DATA MINING FINAL PROJECT

BANKRUPTCY PREDICTION

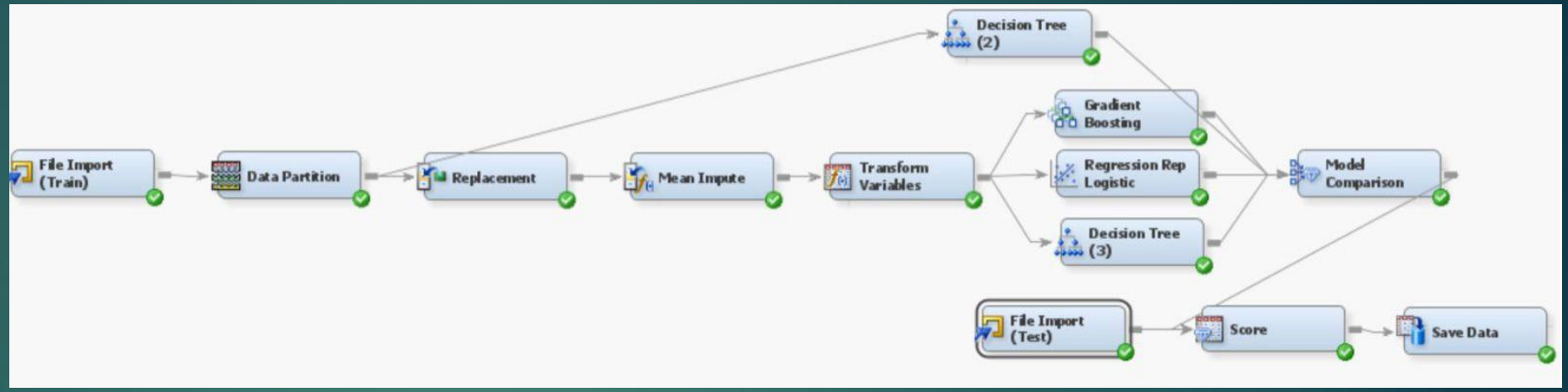
Team 1:
- Ajay Shankar
- Krithiga Rajan Sangeetha Rajan
- Keerthi Anand Sangeetha Rajan

INTRODUCTION

- Develop a machine learning model to predict firm bankruptcy using financial indicators.
- Training set with 64 predictors and a target variable
- Class imbalance with only 211 bankrupt cases out of 10,000.
- Addressed class imbalance using oversampling (replicating minority class records).
- Developed an ensemble model achieving a score of 96.088

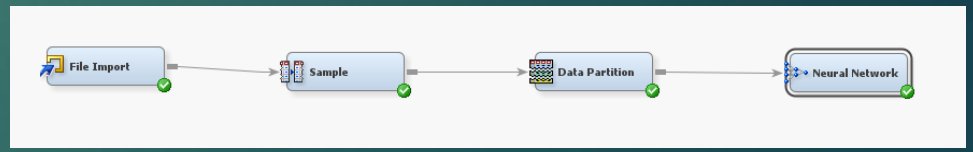
TRIED AND TESTED METHODS

- First Model:

