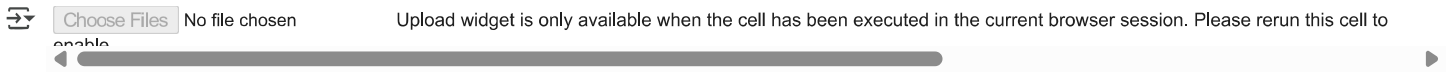
```
from google.colab import files
uploaded = files.upload()
```


```
from google.colab import files
uploaded = files.upload()
```




```
import pandas as pd
import pandas as pd
import pandas as pd
import pandas as pd

# Load dataset (update filename if needed)
df = pd.read_csv('sarathi.....csv')

# Show first few rows
df.head()
df.head()
df.head()
df.head()
```

	Annotator A ID	Annotator B ID	Parition ID	Corpora ID	Sentence ID	Text	Annotator A Text	Annotator B Text	Length	Error	Alignment Score	Agreement
0	1	7.0	2.0	2.0	7506.0	what is the price for round trip between toron...	what is the price for round trip between toron...	what is the price for round trip between toron...	61.0	0.0	1.000000	1.0
1	1	7.0	2.0	2.0	7507.0	trying to find a good deal and good timing to ...	trying to find a good deal and good timing to ...	[trying to find a]good deal and good timing t...	81.0	17.0	0.790123	0.0
2	1	7.0	2.0	2.0	7509.0	trying to plan route penn ny to montgomery ny	trying to plan route penn ny to montgomery ny	trying to plan route penn ny to montgomery ny	63.0	16.0	0.746032	0.0

```
# Shape and structure
print("Shape:", df.shape)
print("Columns:", df.columns.tolist())
df.info()
df.describe()
```

```
Shape: (2663, 12)
Columns: ['Annotator A ID', 'Annotator B ID', 'Parition ID', 'Corpora ID', 'Sentence ID', 'Text', 'Annotator A Text', 'Annotator B Text'
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2663 entries, 0 to 2662
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Annotator A ID    2662 non-null   object
 1   Annotator B ID    2662 non-null   float64
 2   Parition ID       2658 non-null   float64
 3   Corpora ID        2658 non-null   float64
 4   Sentence ID       2658 non-null   float64
 5   Text              2658 non-null   object
 6   Annotator A Text  2658 non-null   object
 7   Annotator B Text  2658 non-null   object
 8   Length            2658 non-null   float64
 9   Error             2658 non-null   float64
 10  Alignment Score   2658 non-null   float64
 11  Agreement         2658 non-null   float64
dtypes: float64(8), object(4)
memory usage: 249.8+ KB
```

	Annotator B ID	Parition ID	Corpora ID	Sentence ID	Length	Error	Alignment Score	Agreement
count	2662.000000	2658.0	2658.0	2658.000000	2658.000000	2658.000000	2658.000000	2658.000000
mean	5.029943	2.0	2.0	8009.269752	97.922122	9.459744	0.923570	0.787434
std	2.435083	0.0	0.0	286.698423	39.827029	20.537986	0.149602	0.409200
min	0.076430	2.0	2.0	7506.000000	43.000000	0.000000	0.000000	0.000000
25%	3.000000	2.0	2.0	7758.000000	68.000000	0.000000	0.905325	1.000000
50%	5.000000	2.0	2.0	8012.500000	84.500000	0.000000	1.000000	1.000000
75%	7.000000	2.0	2.0	8255.000000	117.000000	9.750000	1.000000	1.000000
max	97.922122	2.0	2.0	8505.000000	200.000000	198.000000	1.000000	1.000000

