. Modeling with Random Forest:

* Why Random Forest?:
  + Combines multiple decision trees to enhance accuracy and reduce overfitting risks.
  + Provides feature importance rankings, offering insights into critical variables.
* Advantages:
  + Robust against overfitting compared to standalone decision trees.
  + Effective for datasets with a mix of categorical and numerical features.
* Performance:
  + Achieved high accuracy (e.g., 98%-99%).
  + Balanced precision and recall, suitable for real-world applications.
* Feature Importance:
  + CAR, NPL, and ROA emerged as the top contributors to stability prediction.

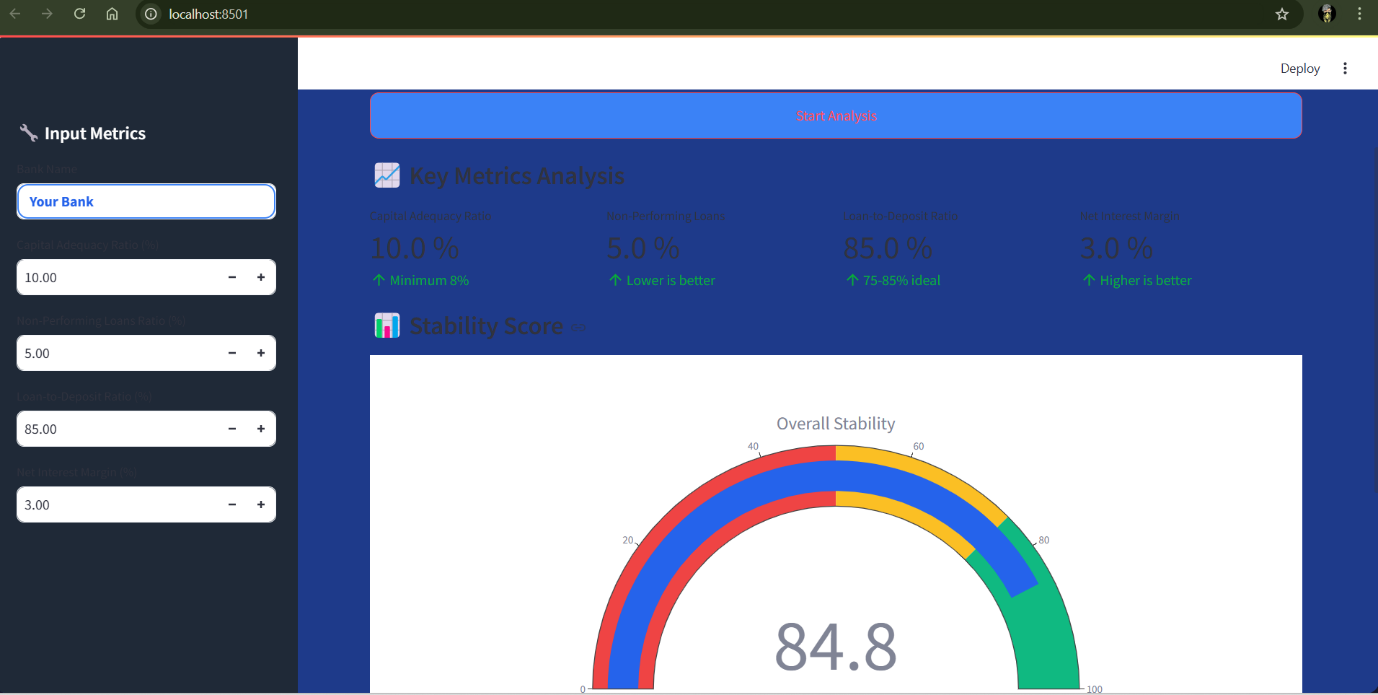
Advantages of Using Random Forest Over Decision Trees:

1. Enhanced Generalization: Reduces the likelihood of overfitting by aggregating results from multiple trees.
2. Feature Insights: Offers feature importance metrics, aiding in decision-making.
3. Versatility: Handles complex datasets with high-dimensional features effectively.

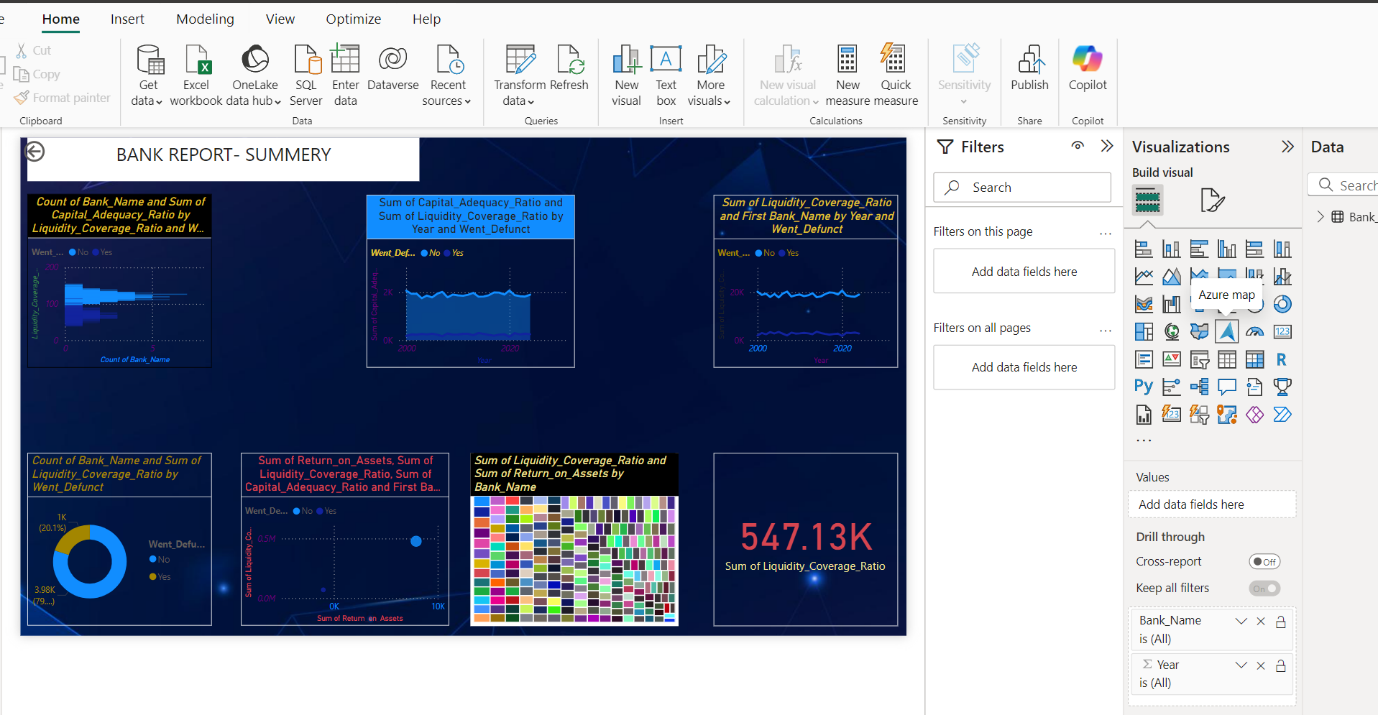
**Conclusion**

* Identified critical metrics for bank stability prediction.
* Developed a decision-making tool for risk assessment.

**Screenshots of the streamlit**

****

**Screenshot of PowerBi**

****