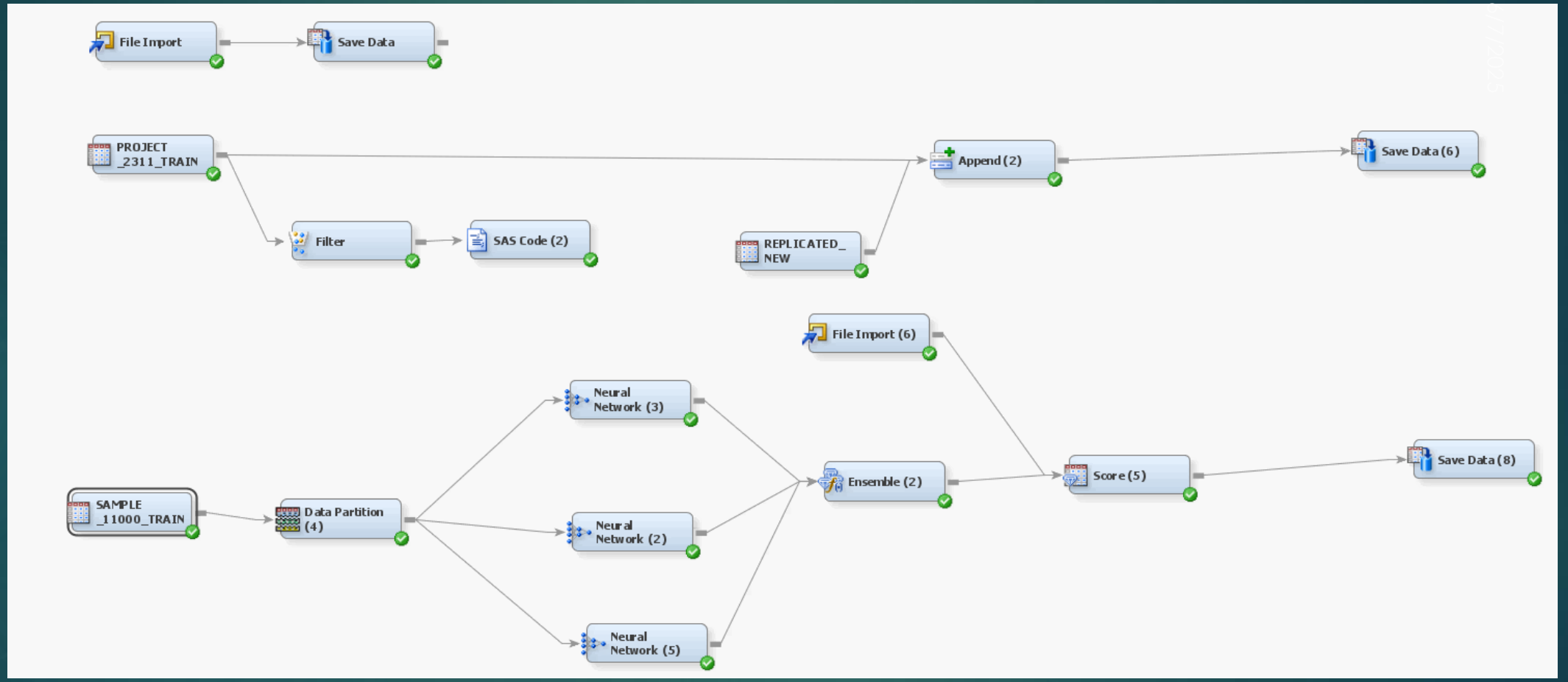


- Second Model:

```
*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*


Summary Statistics for Class Targets
(maximum 500 observations printed)

Data=DATA

            Numeric     Formatted     Frequency
Variable     Value       Value         Count       Percent     Label

 class         0           0            9789        97.89
 class         1           1             211         2.11


Data=SAMPLE

            Numeric     Formatted     Frequency
Variable     Value       Value         Count       Percent     Label

 class         0           0             211          50
 class         1           1             211          50
```



