**End to End ML Project Requirements**

Here’s an **expanded and exhaustive list** of tools for each stage of an end-to-end machine learning project, covering **mandatory tools** and as many **optional tools** as possible. This list applies to projects like **sentiment analysis** as well as other machine learning applications, ensuring complete estimation and flexibility for scaling.

**1. Data Collection**

| **Category** | **Tools** |
| --- | --- |
| **Web Scraping** | BeautifulSoup, Scrapy, Selenium, Puppeteer, Playwright, Octoparse |
| **APIs** | Tweepy (Twitter), Facebook Graph API, Reddit API, Google Places API, OpenAI API, RapidAPI |
| **Survey/Forms** | Google Forms, Typeform, JotForm |
| **Streaming Data** | Apache Kafka, RabbitMQ, Amazon Kinesis |
| **IoT Data** | MQTT, AWS IoT Core |
| **Custom Scripts** | Python’s requests, urllib, http.client for HTTP requests |

**2. Data Storage**

| **Category** | **Tools** |
| --- | --- |
| **Relational Databases** | MySQL, PostgreSQL, SQLite, Microsoft SQL Server, Oracle Database |
| **NoSQL Databases** | MongoDB, Cassandra, Firebase, DynamoDB, Couchbase |
| **Big Data Storage** | Apache HDFS, Amazon S3, Google Cloud Storage, Azure Blob Storage |
| **Data Warehousing** | Snowflake, Google BigQuery, AWS Redshift, Apache Hive, ClickHouse |
| **Time-Series Storage** | InfluxDB, TimescaleDB |

**3. Data Preprocessing**

| **Category** | **Tools** |
| --- | --- |
| **Data Cleaning** | Pandas, NumPy, OpenRefine, Pyjanitor |
| **Text Preprocessing** | NLTK, spaCy, TextBlob, gensim |
| **Data Profiling** | Pandas Profiling, Sweetviz, DataProfiler, YData-Profiling |
| **ETL Tools** | Talend, Informatica, Apache NiFi, Apache Spark, AWS Glue |
| **Data Validation** | Great Expectations, TensorFlow Data Validation (TFDV), Deequ |

**4. Feature Engineering**

| **Category** | **Tools** |
| --- | --- |
| **Feature Scaling** | Scikit-learn (MinMaxScaler, StandardScaler, RobustScaler), PyCaret |
| **Feature Selection** | Scikit-learn, Feature-engine, Boruta, SHAP, LIME |
| **Dimensionality Reduction** | PCA (Scikit-learn), t-SNE, UMAP, Autoencoder-based techniques |
| **Text Vectorization** | CountVectorizer, TfidfVectorizer, gensim, fastText, Hugging Face |
| **Image Features** | OpenCV, PyTorch torchvision, TensorFlow Hub, YOLO |

**5. Model Building**

| **Category** | **Tools** |
| --- | --- |
| **Traditional ML** | Scikit-learn, Statsmodels, MLlib (Apache Spark), XGBoost, CatBoost, LightGBM |
| **Deep Learning** | TensorFlow, Keras, PyTorch, MXNet, JAX, PaddlePaddle |
| **Reinforcement Learning** | Stable-Baselines3, Ray RLlib |
| **AutoML** | H2O.ai, Auto-sklearn, Google Cloud AutoML, PyCaret, DataRobot |
| **Custom ML Algorithms** | Python, R, Julia |

**6. Evaluation**

| **Category** | **Tools** |
| --- | --- |
| **Performance Metrics** | Scikit-learn (accuracy, precision, recall, F1-score), Keras metrics, TensorFlow Metrics |
| **Visualization** | Matplotlib, Seaborn, Plotly, Bokeh, Altair, ggplot2 (R) |
| **Explainability** | SHAP, LIME, ELI5, InterpretML |
| **Error Analysis** | Yellowbrick, Error Analysis Toolkit (EAT), Fairlearn |

**7. Model Deployment**

| **Category** | **Tools** |
| --- | --- |
| **API Frameworks** | Flask, FastAPI, Django |
| **Model Serving** | TensorFlow Serving, TorchServe, MLflow, AWS SageMaker Endpoint, Vertex AI |
| **Containerization** | Docker, Podman |
| **Orchestration** | Kubernetes, OpenShift |
| **Edge Deployment** | TensorFlow Lite, PyTorch Mobile, ONNX Runtime |

**8. User Interface**

| **Category** | **Tools** |
| --- | --- |
| **Web Frameworks** | Flask, Django, FastAPI |
| **Frontend Frameworks** | React.js, Angular.js, Vue.js |
| **Data Visualization** | Dash, Streamlit, Tableau, Power BI, Google Data Studio |

**9. Monitoring & Logging**

| **Category** | **Tools** |
| --- | --- |
| **Logging** | Python logging, Logstash, Fluentd |
| **Monitoring** | Prometheus, Grafana, New Relic |
| **ML Monitoring** | Evidently AI, WhyLabs |

**10. Workflow Automation**

| **Category** | **Tools** |
| --- | --- |
| **Pipeline Orchestration** | Apache Airflow, Prefect, Luigi |
| **DevOps** | Jenkins, GitHub Actions, CircleCI |

**Additional Tools by Use Case**

| **Use Case** | **Tools** |
| --- | --- |
| **Time-Series Analysis** | Prophet, statsmodels, Darts |
| **Natural Language Processing** | Hugging Face Transformers, spaCy, NLTK, gensim, AllenNLP |
| **Image Processing** | OpenCV, Pillow, TensorFlow, PyTorch torchvision |
| **Recommendation Systems** | Surprise, LightFM, TensorFlow Recommenders, Nvidia Merlin |
| **Big Data Processing** | Apache Spark, Hadoop, Flink |

**Key Points for Estimation**

* **Mandatory Tools**: Tools like Scikit-learn, Pandas, Flask, and TensorFlow/Keras are essential for almost every ML project.
* **Project-Specific Tools**: Depending on the complexity (e.g., real-time streaming, IoT data), additional tools like Apache Kafka or TensorFlow Lite might be necessary.
* **Scalability**: For large-scale deployment or monitoring, tools like Kubernetes, Prometheus, and Docker become crucial.