Planning Logic

Prosperity Prognosticator: ML For Startup Success Prediction

1. Project Phases and Logic

The project is divided into 5 core development phases, each building upon the previous. The phases follow a linear ML development lifecycle from data preparation to deployment.

2. Phase Breakdown

Phase	Phase Name	Activities	Duration
Phase 1	Data Collection & Preparation	Download dataset, clean data, handle nulls, encode features	Week 1
Phase 2	Exploratory Data Analysis	Statistical analysis, visualizations, correlation study	Week 1-2
Phase 3	Model Building	Train 6 algorithms, evaluate, compare accuracy	Week 2-3
Phase 4	Performance Testing & Tuning	GridSearchCV tuning, feature selection, finalize model	Week 3
Phase 5	Model Deployment	Save model, build Flask app, create HTML templates, test	Week 4

3. Decision Logic

3.1 Algorithm Selection Logic

- Train all 6 algorithms on the same dataset split
- Compare accuracy scores using the evaluate_model() function
- Select the algorithm with the highest accuracy for hyperparameter tuning
- Default expectation: Random Forest performs best for tabular classification

3.2 Feature Selection Logic

- After tuning, extract feature_importances_ from the Random Forest model
- Rank all features by importance score in descending order
- Select the top 10 most important features

- Retrain the model using only selected features
- Compare accuracy before and after feature selection

3.3 Model Saving Logic

- Save the final tuned model to random_forest_model.pkl using pickle.dump()
- Save the top feature list to features.pkl for use in app.py form generation
- Validate the saved model by loading and running a test prediction

4. Risk Planning

Risk	Likelihood	Impact	Mitigation
Low model accuracy	**Medium**	**High**	**Apply hyperparameter tuning and feature selection**
Missing/null data in CSV	**High**	**Medium**	**Fill nulls with median; drop columns with >50% nulls**
Flask routing errors	**Low**	**High**	**Test all routes manually before final submission**
Feature mismatch in prediction	**Medium**	**High**	**Save feature list as features.pkl and use consistently**
Overfitting	**Medium**	**High**	**Use cross-validation and test on hold-out test set**