PROJECT WORKFLOW

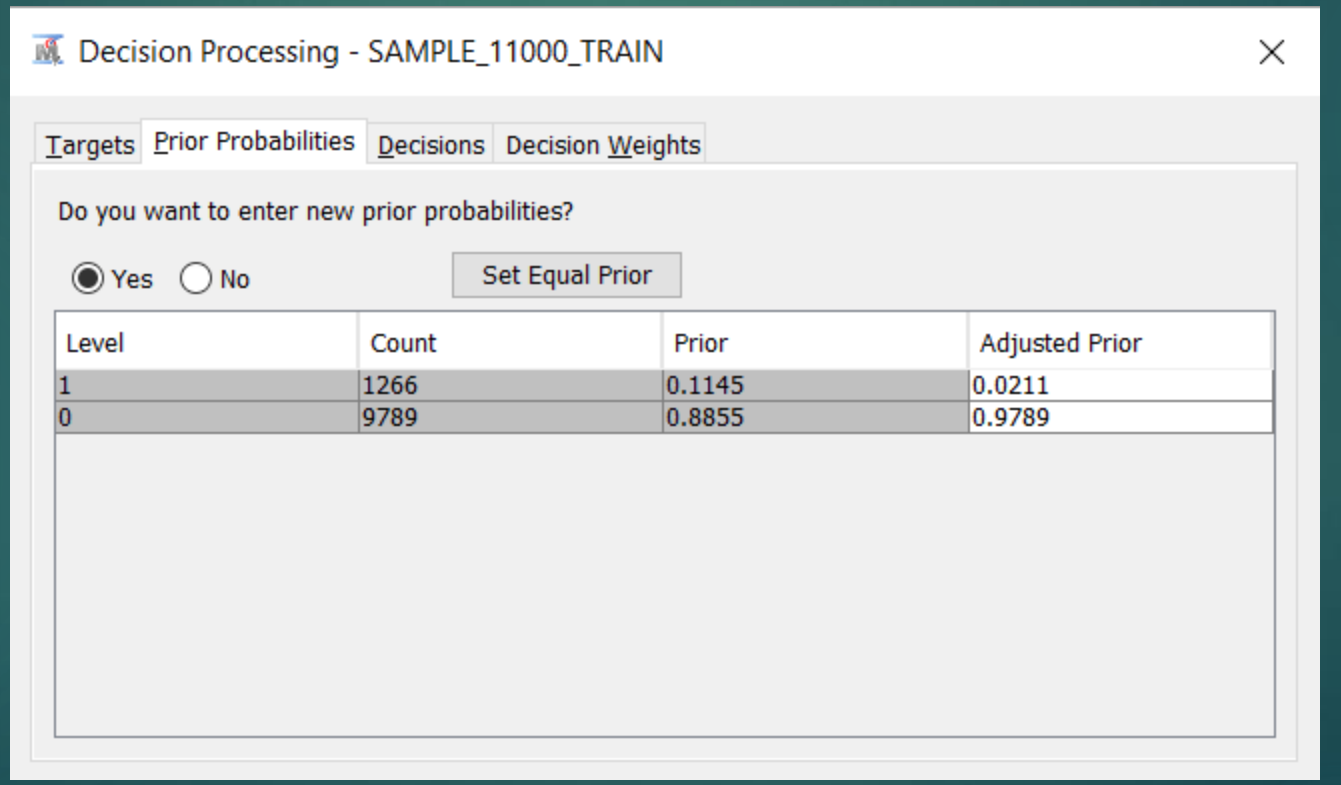


FINAL MODEL – DATA PREPROCESSING

- **Class imbalance issue:** Initially, the dataset had 211 bankrupt firms (Class=1) compared to 9789 non-bankrupt firms (Class=0), making it highly skewed.
- **Oversampling technique:** Filtered Class=1 records and replicated them 5 times using SAS code node to balance the dataset.
- **Adjusting Prior Probabilities:** Aligned the model with real-world class distribution (Class=1: 2.11%, Class=0: 97.89%) to prevent bias from oversampling and ensure reliable predictions.
- **Appended oversampled data:** Combined the replicated data with the original dataset to ensure a balanced training set.

```
data replicated;
    set &EM_IMPORT_DATA;
    do i = 1 to 5; /* Replicate each row 30 times */
        output;
    end;
run;
```

