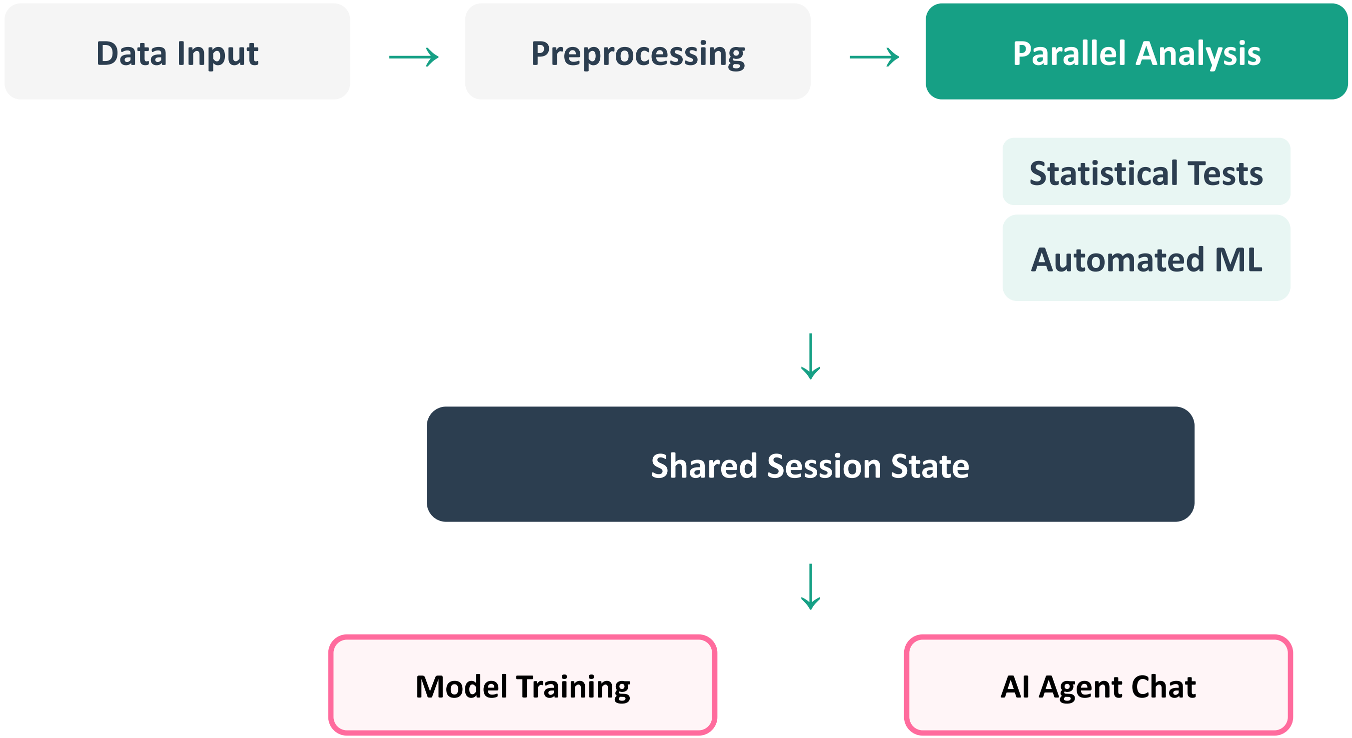
Statistical AI Agent for Dataset Analysis

A Parallel-Validation Statistical Analysis System
with AI-Powered Chat Interface

Lei GAO (s327756) • Lan DENG (s338219)