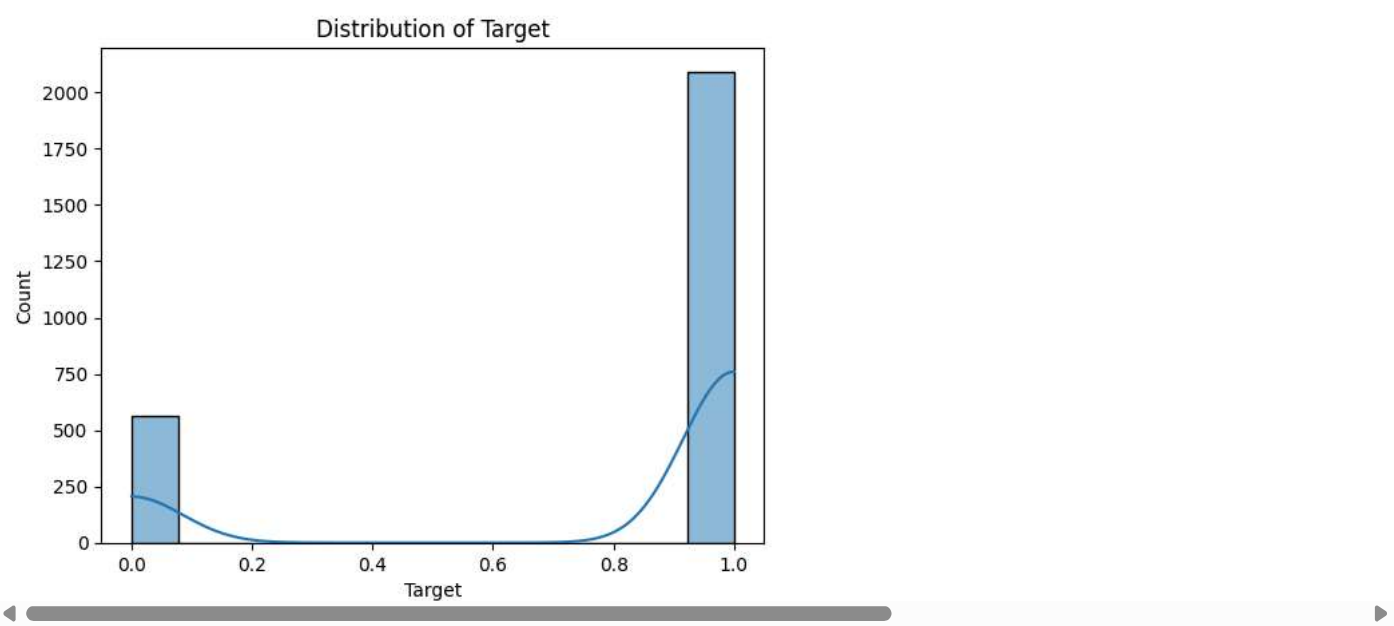
```
print("Missing Values:\n", df.isnull().sum())
print("Duplicate Rows:", df.duplicated().sum())
```

```
Missing Values:
 Annotator A ID      1
Annotator B ID      1
Parition ID         5
Corpora ID          5
Sentence ID         5
Text                5
Annotator A Text    5
Annotator B Text    5
Length              5
Error               5
Alignment Score     5
Agreement           5
dtype: int64
Duplicate Rows: 0
```

```
import seaborn as sns
import matplotlib.pyplot as plt

# Replace 'target_column' with the actual name of your target column
target_column = df.columns[-1]

sns.histplot(df[target_column], kde=True)
plt.title("Distribution of Target")
plt.xlabel("Target")
plt.show()
```




```
target = df.columns[-1]
features = df.columns.drop(target)
print("Features:", features)
```

```
Features: Index(['Annotator A ID', 'Annotator B ID', 'Parition ID', 'Corpora ID',
       'Sentence ID', 'Text', 'Annotator A Text', 'Annotator B Text', 'Length',
       'Error', 'Alignment Score'],
      dtype='object')
```

```
categorical_cols = df.select_dtypes(include=['object']).columns
print("Categorical Columns:", categorical_cols.tolist())

df_encoded = pd.get_dummies(df, drop_first=True)
```

```
Categorical Columns: ['Annotator A ID', 'Text', 'Annotator A Text', 'Annotator B Text']
```

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
X_scaled = scaler.fit_transform(df_encoded.drop(target, axis=1))
y = df_encoded[target]



from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)


from sklearn.preprocessing import StandardScaler

from sklearn.impute import SimpleImputer


target_column = df.columns[-1]


imputer = SimpleImputer(strategy='mean')
df_imputed = pd.DataFrame(imputer.fit_transform(df_encoded), columns=df_encoded.columns)


