Getting Started with Kaggle

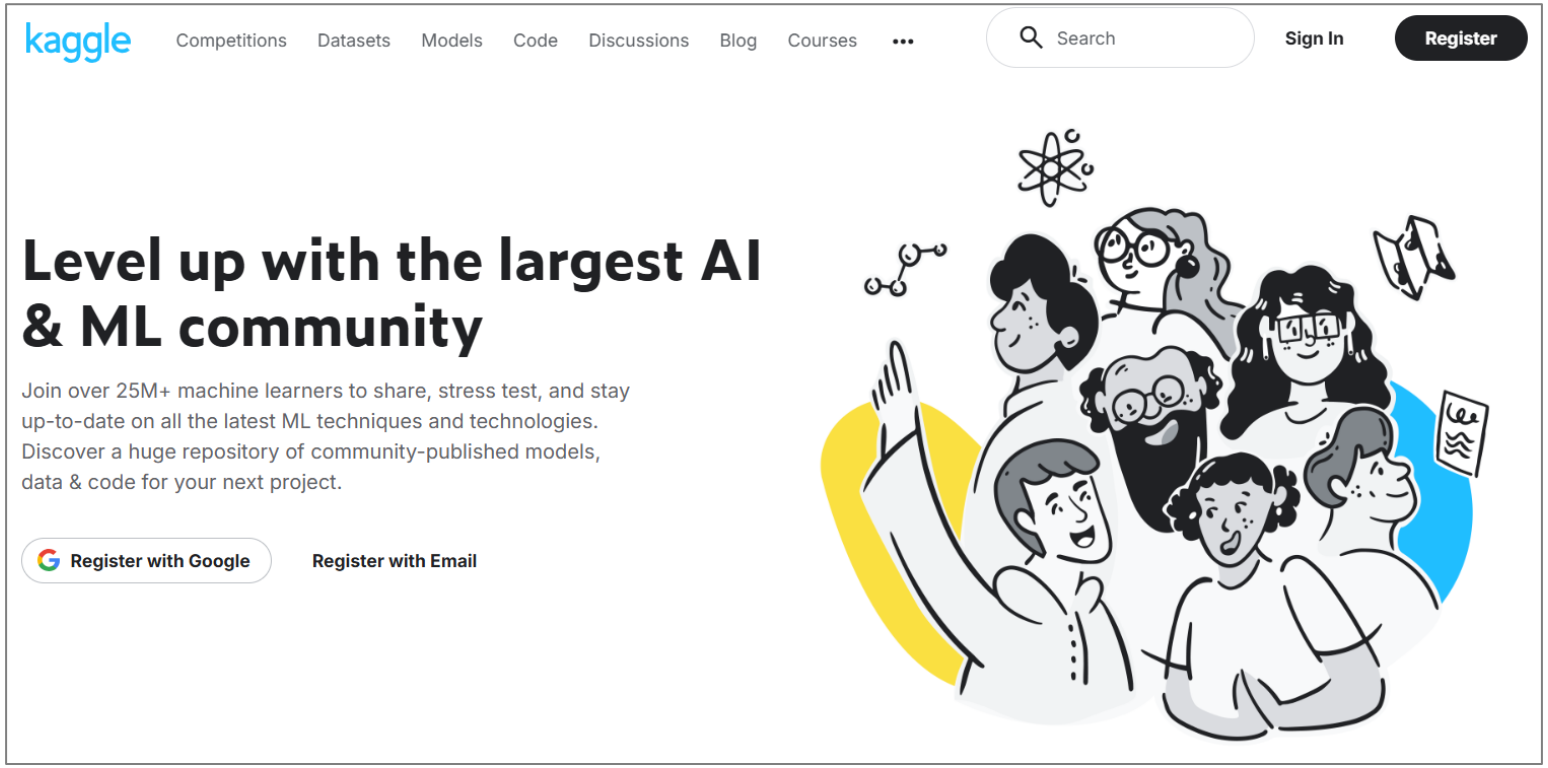
Iftekharul Islam

Agenda

- Intro to Kaggle & Competitions
- Titanic Demo (Minimal Walkthrough)
- Project Options Overview
- Project Requirements & Deliverables
- Tips to Improve Performance

Kaggle Overview

- Online platform for **machine learning competitions**
- Datasets, notebooks (with free GPUs), leaderboards
- Great for learning by doing