System Architecture Overview



Core Design Principles

Shared Analysis Cache

Eliminate duplicate computation by storing analysis results

Parallel Feature Ranking

Statistical and ML methods run independently for unbiased comparison

Branching Consumption

One analysis, multiple uses: Training + Chat both consume shared results

Session State Management

Cross-page data sharing enables flexible workflows

Data Processing Pipeline

CSV Upload → Data Validation → Preprocessing

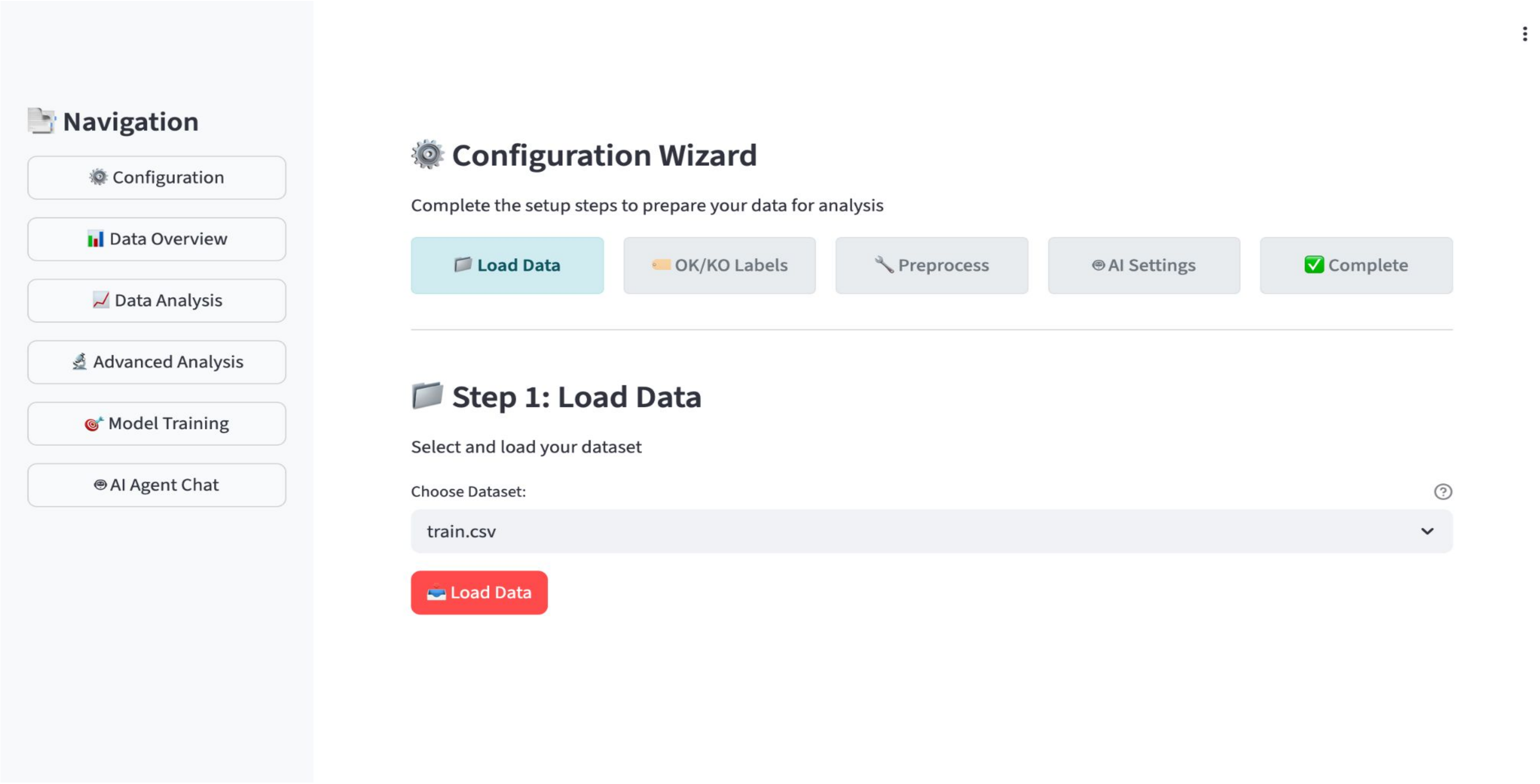
Preprocessing Steps

- Missing Values: mean/median/mode/drop
- Encoding: Label Encoding / One-Hot Encoding
- Scaling: StandardScaler / MinMaxScaler
- Label Creation: OK/KO Classification