X = df_imputed.drop(target_column, axis=1)
y = df_imputed[target_column]
```

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)


from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score


X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)



print("MSE:", mean_squared_error(y_test, y_pred))
print("R² Score:", r2_score(y_test, y_pred))
```

```
MSE: 0.10665644669089891
R² Score: 0.34817086308129597
```

```
new_sample = df.drop(target, axis=1).iloc[0]
new_df = pd.DataFrame([new_sample])


df_temp = pd.concat([df.drop(target, axis=1), new_df], ignore_index=True)
df_temp_encoded = pd.get_dummies(df_temp, drop_first=True)
df_temp_encoded = df_temp_encoded.reindex(columns=df_encoded.drop(target, axis=1).columns, fill_value=0)


new_input_scaled = scaler.transform(df_temp_encoded.tail(1))
predicted_value = model.predict(new_input_scaled)

print("🎯 Predicted Value:", round(predicted_value[0], 2))
```

```
🎯 Predicted Value: 1.0
```

```
!pip install gradio

import gradio as gr

def predict_from_input(**kwargs):
    input_df = pd.DataFrame([kwargs])
    df_temp = pd.concat([df.drop(target, axis=1), input_df], ignore_index=True)
    df_temp_encoded = pd.get_dummies(df_temp, drop_first=True)
    df_temp_encoded = df_temp_encoded.reindex(columns=df_encoded.drop(target, axis=1).columns, fill_value=0)
    scaled_input = scaler.transform(df_temp_encoded.tail(1))
    prediction = model.predict(scaled_input)
    return round(prediction[0], 2)

# Auto-generate Gradio inputs
inputs = []
for col in df.drop(columns=target).columns:
    if df[col].dtype == 'object':
        inputs.append(gr.Dropdown(choices=df[col].dropna().unique().tolist(), label=col))
    else:
        inputs.append(gr.Number(label=col))

output = gr.Number(label=f"🎯 Predicted {target}")

gr.Interface(fn=predict_from_input, inputs=inputs, outputs=output, title="📈 Sarathi Prediction App").launch()
```

```
Collecting gradio
  Downloading gradio-5.29.1-py3-none-any.whl.metadata (16 kB)
Collecting aiofiles<25.0,>=22.0 (from gradio)
  Downloading aiofiles-24.1.0-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: anyio<5.0,>=3.0 in /usr/local/lib/python3.11/dist-packages (from gradio) (4.9.0)
Collecting fastapi<1.0,>=0.115.2 (from gradio)
  Downloading fastapi-0.115.12-py3-none-any.whl.metadata (27 kB)
Collecting ffmpy (from gradio)
  Downloading ffmpy-0.5.0-py3-none-any.whl.metadata (3.0 kB)
Collecting gradio-client==1.10.1 (from gradio)
  Downloading gradio_client-1.10.1-py3-none-any.whl.metadata (7.1 kB)
Collecting groovy~=0.1 (from gradio)
  Downloading groovy-0.1.2-py3-none-any.whl.metadata (6.1 kB)
Requirement already satisfied: httpx>=0.24.1 in /usr/local/lib/python3.11/dist-packages (from gradio) (0.28.1)
Requirement already satisfied: huggingface-hub>=0.28.1 in /usr/local/lib/python3.11/dist-packages (from gradio) (0.31.2)
Requirement already satisfied: jinja2<4.0 in /usr/local/lib/python3.11/dist-packages (from gradio) (3.1.6)
Requirement already satisfied: markupsafe<4.0,>=2.0 in /usr/local/lib/python3.11/dist-packages (from gradio) (3.0.2)
Requirement already satisfied: numpy<3.0,>=1.0 in /usr/local/lib/python3.11/dist-packages (from gradio) (2.0.2)
Requirement already satisfied: orjson~=3.0 in /usr/local/lib/python3.11/dist-packages (from gradio) (3.10.18)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-packages (from gradio) (24.2)
Requirement already satisfied: pandas<3.0,>=1.0 in /usr/local/lib/python3.11/dist-packages (from gradio) (2.2.2)
Requirement already satisfied: pillow<12.0,>=8.0 in /usr/local/lib/python3.11/dist-packages (from gradio) (11.2.1)
Requirement already satisfied: pydantic<2.12,>=2.0 in /usr/local/lib/python3.11/dist-packages (from gradio) (2.11.4)
Collecting pydub (from gradio)
  Downloading pydub-0.25.1-py2.py3-none-any.whl.metadata (1.4 kB)
Collecting python-multipart>=0.0.18 (from gradio)
  Downloading python_multipart-0.0.20-py3-none-any.whl.metadata (1.8 kB)
Requirement already satisfied: pyyaml<7.0,>=5.0 in /usr/local/lib/python3.11/dist-packages (from gradio) (6.0.2)
Collecting ruff>=0.9.3 (from gradio)
  Downloading ruff-0.11.10-py3-none-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (25 kB)
Collecting safehttpx<0.2.0,>=0.1.6 (from gradio)
  Downloading safehttpx-0.1.6-py3-none-any.whl.metadata (4.2 kB)
Collecting semantic-version~=2.0 (from gradio)
