# VIVA Quick Reference Cheat Sheet

## Project Summary (30-second pitch)

**Two classification problems:** 1. **Forest Cover Type**: 7-class prediction, 581K instances, 54 features → **MLP: 92.21%** 2. **Smoking Prediction**: Binary classification, 39K train, 23 features → **MLP: 75.30%**

**Models used**: Logistic Regression, SVM, MLP Neural Network

---

## Key Numbers to Remember

### Forest Cover Type

- **Dataset**: 581,012 instances, 54 features (10 continuous + 44 binary)
- **Classes**: 7 (imbalanced: Class 2 most common)
- **Best Model**: MLP (100, 100) → **92.21% accuracy**
- **Preprocessing**: StandardScaler on 10 continuous features only

### Smoking Prediction

- **Dataset**: 38,984 train, 16,708 test, 23 original features
- **Classes**: Binary (63% non-smoker, 37% smoker)
- **Best Model**: MLP (256, 128) → **75.30% accuracy**
- **Feature Engineering**: 31-56 features per model (extensive)

---

## Model Quick Facts

### Logistic Regression

- **How**: Sigmoid function → probability → classification
- **Pros**: Fast, interpretable, good baseline
- **Cons**: Linear decision boundary
- **Forest**: 72.34% | **Smoking**: 73.52%

### SVM

- **How**: Finds optimal hyperplane with maximum margin
- **Pros**: Handles non-linearity (kernels), high-dimensional
- **Cons**: Slow on large data, sensitive to hyperparameters
- **Forest**: 71.14% | **Smoking**: 60.54% (poor tuning)

### MLP Neural Network

- **How**: Multi-layer feedforward network with backpropagation
- **Pros**: Non-linear, learns complex patterns, best performance

- **Cons**: Black box, longer training, needs tuning
- **Forest**: 92.21% | **Smoking**: 75.30%

---

## Key Concepts (One-Liners)

| Concept | Explanation |
|---|---|
| **Overfitting** | Model learns training data too well, fails on new data |
| **Regularization** | Penalty to prevent overfitting (L1/L2) |
| **Cross-Validation** | K-fold: train on k-1 folds, test on 1, repeat k times |
| **Stratified Split** | Maintains class distribution in train/test |
| **StandardScaler** | (x - mean) / std → normalizes features |
| **Early Stopping** | Stop training when validation score stops improving |
| **Backpropagation** | Calculate gradients backward through network |
| **Feature Engineering** | Creating new features from existing ones (ratios, interactions) |

---

## Top 5 VIVA Answers

**Q: Why MLP best?**

**A**: Non-linear patterns, complex interactions (elevation×soil, medical ratios), deep architecture captures hierarchical features.

**Q: Why different feature sets?**

**A**: Simpler models (LR) need fewer features to avoid overfitting. Complex models (MLP) benefit from comprehensive features.

**Q: How handle class imbalance?**

**A**: Stratified train-test split maintains distribution. Could use class weights (future work).

**Q: Why scale only continuous features?**

**A**: Binary features already 0/1. Continuous features have different scales (elevation ~3000, slope ~30).

**Q: Key insights from EDA?**

**A**: - **Forest**: Elevation most important, hillshade patterns vary by type - **Smoking**: Liver enzymes (GTP, ALT, AST) strongest predictors, HDL lower in smokers

---

## Performance Summary

### Forest Cover Type

```
MLP:           92.21%
Logistic Reg:  72.34%
SVM:           71.14%
```

### Smoking Prediction

```
MLP:           75.30%
Logistic Reg:  73.52%
SVM:           60.54%
```

---

## Model Configurations

### Forest Cover - MLP

- Architecture: (100, 100)
- Activation: ReLU
- Solver: Adam
- Learning Rate: Adaptive (0.001)
- Early Stopping: Yes

### Smoking - MLP

- Architecture: (256, 128)
- Activation: Logistic (sigmoid)
- Solver: Adam
- Learning Rate: 0.0048 (adaptive)
- Early Stopping: Yes

---

## Feature Engineering Highlights

### Smoking Prediction

- **Medical Ratios**: Chol/HDL, AST/ALT, Trig/HDL (clinically meaningful)

- **Polynomial**: age × BMI, age × systolic (interactions)
- **Log Transforms**: log(GTP), log(ALT) (handle skewness)
- **Composite Scores**: Metabolic risk, CV risk (combine factors)
- **Averaging**: eyesight_avg, hearing_sum (bilateral measurements)

**Forest Cover**

- **No feature engineering**: Used raw features directly
- **Preprocessing**: Only scaling continuous features

---

## Common Pitfalls to Avoid

1. Don't say "I don't know" - say "I would investigate…"
2. Don't blame the data - explain what you did to handle challenges
3. Don't memorize - understand concepts and explain in your words
4. Don't ignore limitations - acknowledge and suggest improvements

---

## Strengths to Emphasize

1. Comprehensive EDA with visualizations
2. Multiple models compared systematically
3. Proper preprocessing (scaling, stratified splits)
4. Hyperparameter tuning (RandomizedSearchCV, 5-fold CV)
5. Domain knowledge applied (medical feature engineering)
6. Best practices (early stopping, regularization)

---

## Improvement Suggestions (If Asked)

1. **More hyperparameter tuning** (especially SVM)
2. **Ensemble methods** (combine multiple models)
3. **Feature selection** (remove redundant features)
4. **Class balancing** (SMOTE, class weights)
5. **Deeper networks** (try more layers)
6. **More data** (if available)

---

## Evaluation Metrics

- **Accuracy**: Overall correctness
- **Precision**: Of predicted positives, how many correct
- **Recall**: Of actual positives, how many found

- **F1-Score**: Harmonic mean of precision and recall
- **Confusion Matrix**: Shows TP, TN, FP, FN per class

---

## Final Checklist

Before VIVA, make sure you can explain: - [ ] Why each model was chosen - [ ] Why MLP performed best - [ ] How preprocessing was done - [ ] Feature engineering rationale - [ ] How class imbalance was handled - [ ] Key EDA insights - [ ] Model configurations - [ ] Limitations and improvements

---

**Remember**: Be confident, explain your choices, connect theory to practice!