Here is a structured summary and interview prep for your multi-notebook CATSnDOGS classification and clustering project including simulation study:

**Suggested Project Title:**

**"Comprehensive Evaluation of Classification and Clustering Algorithms on CATSnDOGS Image Dataset with Simulation Studies"**

**Project Summary**

**Part 1 – Supervised Learning with Multiple Classifiers**

* **Data:** 64x64 grayscale images of cats and dogs labeled accordingly.
* **Deep Learning Model:** CNN trained on raw images achieved up to 95% validation accuracy but showed some overfitting.
* **Classical ML Models:** SVM, kNN, Logistic Regression, Naive Bayes tested on flattened images with metrics and confusion matrices. SVM performed best with ~80-85% accuracy.
* **Misclassification Analysis:** Visualized misclassified images for CNN and SVM to gain insights on error types.
* **Feature Importance:** For linear models and logistic regression variants, heatmaps and bar plots showed pixel importance patterns.

**Part 2 – Unsupervised Learning / Clustering**

* **PCA Reduction:** Applied after feature scaling for visualization and dimensionality reduction.
* **K-Means Clustering:** Evaluated cluster numbers using Elbow method, silhouette score, NMI, and accuracy relative to labels. Accuracy was near random, silhouette scores low (~0.13 at best).
* **K-Medoids (PAM):** Similar evaluation as K-Means showed unstable and low accuracy (~0.6 max for 2 clusters) with poor performance as clusters increase.
* **Hierarchical Clustering:** Performed Ward linkage, dendrogram exploration, and clustering with varying cluster numbers. Accuracy and silhouette scores were low, confirming limited separability.
* **Visualizations:** Clusters and centroids plotted in PCA space with silhouette plots to assess cohesion and separation.

**Part 3 – Simulation Studies on Sample Size and Novel Approaches**

* **Algorithms Compared:** KNN, SVM, QDA, Logistic Regression, Gradient Boosting, Random Forest, Neural Networks.
* **Accuracy vs Sample Size:** Models trained with increasing sample sizes (10-190) in repeated runs.
  + SVM and KNN performed best and improved steadily with more data.
  + QDA struggled due to unrealistic assumptions.
  + Neural networks showed inconsistent accuracy, indicating data/architecture challenges.
* **Patch-wise Classification:** Images split into 16 non-overlapping 16x16 patches; models run independently per patch.
  + Accuracy maps showed consistent importance of upper head patches across models; lower patches less predictive.
  + SVM was top-performing; QDA and Logistic Regression performed worst.
* **Decision Boundary Visualization:** PCA-reduced data visualized with decision boundaries for SVM, QDA, Logistic Regression demonstrating their separations.

**Part 4 – Imbalanced Data Simulation**

* **Class Proportion Variation:** Increased dog class proportion (from ~50% to majority) by removing cat samples.
* **Error Metric:** Proportion of cats misclassified as dogs monitored.
* **Findings:** As imbalance increased, all models tended to overclassify as dogs, confirming known bias issues in imbalanced datasets.
* **Implications:** Highlights the brittleness of classifiers when faced with class imbalance in real applications.

**Interview Preparation: Key Concepts**

**Supervised Learning**

* **Why use CNN for image data, and why compare with classical ML?**  
  CNNs exploit spatial structure directly, usually outperforming classical models on image tasks. Classical ML provides interpretability and baseline comparison.
* **Why flatten images for SVM, kNN, Logistic Regression?**  
  These models require fixed-size vector input; flattening preserves pixel information in tabular form.
* **What causes overfitting in CNN and how to observe it?**  
  High train accuracy but declining or variable validation accuracy; possible due to model complexity or limited data.
* **Why visualize misclassified images?**  
  To diagnose error patterns and understand if mistakes are random or systematic.

**Clustering**

* **Why use PCA before clustering?**  
  Reduce dimensionality and noise; help visualization and improve clustering stability.
* **What do silhouette scores and NMI tell us in clustering?**  
  Silhouette measures cluster cohesion/separation; NMI compares cluster assignments to true labels, assessing clustering accuracy.
* **Why do K-Means and PAM struggle here?**  
  High-dimensional, noisy image data often lacks compact cluster structure in original feature space; labels may not correspond well to natural clusters.
* **What is dendrogram's role in hierarchical clustering?**  
  Visualizes cluster merge distances at each step, helping identify optimal cluster numbers.

**Simulation Studies**

* **Why test accuracy as a function of sample size?**  
  To evaluate how models learn as more data become available, estimating data efficiency and performance trends.
* **Why patch-wise classification?**  
  To explore spatial informativeness of image regions and reduce input dimensionality for models.
* **Why simulate class imbalance?**  
  Real-world data often have imbalanced classes; assessing robustness to imbalance is critical.
* **Which models handle imbalance better and why?**  
  Models with regularization or ensemble methods often resist bias, but most tend to overpredict majority class without adjustment.

**General ML Concepts**

* **Why cross-validate and use grid search?**  
  To optimize hyperparameters and fairly estimate performance, reducing overfitting risk.
* **What’s the impact of kernel choice in SVM?**  
  Determines the decision boundary complexity; RBF kernel captures non-linear boundaries better than linear.
* **How does PCA affect clustering and classification?**  
  Can improve performance by removing noisy dimensions and focusing on key variance directions, but may also remove subtle discriminative features.

If you want, I can format this into a polished Markdown report with sections, tables, and concise Q&A for interview prep!