  Downloading semantic_version-2.10.0-py2.py3-none-any.whl.metadata (9.7 kB)
Collecting starlette<1.0,>=0.40.0 (from gradio)
  Downloading starlette-0.46.2-py3-none-any.whl.metadata (6.2 kB)
Collecting tomlkit<0.14.0,>=0.12.0 (from gradio)
  Downloading tomlkit-0.13.2-py3-none-any.whl.metadata (2.7 kB)
Requirement already satisfied: typer<1.0,>=0.12 in /usr/local/lib/python3.11/dist-packages (from gradio) (0.15.3)
Requirement already satisfied: typing-extensions~=4.0 in /usr/local/lib/python3.11/dist-packages (from gradio) (4.13.2)
Collecting uvicorn>=0.14.0 (from gradio)
  Downloading uvicorn-0.34.2-py3-none-any.whl.metadata (6.5 kB)
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages (from gradio-client==1.10.1->gradio) (2025.3.2)
Requirement already satisfied: websockets<16.0,>=10.0 in /usr/local/lib/python3.11/dist-packages (from gradio-client==1.10.1->gradio)
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.11/dist-packages (from anyio<5.0,>=3.0->gradio) (3.10)
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.11/dist-packages (from anyio<5.0,>=3.0->gradio) (1.3.1)
Requirement already satisfied: certifi in /usr/local/lib/python3.11/dist-packages (from httpx>=0.24.1->gradio) (2025.4.26)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.11/dist-packages (from httpx>=0.24.1->gradio) (1.0.9)
Requirement already satisfied: h11>=0.16 in /usr/local/lib/python3.11/dist-packages (from httpcore==1.*->httpx>=0.24.1->gradio) (0.16
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.28.1->gradio) (3.18.0)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.28.1->gradio) (2.32.3)
Requirement already satisfied: tqdm>=4.42.1 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.28.1->gradio) (4.67.1
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas<3.0,>=1.0->gradio) (2.9
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas<3.0,>=1.0->gradio) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas<3.0,>=1.0->gradio) (2025.2)
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic<2.12,>=2.0->gradio) (
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.11/dist-packages (from pydantic<2.12,>=2.0->gradio) (2
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from pydantic<2.12,>=2.0->gradio)
Requirement already satisfied: click>=8.0.0 in /usr/local/lib/python3.11/dist-packages (from typer<1.0,>=0.12->gradio) (8.2.0)
Requirement already satisfied: shellingham>=1.3.0 in /usr/local/lib/python3.11/dist-packages (from typer<1.0,>=0.12->gradio) (1.5.4)
Requirement already satisfied: rich>=10.11.0 in /usr/local/lib/python3.11/dist-packages (from typer<1.0,>=0.12->gradio) (13.9.4)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas<3.0,>=1.0->gr
Requirement already satisfied: markdown-it-py>=2.2.0 in /usr/local/lib/python3.11/dist-packages (from rich>=10.11.0->typer<1.0,>=0.12
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /usr/local/lib/python3.11/dist-packages (from rich>=10.11.0->typer<1.0,>=0.
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub>=0
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub>=0.28.1-
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.11/dist-packages (from markdown-it-py>=2.2.0->rich>=10.11.0->type
Downloading gradio-5.29.1-py3-none-any.whl (54.1 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 54.1/54.1 MB 15.9 MB/s eta 0:00:00
Downloading gradio_client-1.10.1-py3-none-any.whl (323 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 323.1/323.1 kB 17.8 MB/s eta 0:00:00
Downloading aiofiles-24.1.0-py3-none-any.whl (15 kB)
Downloading fastapi-0.115.12-py3-none-any.whl (95 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 95.2/95.2 kB 5.5 MB/s eta 0:00:00
Downloading groovy-0.1.2-py3-none-any.whl (14 kB)
Downloading python_multipart-0.0.20-py3-none-any.whl (24 kB)
Downloading ruff-0.11.10-py3-none-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (11.6 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 11.6/11.6 MB 79.0 MB/s eta 0:00:00
Downloading safehttpx-0.1.6-py3-none-any.whl (8.7 kB)
Downloading semantic_version-2.10.0-py2.py3-none-any.whl (15 kB)
Downloading starlette-0.46.2-py3-none-any.whl (72 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 72.0/72.0 kB 5.0 MB/s eta 0:00:00
Downloading tomlkit-0.13.2-py3-none-any.whl (37 kB)
Downloading uvicorn-0.34.2-py3-none-any.whl (62 kB)
```