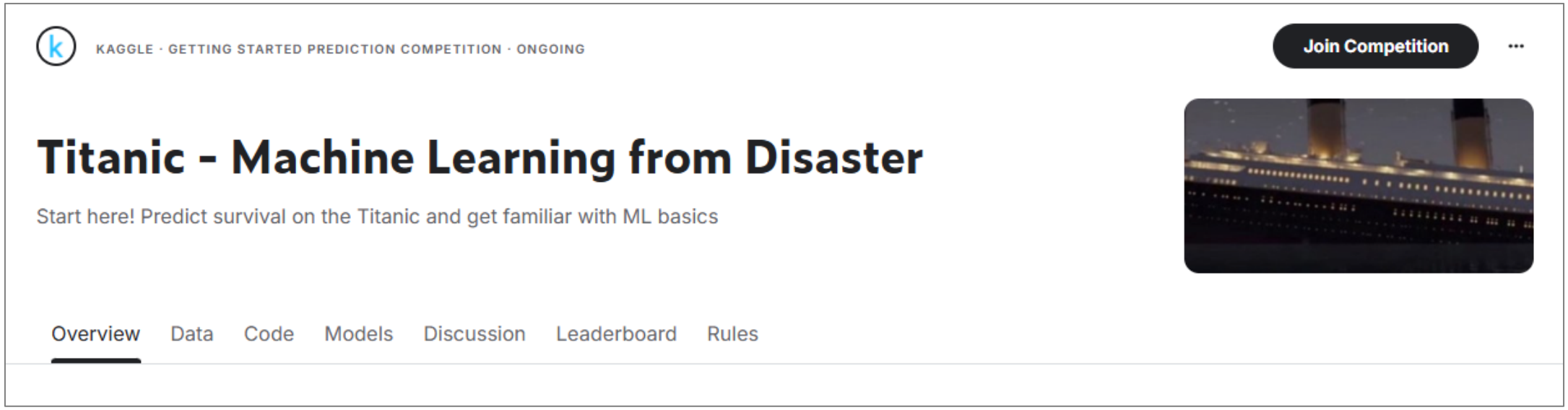


Titanic Walkthrough (Overview)

- Join competition & open Kaggle notebook
- Load dataset
- Minimal preprocessing
- Training Random Forest
- Create submission file
- Upload predictions

Join Competition

- Go to [Titanic Competition Page]
- Click **Join Competition**
- Open **Code → New Notebook**



Load & Inspect Data

```
import pandas as pd
train_data = pd.read_csv("/kaggle/input/titanic/train.csv")
test_data  = pd.read_csv("/kaggle/input/titanic/test.csv")

train_data.head()
```

Train Random Forest

- A **Random Forest** is made up of many **decision trees**.
  - Each tree looks at different parts of the data and makes its **own prediction**.
  - Once all the trees have "voted," the forest takes the **majority decision**.
  - For example, if 70 out of 100 trees predict "survived," the final prediction is **"survived."**
- This approach reduces the risk of **overfitting** and usually performs better than a single decision tree.

