**Titanic Survival Prediction**

A machine learning project that predicts whether passengers on the Titanic would have survived the disaster. This is a classic dataset that's perfect for learning data science basics.

**What This Project Does**

Back in 1912, the Titanic sank on its maiden voyage. We're using passenger data to build a model that can predict survival chances based on things like age, gender, ticket class, and other factors. It's fascinating how data can tell the story of this historic tragedy.

**The Dataset**

The data includes information about passengers like:

* Whether they survived (this is what we're trying to predict)
* Their ticket class (1st, 2nd, or 3rd class)
* Age and gender
* How much they paid for their ticket
* Whether they were traveling with family
* Where they got on the ship

There's also some missing information, which is pretty common with historical data. Part of the project involves figuring out how to handle these gaps.

**Getting Started**

You'll need these Python packages:

pandas, numpy, matplotlib, seaborn, scikit-learn

Just install them with pip if you don't have them already:

pip install pandas numpy matplotlib seaborn scikit-learn

For the dataset, you can either:

1. Download it using kagglehub (the code shows you how)
2. Let the script use the built-in seaborn version (it does this automatically if it can't find your downloaded file)

**How Well Does It Work?**

I tested three different algorithms:

* **Random Forest**: Usually gets around 82-85% accuracy. This one's pretty good at handling mixed data types.
* **Logistic Regression**: Simpler but still effective, around 80-83% accuracy
* **SVM**: About 81-84% accuracy, good for more complex patterns

The Random Forest tends to perform best, so that's usually the one it picks.

**What I Learned From the Data**

Some interesting patterns emerged:

* Women had way better survival rates than men (about 74% vs 19%) - the "women and children first" rule was real
* First-class passengers definitely had advantages
* Kids had better chances of survival
* People traveling completely alone or in really big groups didn't do as well
* Higher ticket prices generally meant better survival odds

**Making Predictions**

There's a function called predict\_survival() that you can use to test different scenarios:

# Let's say we have a 25-year-old woman in first class

prediction, probability = predict\_survival(

age=25,

fare=50,

sex='female',

pclass=1

)

You can play around with different combinations to see how survival chances change.

**The Process**

Here's basically what the code does:

1. Loads and explores the data (lots of graphs to understand what happened)
2. Cleans up missing values and creates some new useful features
3. Trains multiple models to see which works best
4. Shows you which factors were most important for survival
5. Gives you a function to make your own predictions

**Things That Could Be Better**

If I were to keep working on this, I'd probably:

* Try tuning the model parameters more carefully
* Add cross-validation to get more reliable accuracy estimates
* Maybe try some newer algorithms like XGBoost
* Build a simple web interface so people could easily try different scenarios
* Look more closely at the cabin data (though there's a lot missing there)

**A Few Notes**

This project is great for learning because the Titanic dataset is well-understood and not too complicated. The patterns in the data really do reflect the historical reality of what happened that night.

The code includes plenty of comments and visualizations, so it's hopefully easy to follow along and understand what's happening at each step.

**Running the Code**

Just run the Python script and it'll walk through the entire process - from loading the data to training models and making sample predictions. The whole thing takes a couple minutes to run and shows you lots of interesting visualizations along the way.

Feel free to modify anything or experiment with different approaches. That's the best way to learn!

This was a really interesting project to work on. There's something powerful about using data science to understand historical events, even tragic ones like the Titanic disaster. The patterns in the data tell a very human story about how social class, age, and gender affected people's chances of survival.