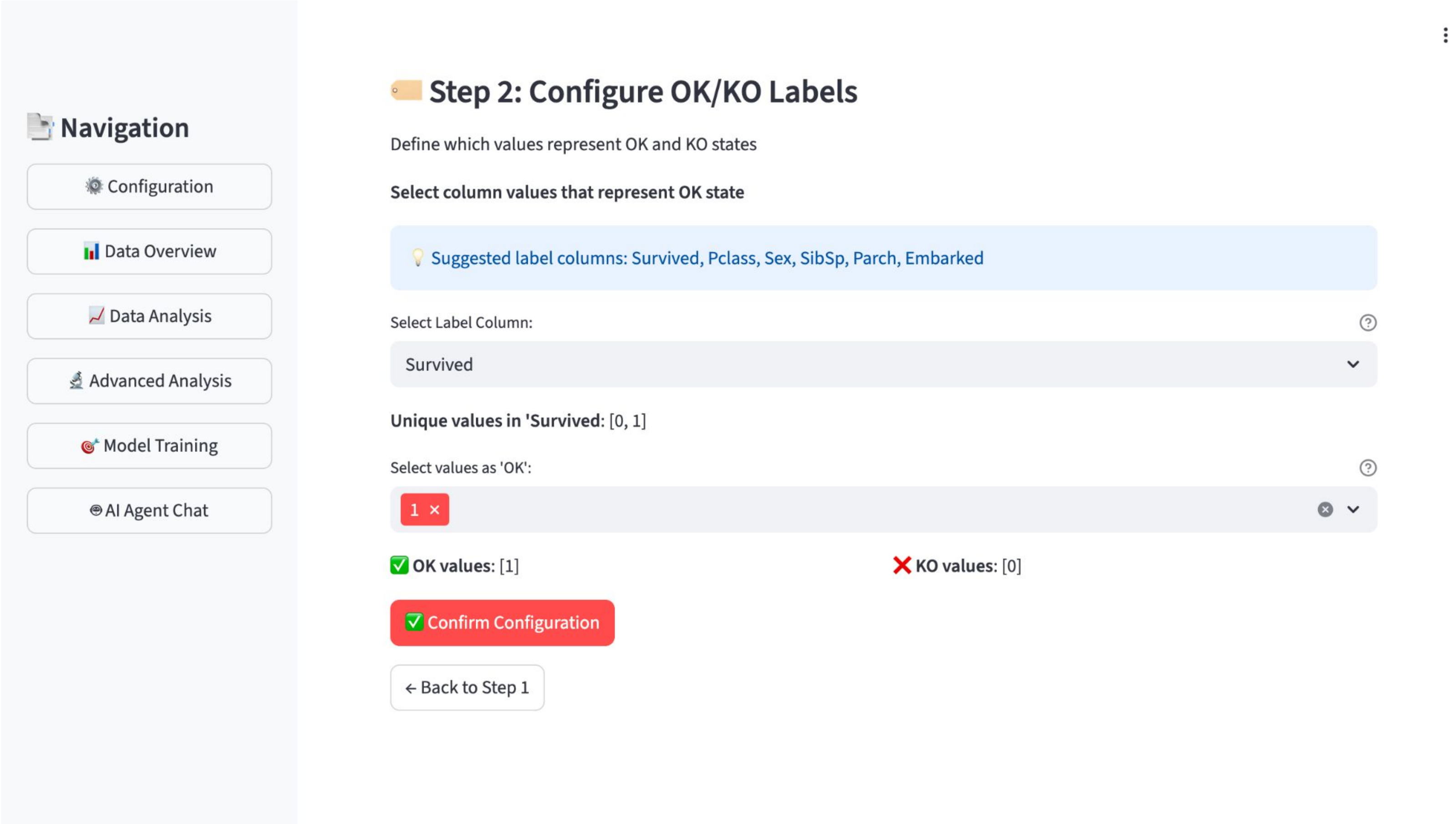
Key Parameters

- Missing Strategy
- Encoding Method
- Scaling Technique

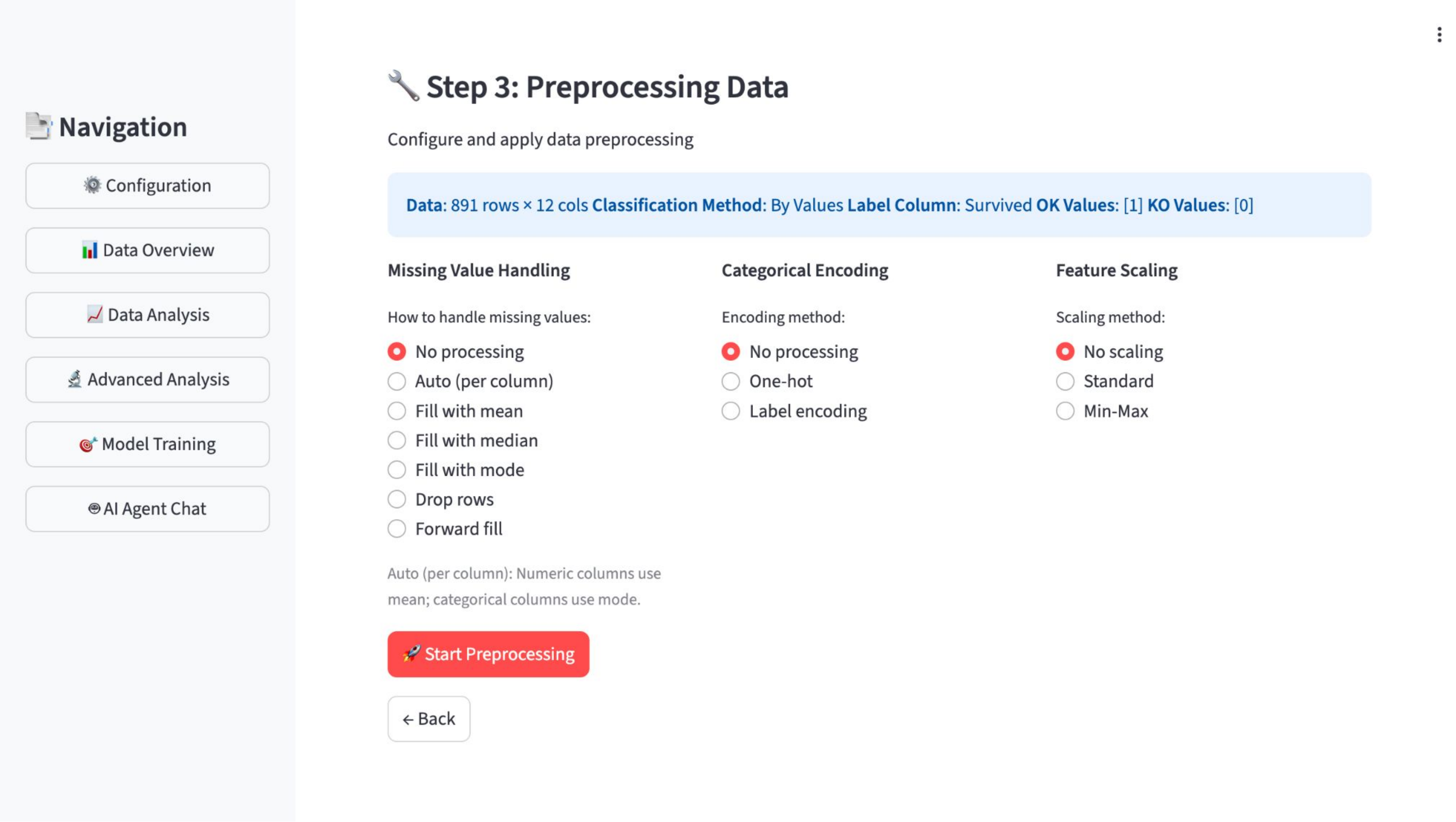
Configuration Interface



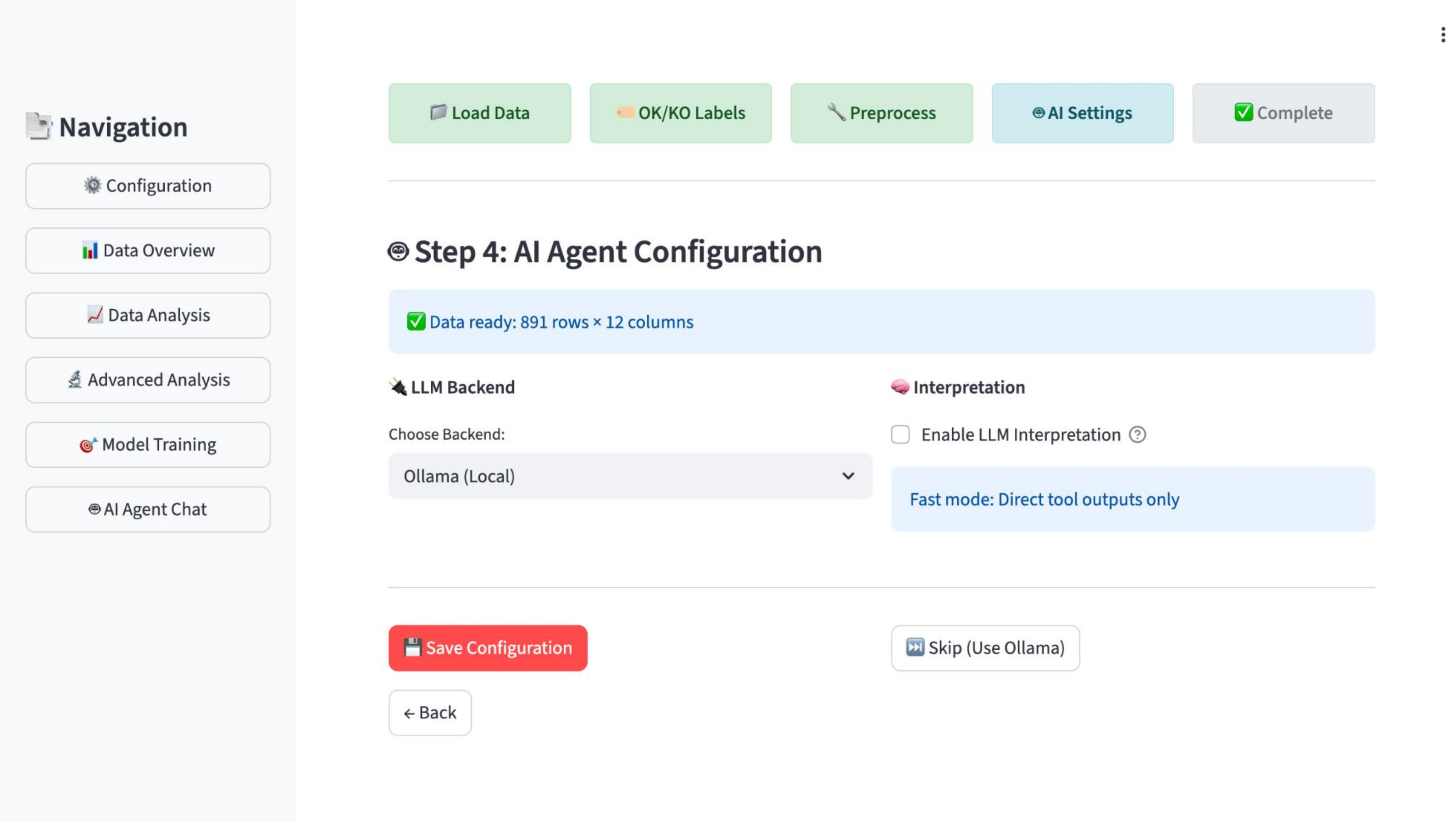
Configuration Interface



Configuration Interface



Configuration Interface



Configuration Interface

📑 Navigation

- ⚙️ Configuration
- 📊 Data Overview
- 📈 Data Analysis
- 🔬 Advanced Analysis
- 🎯 Model Training
- 🤖 AI Agent Chat

📋 Next Steps

✅ Configuration complete! You can now:

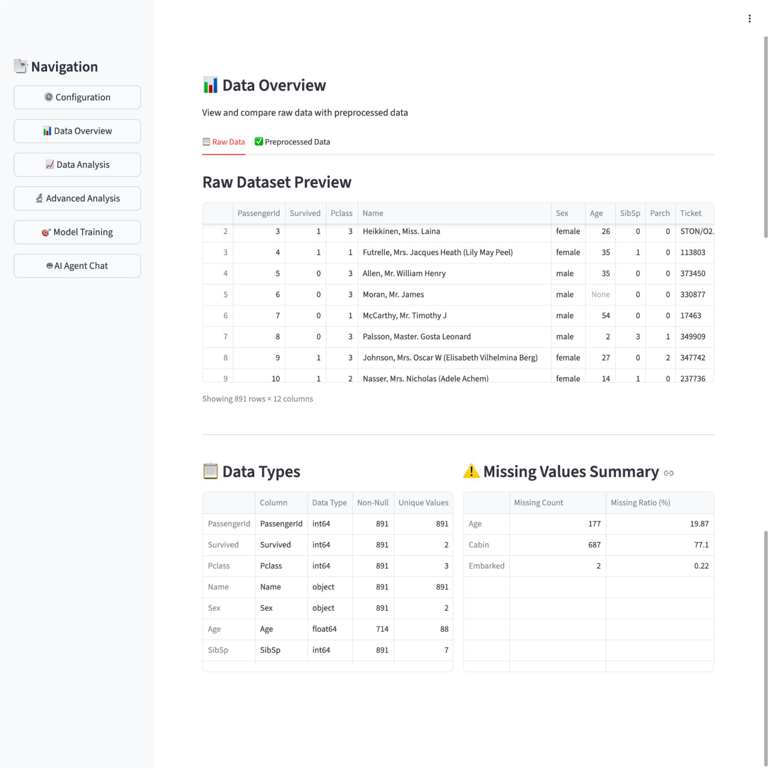
1. 📊 **Data Overview** → View raw data and preprocessing results
2. 📈 **Data Analysis** → Explore features and distributions
3. 🔬 **Advanced Analysis** → Run AutoGluon feature importance
4. 🎯 **Model Training** → Train discriminative models
5. 🤖 **AI Agent Chat** → Ask questions in natural language

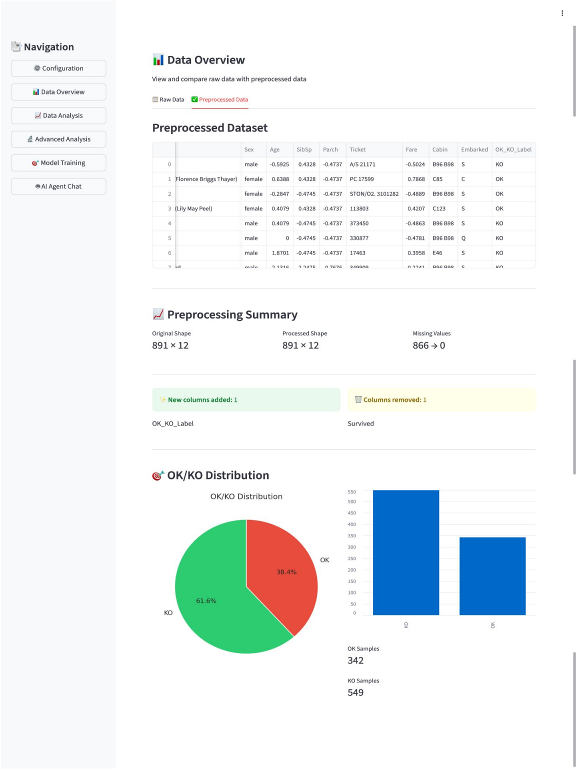
🔄 Edit Configuration

Need to change something? You can return to any step:

📁 Step 1: Load Data | 🏷️ Step 2: Labels | 🔧 Step 3: Preprocess | 🤖 Step 4: AI Settings

Data Overview





Dataset statistics and initial exploration

Data Analysis

Column	Type	Data Type	Unique	Available	Status
PassengerId	Numerical	int64	891	❌	Too many unique (891)
Pclass	Numerical	int64	3	✅	Available
Name	Categorical	object	891	❌	Too many categories (891)
Sex	Categorical	object	2	✅	Available
Age	Numerical	float64	88	✅	Available
SibSp	Numerical	int64	7	✅	Available
Parch	Numerical	int64	7	✅	Available
Ticket	Categorical	object	681	❌	Too many categories (681)
Fare	Numerical	float64	248	✅	Available
Cabin	Categorical	object	147	❌	Too many categories (147)
Embarked	Categorical	object	3	✅	Available
OK_KO_Label	Label	object	2	🎯	Target variable

Data Analysis

Compare feature distributions between OK and KO groups

› 📊 Dataset Summary & Feature Availability

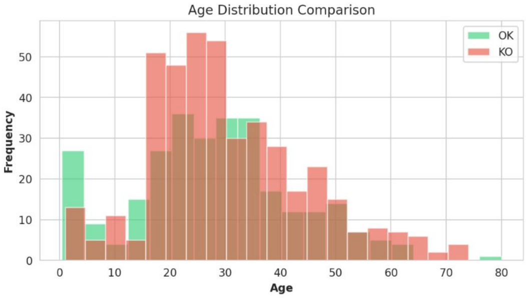
📊 Numerical Features 🏷️ Categorical Features

Select numerical features to analyze:

Age ×

Age - Statistical Comparison

	OK	KO
Mean	28.3437	30.6262
Std Dev	14.951	14.1721
Median	28	28



› 📊 Dataset Summary & Feature Availability

📊 Numerical Features 🏷️ Categorical Features

Select categorical features to analyze:

Sex ×

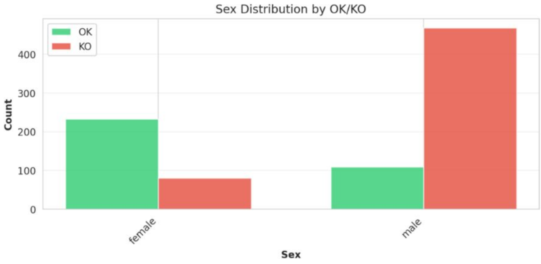
Sex - Category Distribution

Count Table:

Sex	KO	OK
female	81	233
male	468	109

Percentage Table:

Sex	KO	OK
female	14.75%	68.13%
male	85.25%	31.87%



Parallel Validation Approach

Statistical Ranking

Method
Hypothesis Testing

Metrics
p-value + Effect Size

Score
$-\log_{10}(p) \times$ effect

Interpretable & Theory-Based

VS

ML Ranking

Method
Ensemble Learning

Metrics
Permutation Importance

Score
Model-based Importance

Data-Driven & Predictive

Statistical Ranking

Feature Type	Test Method	Metrics	Purpose
Numeric	T-test	p-value, Cohen's d	Mean difference (parametric)
Numeric	Mann-Whitney U	p-value, Effect size	Rank difference (non-parametric)
Numeric	Kolmogorov-Smirnov	p-value, KS statistic	Distribution difference
Categorical	Chi-square	p-value	Independence test
Categorical	Cramér's V	Effect size (0-1)	Association strength