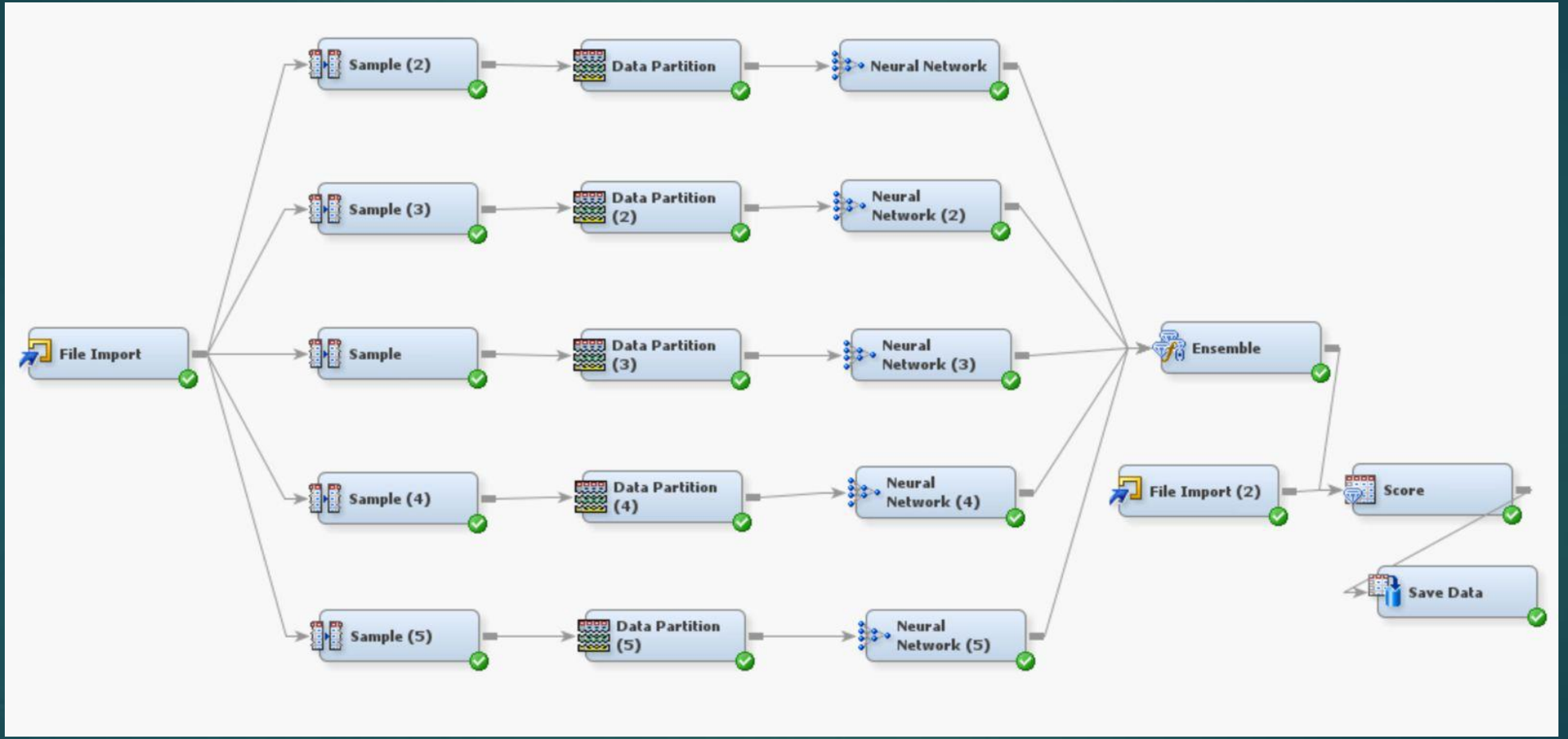
```
/* Save the replicated data to SASUSER */
data SASUSER.REPLICATED_NEW; /* Replace SASUSER.REPLICATED_OUTPUT with your desired dataset name */
    set replicated;
run;
```



MODEL BUILDING

- **Neural Networks setup:** Three Neural Networks with different configurations: Iterations (150, 150, 100) and seeds.
- **Ensemble method:** Combined the outputs of individual Neural Networks to create a robust final model.
- **Optimized for performance:** Improved generalization and achieved a high accuracy score of 94.425.

SECOND BEST MODEL



- **Multiple sampling:** Created 5 samples from the dataset, increasing the percentage of Class=1 from 2% to 11%.
- **Randomization:** Used different random seeds for each sample, ensuring diverse distributions for Class=0.
- **Training and validation:** Partitioned each sample into training and validation datasets.
- **Neural Networks:** Developed Neural Networks with varied configurations (e.g., iterations and levels) for each sample.
- **Model Ensembling:** Combined all 5 Neural Network models into a single ensemble model.
- **Evaluation:** Compared the ensembled model against the test data, achieving a slightly lower score of 94.05 compared to the first model.

CONCLUSION

This project provided an engaging and insightful opportunity to apply advanced machine learning techniques learned in class, exploring their potential to predict firm bankruptcy using financial indicators. Through **Oversampling**, **Prior probability** adjustment, and **Ensemble** modeling, we successfully tackled class imbalance and developed a reliable model for predicting firm bankruptcy. We achieved a prediction accuracy of **96.088 on the private leaderboard**. By tackling the challenge of class imbalance through oversampling, we ensured a balanced representation of the minority class within the training data. Additionally, adjusting prior probabilities allowed the model to better reflect real-world class distributions, enhancing its reliability. Multiple Neural Networks with varied configurations were developed to capture diverse data patterns, and their outputs were effectively combined using ensemble modeling, which helped minimize bias and variance.