Machine Learning Data Scientist Interview Cheat Sheet

1. Data Handling

- Pandas: read_csv, df.info(), df.describe(), df.isnull().sum()
- Train/Test Split: train_test_split(X, y, test_size=0.2, stratify=y)
- Scaling: StandardScaler, MinMaxScaler, OneHotEncoder

2. Classical ML Pipelines (scikit-learn)

- Models: LogisticRegression, RandomForest, XGBClassifier
- Pipeline: Pipeline([('scale', StandardScaler()), ('clf', LogisticRegression())])
- Cross-validation: cross_val_score(model, X, y, cv=5)

3. Deep Learning (Keras/TensorFlow)

- Pretrained Models: ResNet50, VGG16, EfficientNet (keras.applications)
- Custom Model: Sequential -> Conv2D -> MaxPool -> Flatten -> Dense -> Softmax
- Compile & Train: model.compile(optimizer, loss, metrics); model.fit(X_train, y_train)
- Data Loading: ImageDataGenerator / tf.data

4. Model Evaluation

- Classification: accuracy, precision, recall, f1_score, roc_auc
- Regression: rmse, mae, r2_score
- Confusion Matrix & ROC Curve

5. Feature Engineering

- Missing Data: dropna, fillna, SimpleImputer
- Encoding: OneHotEncoder, LabelEncoder, get_dummies()
- Text: CountVectorizer, TfidfVectorizer
- Images: normalization (rescale 1./255), augmentation

6. Quick Tips

- Always split data (train/test/val)
- Check class imbalance → use stratify, SMOTE
- Start with baseline (Logistic, simple NN)
- Use callbacks (EarlyStopping, ModelCheckpoint) for DL
- Document metrics, plots, errors