Composite Ranking Score

$$\text{Score} = -\log_{10}(\text{p-value}) \times \text{effect\_size}$$

ML Ranking

AutoGluon Framework

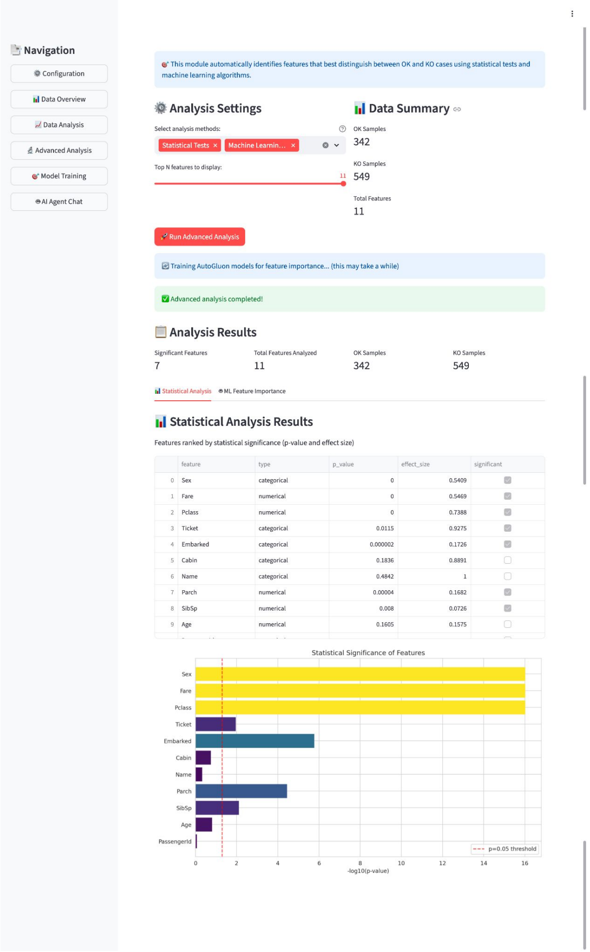
TabularPredictor
Automated ML ensemble training

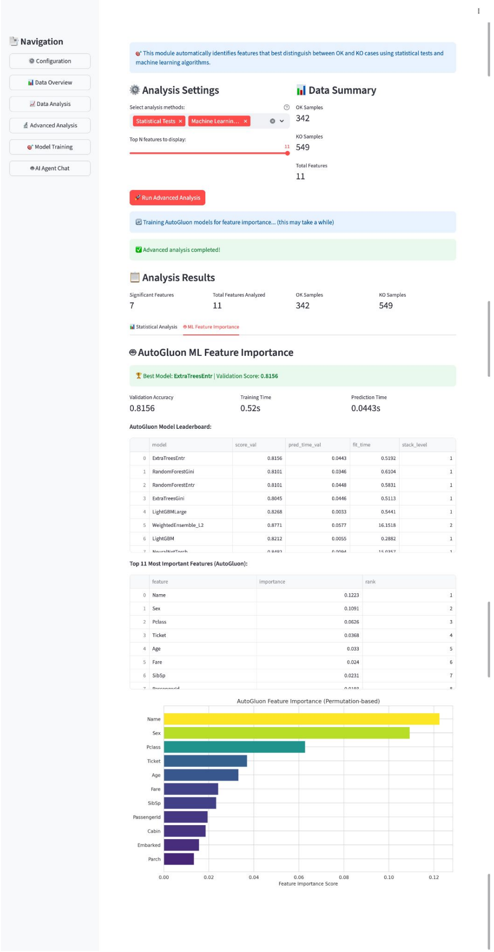
Ensemble Methods
WeightedEnsemble_L2 as best model

Permutation-Based
Feature importance via shuffling

Training Metrics

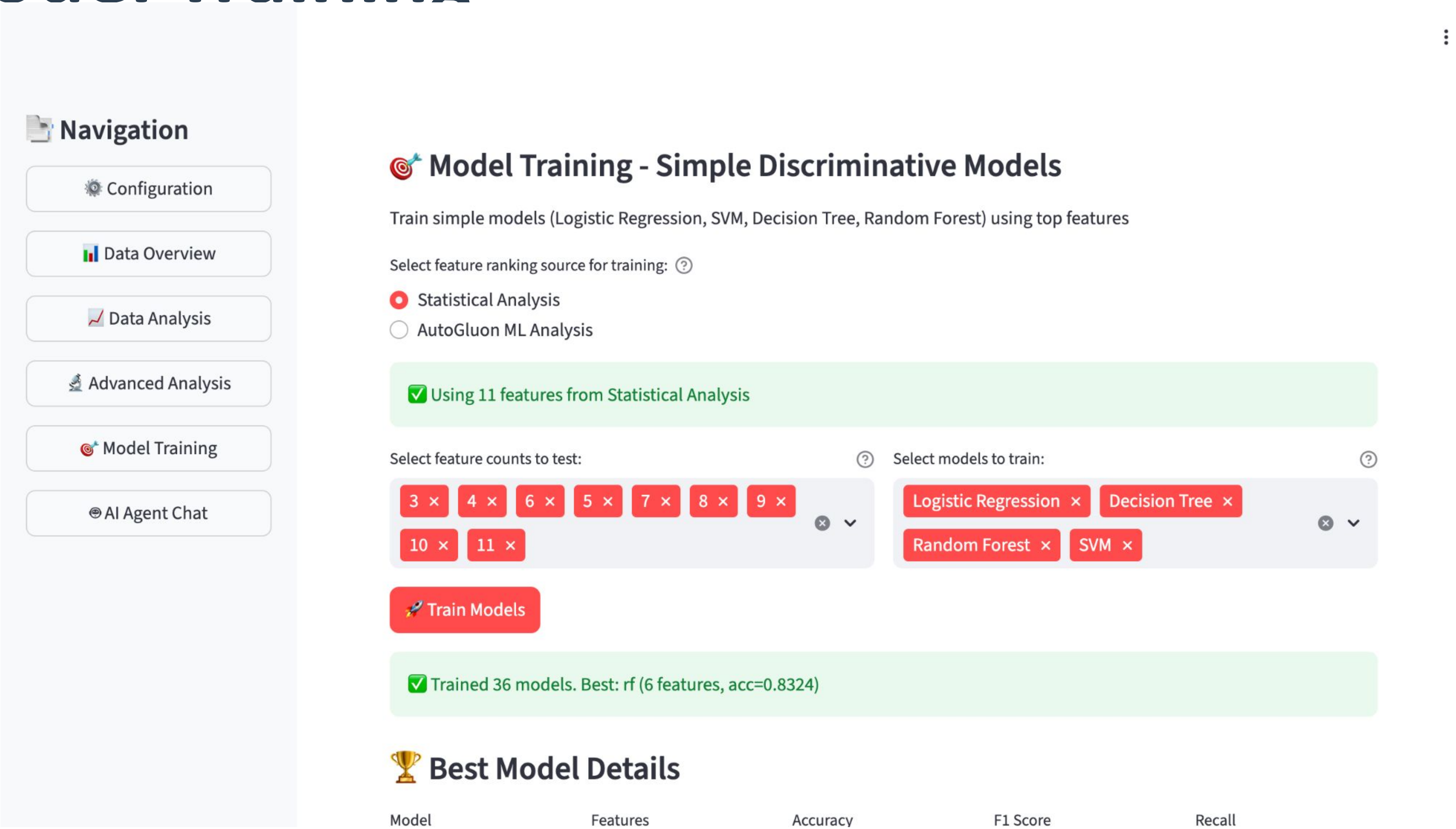
Training Time	**~60-120s**
Best Model	**WeightedEnsemble_L2**
Validation Accuracy	**0.89**





