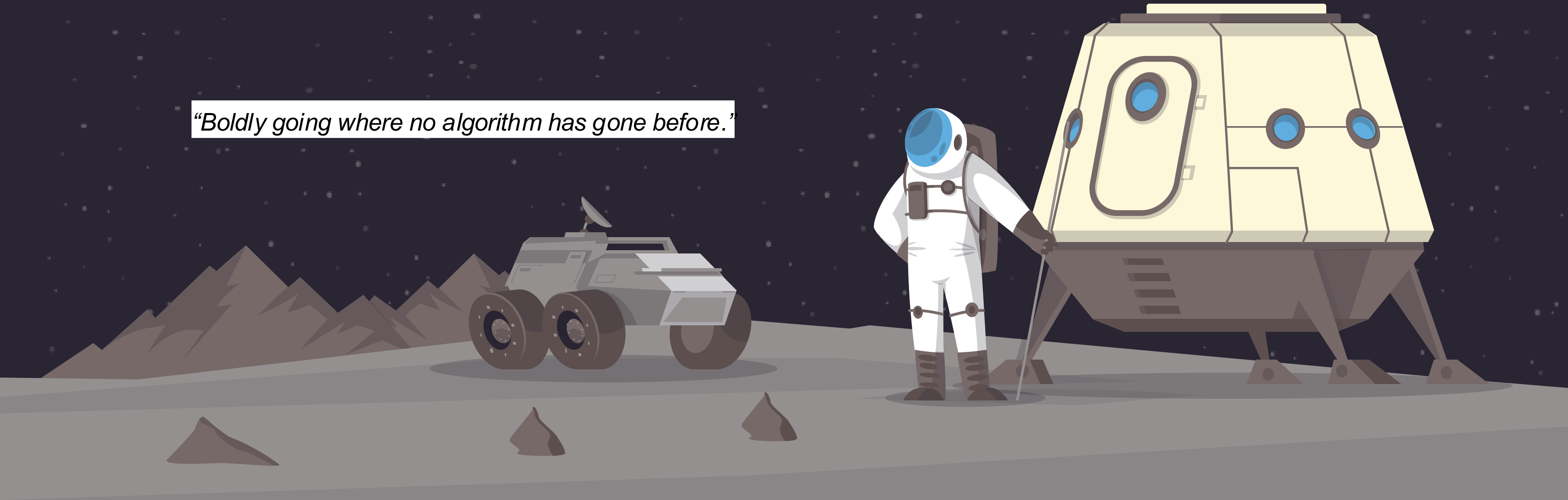
S'kaiNet - Exoplanet Analyzer

By **Team Outlander**

"Boldly going where no algorithm has gone before."



Inspiration

Our project draws inspiration from the spirit of exploration found in **Star Trek** and the logic of **Vulcan philosophy**. The name **S’kai** means “*discovery*” in the Vulcan language — a perfect symbol of our mission to seek new worlds through science and reason. By combining NASA’s real exoplanet data with AI, we imagined how future Starfleet-like technologies could identify habitable planets beyond our solar system. **S’kaiNet** is our bridge between imagination and science — where curiosity meets logic in the pursuit of discovery.

Our Solution

Full pipeline: load → clean → engineer features → grouped CV → train ensembles → evaluate → threshold tuning → interpretability.
- **Group-aware validation** by `kepid` to prevent leakage across KOIs from the same star.
- **Astrophysical features**: ratios, geometry, SNR proxies, uncertainty-aware measures.
- **Strong tabular learners**: XGBoost, LightGBM, CatBoost, RandomForest.
- **Advanced training modes**:
- **Stacking Ensemble** — blend your base learners with a Logistic Regression meta-learner
- **Multi-Step (Hierarchical) Pipeline** — high-recall planet filter (MLP) → high-precision planet type (XGBoost)
- **Binary Planet Model** — simplified CONFIRMED vs FALSE POSITIVE experiment
- **Planet-centric metrics** and **per-class threshold optimization** to increase planet recall under constraints.

Contributions

- **Label leakage → inflated accuracies**
  Inclusion of post-hoc or human-informed fields (e.g., `kepler_name`, `koi_pdisposition`, `koi_score`, and KOI false-positive flags) leaks disposition information into training, producing unrealistically high metrics that won't hold in deployment.

- **Split strategies that don't respect astrophysical grouping**
  Random/stratified CV that ignores `kepid` allows KOIs from the same star to appear in both train and validation, letting models memorize star-specific quirks.

- **Single-stage argmax decisions**
  Using plain `argmax` on multiclass probabilities under-controls discovery/false-alarm trade-offs—especially harmful when **planet recall** is the scientific priority.

- **Limited feature engineering**
  Raw KOI columns miss physically motivated relationships (geometry, SNR normalization, uncertainty ratios) that help separate real transits from systematics.

- **Class imbalance not explicitly handled**
  Without class weighting, minority classes (planets) can be under-served by decision rules and selection metrics.

- **Reproducibility gaps**
  Absent or inconsistent feature lists, seed handling, and fold definitions make results hard to replicate or compare fairly.

Results and Performance

Dataset:
- Trained on **Kepler KOI dataset** — 9,564 valid samples
- 67 total features: 36 existing + 31 newly engineered

Binary Classification Model:
- Default: **95.3% Accuracy**, **92.1% Recall**, **92.7% Precision**
- Planet Detection (Confirmed + Candidate): **96.7% Recall**, **96.4% Precision**

Multistep Model (MLP + XGBoost):
- Default: **88% Accuracy**, **88% Recall**, **87.5% Precision**
- Without dropping koi_fpflag_ columns: **91.2% Accuracy**, **91.2% Recall**, **91.1% Precision**

Ensemble Models:
- XGBoost: **78% Accuracy**, **78% Recall**, **78.2% Precision**, **91.9% Train Accuracy**
- LightGBM: **77.7% Accuracy**, **77.7% Recall**, **78% Precision**, **92.9% Train Accuracy**
- CatBoost: **73.6% Accuracy**, **73.6% Recall**, **77.6% Precision**, **87.4% Train Accuracy**
- RandomForest: **77.4% Accuracy**, **77.4% Recall**, **77% Precision**, **92.7% Train Accuracy**

Stacked Ensemble Model:
- Default: **78.6% Accuracy**, **79% Recall**, **77% Precision**

Our Team

Jemshit
Iskanderov

Nurmyrat
Amanmadov

Tarlan
Abdullayev

Parahat
Iljanov

Links

App: https://skainetweb.vercel.app/

GitHub: https://github.com/jemshit/NASA_exoplanet_detection