Train Random Forest

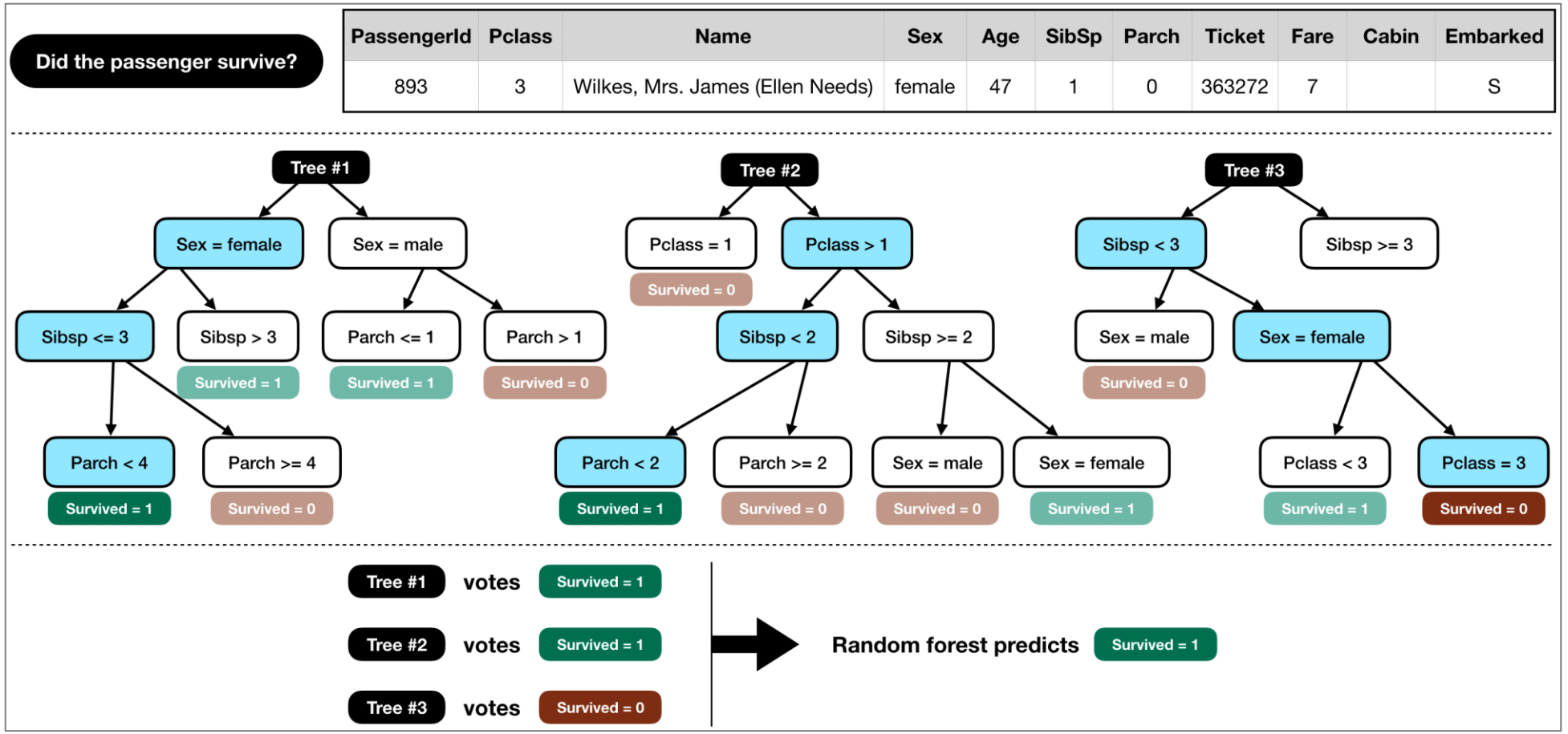


Image: https://www.kaggle.com/code/alexisbcook/titanic-tutorial

Train Random Forest

```
from sklearn.ensemble import RandomForestClassifier


y = train_data["Survived"]
features = ["Pclass", "Sex", "SibSp", "Parch"]
X = pd.get_dummies(train_data[features])
X_test = pd.get_dummies(test_data[features])


model = RandomForestClassifier(n_estimators=100, max_depth=5, random_state=1)
model.fit(X, y)
```

Predict & Save Submission

```
predictions = model.predict(X_test)
output = pd.DataFrame({"PassengerId": test_data.PassengerId, "Survived": predictions})
output.to_csv("submission.csv", index=False)
print("submission.csv saved! Ready to submit on Kaggle.")
```

Submit Prediction

- Once your notebook is ready, click "**Save Version"** in the top-right corner.
- In the pop-up window, choose "**Save and Run All"** so the entire notebook executes from start to finish, then click Save.
- After the run is complete, you'll see the new version listed on the right side of the screen. Click the **ellipsis (...)** next to the latest version and select **"Open in Viewer."**
- Finally, go to the **Data** (or **Output**) tab at the top of the notebook page and click **"Submit"** to upload your predictions to the competition.

Try These Ideas at Home

- **Explore more features**: Age, Fare, Embarked
- **Handle missing values**
  - Drop rows?
  - Fill with: Mean? Median? Most common value?
- **Feature engineering**: Create new features based on existing features
- **Compare different models**: Logistic Regression, Gradient Boosting etc.

ML Project Workflow

- **Data → EDA → Preprocessing → Baseline Model → Submission → Improvements**
- A more detailed walk-through (EDA, Feature Engineering, more models)

Acknowledgement

- https://www.kaggle.com/code/alexisbcook/titanic-tutorial