Model performance comparison: Different algorithms × Feature selection methods

Model Training

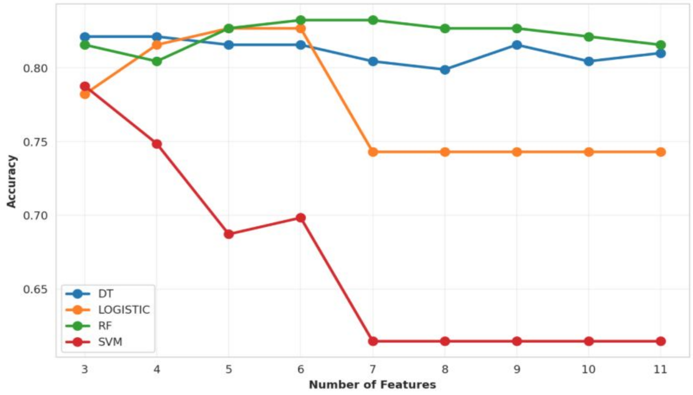


Model Training Results - Statistical

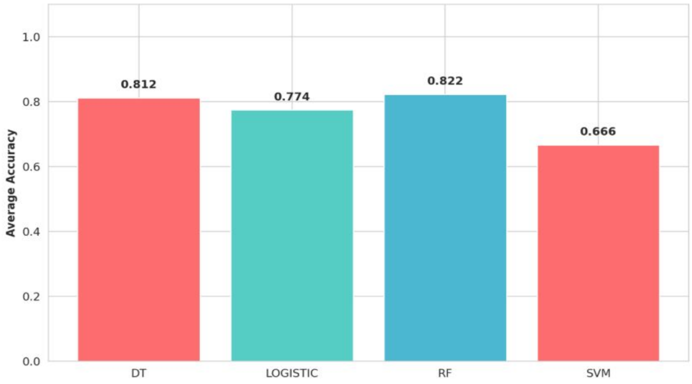
🏆 Best Model Details

Model	Features	Accuracy	F1 Score	Recall
RF	6/11	0.8324	0.7727	0.6892

📈 Accuracy vs Feature Count



🎯 Model Comparison



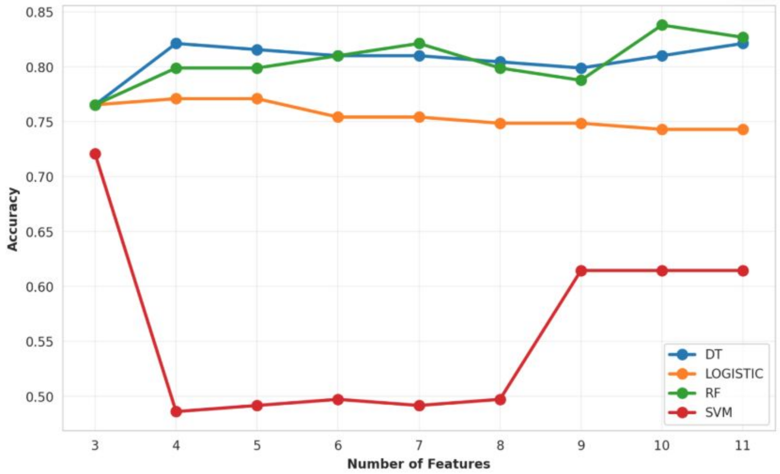
Model performance comparison: Different algorithms × Feature selection methods

Model Training Results - AutoGluon

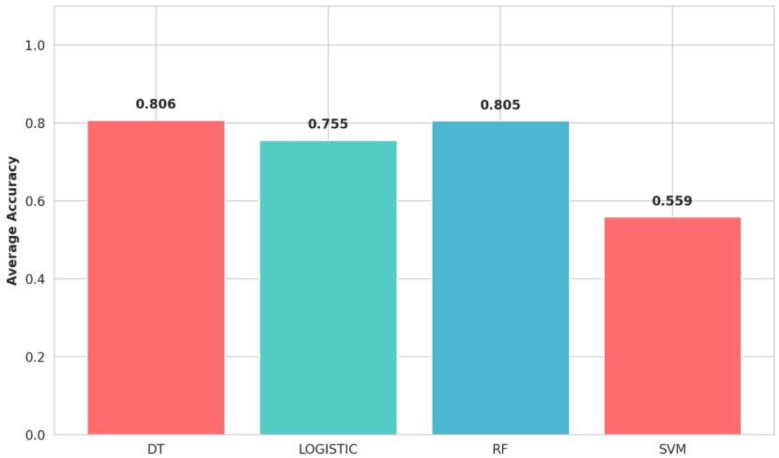
🏆 Best Model Details

Model	Features	Accuracy	F1 Score	Recall
RF	10/11	0.8380	0.7786	0.6892

📈 Accuracy vs Feature Count



🎯 Model Comparison



AI-Powered Chat Interface: Architecture

User Natural Language Query

↓

StatisticalAgent (Intent Parser)

Rule-based + LLM Fallback (LLaMA-3)

↓

Tool Routing

- get_feature_importance()
- get_statistical_summary()
- plot_distribution()
- plot_time_series()
- plot_frequency_spectrum()

Data Context

← Session State

- processed_df
- analysis_results
- feature_ranking
- statistical_summary

↓

Response Generation: Plot + Text Summary

LLM Backend Design

Backend	Deployment	Latency	Use Case
Ollama (LLaMA-3)	Local	2-5s	Privacy-first deployment
OpenAI GPT-4	Cloud	0.5-1s	High accuracy research
Claude / Gemini	Cloud	0.5-1s	Alternative cloud options
DeepSeek	Cloud	1-2s	Cost-effective option

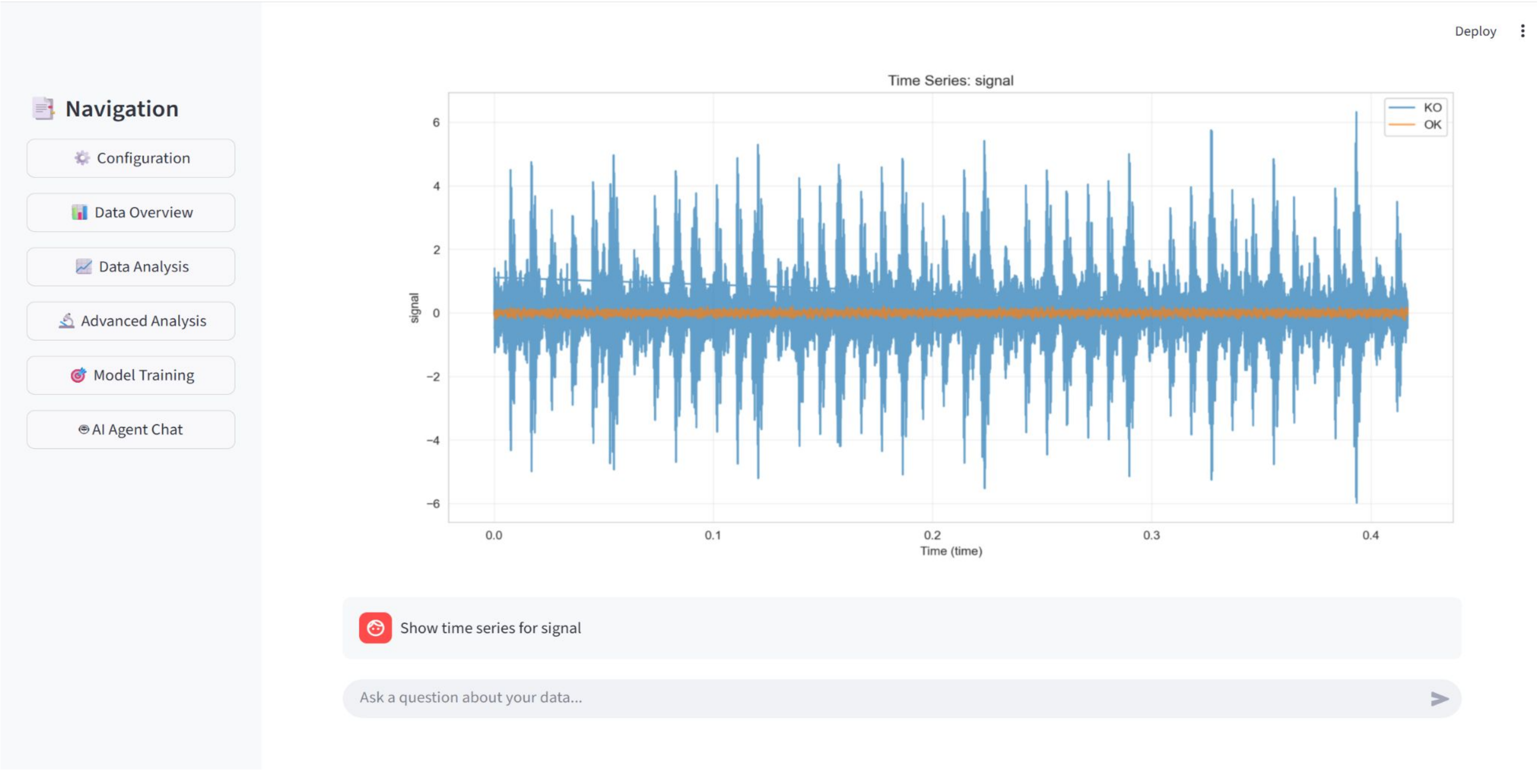
ConversationManager

Chat History
Maintains context

Context Window
Manages tokens

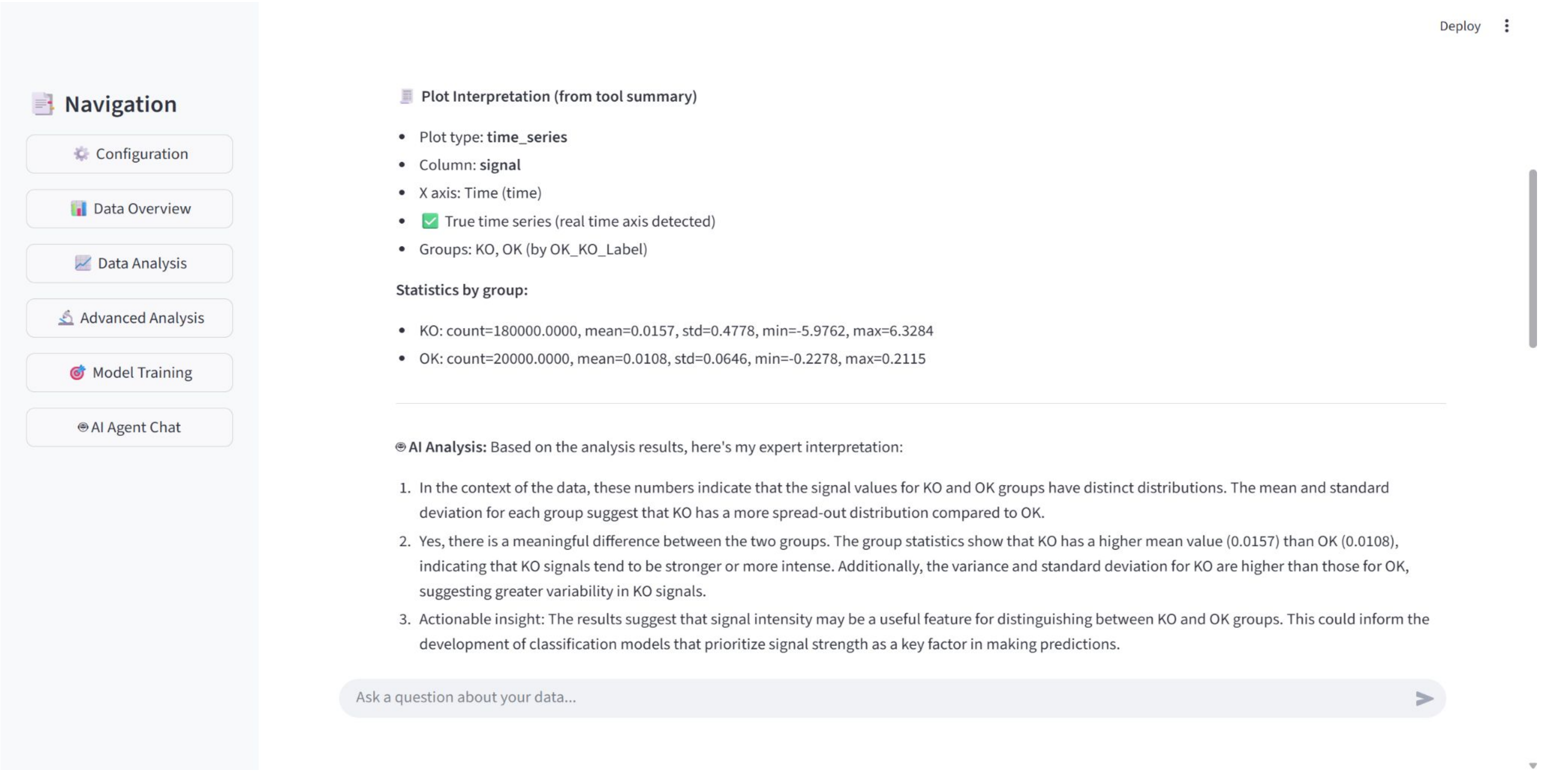
Session State
Links to analysis

AI Agent Use Case : Time Series Visualization



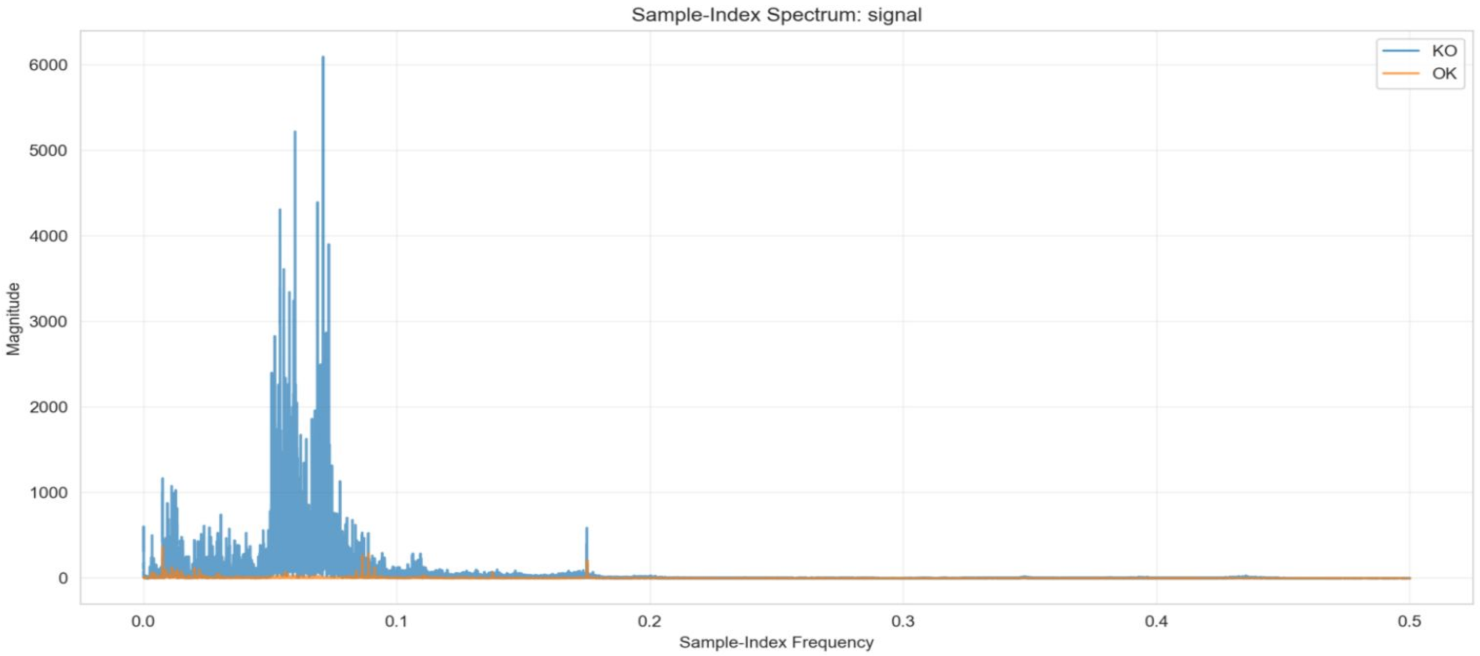
User query: "Compare distribution of rms between OK and KO"

AI Agent Use Case : Time Series Visualization



User query: "Show time series of signal"

AI Agent Use Case: Frequency Spectrum (FFT)



Plot FFT for signal

User query: "Show frequency spectrum" - FFT with peak detection

AI Agent Use Case: Frequency Spectrum (FFT)

✅ Generated frequency spectrum (FFT) plot for **signal**

🧾 **Plot Interpretation (from tool summary)**

- Plot type: **sample_index_spectrum**
- Column: **signal**
- ⚠️ Feature table data (NOT physical waveform)
- Groups: KO, OK (by OK_KO_Label)

Dominant peaks (sample-index):

- KO: 0.07, 0.07, 0.07
- OK: 0.01, 0.09, 0.09

Note: ⚠️ Feature table spectrum (NOT physical frequency). This shows patterns in sample order, not real Hz. Top-5 peaks by magnitude.

🤖 **AI Analysis:** Based on the analysis results:

1. The top-5 peaks by magnitude in the feature table spectrum indicate patterns in sample order, which may reflect differences in signal characteristics between KO and OK groups.
2. Since groups exist (KO, OK) and there are meaningful differences between them, this suggests that the signal features captured in the spectrum can be used to distinguish between KO and OK samples.
3. The actionable insight is that these signal features can potentially be used as input features for a classification model to predict whether a sample belongs to the KO or OK group.

User query: "What are the statistics for rms?" - Multi-metric view

AI Agent: Technical Highlights

Data Encapsulation

No DataFrame in prompts ⟶ prevents hallucination

Robust Time Handling

Auto-detects time columns with graceful fallback

Context-Aware

Receives full analysis results for comprehensive queries

Flexible Visualization

Natural language ⟶ plot type selection automatically

Structured Output Pattern

Plot + Summary + AI Interpretation ⟶ Deterministic computation (no AI guessing)

System Design Philosophy

1

Parallel Validation

Statistical + ML methods
run independently
→ user selects best

2

Shared State

One analysis
→ multiple consumers
No duplication

3

Anti-Hallucination

Tool-based execution
→ deterministic results
No AI guessing

Limitations & Future Directions

Current Limitations

- Train/eval on same data (no k-fold)
- Class imbalance not tuned
- AutoGluon training time ~60-120s
- Single-session (no persistence)

Future Improvements

- Cross-validation + proper train/test split
- SMOTE/class weights for balancing
- Hyperparameter tuning UI controls
- Session persistence & report export

Thank You

Key Contributions

✓ Parallel-validation architecture for feature ranking

✓ Multi-backend LLM integration with local deployment

✓ Context-aware AI agent with anti-hallucination design

✓ Interactive visualization toolkit for exploratory analysis

Lei GAO (s327756) • Lan DENG (s338219)