Based on the detailed description and analysis you shared about the MNIST dataset project, here is a comprehensive and structured summary report aligned with your work:

**Project Title:**

**Advanced Analysis and Classification of Modified MNIST Digit Dataset Using Dimensionality Reduction and Clustering**

**Summary Report**

**Dataset Overview**

* Utilized a large-scale MNIST variant with 50,000 images of handwritten digits (0-9).
* Each image is represented as a 28x28 pixel grid, flattened into input features.
* The dataset is balanced across digits (~10% per class), enabling use of accuracy as a primary metric.
* Due to large sample size and high dimensionality, efficient subsampling and dimensionality reduction are critical.

**a) Dimensionality Reduction and Data Exploration**

1. **Subsampling Techniques:**
   * Employed *random subsampling* (5,000 samples) and *diversified sampling* (using fast diversified sampling algorithm) to capture representative data subsets.
   * Diversified sampling better preserves feature distribution spread, facilitating more reliable downstream analysis.
2. **Dimensionality Reduction Methods:**
   * Explored linear methods: PCA, Truncated SVD, Sparse PCA, and NMF.
   * Explored nonlinear methods: Kernel PCA (KPCA, with RBF kernel), t-SNE (with various perplexities), and UMAP.
   * Assessed through:
     + Visualization of components (2D and 3D scatter plots).
     + Scree plots depicting explained variance (or equivalent for nonlinear methods).
3. **Findings:**
   * PCA, Truncated SVD, and KPCA showed a similar decay in explained variance, motivating choosing ~60 components for modeling.
   * KPCA outperformed PCA and SVD in explained variance and cluster separation; Sparse PCA and NMF were less informative.
   * t-SNE and UMAP yielded superior visualizations with clear clustering of some classes; UMAP slightly better at separating overlapping classes (e.g., digits 6 and 9).
   * Some classes (digits 2, 3, 8) exhibited diffuse clustering, indicating intrinsic difficulty and possible label noise.

**b) Classification Performance vs Training Size**

1. **Sampling Strategy:**
   * Used *m out of n bootstrap* sampling (10 bootstrap sets of 1,000 samples each) for feasible and diverse model training.
2. **Dimensionality Reduction for Classification:**
   * Tested KPCA, Truncated SVD, and UMAP embedded in classification pipelines.
   * KPCA (60 components) achieved best validation accuracy (~0.83), followed by SVD (~0.81) and UMAP (~0.73).
   * Selected KPCA for final classification models.
3. **Classifiers:**
   * Trained six classifiers: KNN, QDA, Support Vector Machines (linear and kernel), Logistic Regression (kernelized), Random Forest (RF), and Neural Networks (NN).
   * Hyperparameters tuned using cross-validation and grid search; mostly stable except RF depth showing some variability.
4. **Results:**
   * Training set accuracy remained steady for all but KNN (which increased with training size).
   * Validation accuracy declined with smaller training size across models.
   * QDA degraded quickly, likely due to insufficient data to estimate Gaussian class distributions reliably.
   * RF, SVM, and NN remained more robust with moderate drops.

**c) Clustering Analysis and Class Recovery**

1. **Methods Explored:**
   * *K-means clustering* on KPCA embeddings with varying k.
   * *DBSCAN* with extensive parameter tuning—found clustering poor due to density heterogeneity.
   * *Agglomerative hierarchical clustering* with multiple linkage criteria (single, complete, average, Ward).
2. **Findings:**
   * K-means showed no sharp elbow in WCSS plot but silhouette scores suggested ~8 clusters optimal.
   * After label alignment, K-means clusters corresponded well with some digits (e.g., 0, 1), but merged several others (e.g., 4 & 7, 5 & 8, 6 & 9).
   * DBSCAN largely failed to detect meaningful clusters for this dataset, often producing single large clusters or excessive noise points.
   * Agglomerative clustering with Ward linkage performed better than other linkages; identified 8 clusters at suitable distance threshold.
   * However, its adjusted clustering accuracy was lower than K-means, indicating sensitivity to feature space representation.

**Conclusions and Insights**

* Diversified sampling provides reliable representative subsets for exploring vast datasets.
* KPCA combines nonlinear mapping and variance preservation, yielding effective embeddings for classification.
* UMAP excels at visually uncovering complex cluster structure but may underperform slightly in direct classification tasks.
* Classifiers leveraging nonlinear embeddings (KPCA + SVM, RF, NN) offer best trade-off between accuracy and stability.
* Clustering partially recovers known class structure; however, overlapping digits and label noise limit perfect recovery.
* Future improvements could explore ensemble methods, refined embeddings, and active sample cleaning to handle mislabeled or ambiguous samples.

**Key Interview Questions and Suggested Answers**

1. **Q:** How did you handle the challenges posed by the large sample size and high dimensionality of MNIST?  
   **A:** I applied sampling techniques like diversified and bootstrap sampling to reduce computational load while maintaining data representativeness, combined with dimensionality reduction (KPCA, UMAP) to reduce feature space complexity before modeling.
2. **Q:** Why choose KPCA over simpler methods like PCA for dimensionality reduction?  
   **A:** KPCA captures nonlinear relationships better than PCA, producing embeddings that improved classification accuracy and cluster separation for complex image data with intricate variances not represented linearly.
3. **Q:** Which classifiers worked best with this dataset, and why?  
   **A:** KPCA embeddings fed to SVM, Random Forest, and Neural Networks performed best, likely due to their ability to handle complex class boundaries and mitigate overfitting in high-dimensional spaces.
4. **Q:** How effective was clustering in recovering digit classes?  
   **A:** K-means recovered major clusters well but merged several digit groups; hierarchical clustering with Ward linkage had fewer clusters but more ambiguity. Density-based clustering was ineffective, suggesting cluster mixture and noise complexity.
5. **Q:** What are potential limitations experienced?  
   **A:** Label noise and intrinsic similarity (e.g., between digits 6 & 9) complicated classification and clustering. Computational constraints limited exhaustive hyperparameter tuning and data size per training iteration.

If you'd like, I could now help format this into presentation slides, a detailed report, or a concise executive summary. Let me know!

Add to follow-up

Check sources

1. <https://www.kaggle.com/code/pranaysingh25/dimensionality-reduction-visualization-mnist>
2. <https://colah.github.io/posts/2014-10-Visualizing-MNIST/>
3. <https://m-mburu.github.io/datacamp/introduction-to-advanced-dimensionality-reduction/introduction_dimensionality_reduction.html>
4. <https://towardsdatascience.com/6-dimensionality-reduction-techniques-how-and-when-to-use-them-e4891c10b5db/>
5. <https://sites.duke.edu/dimensionreduction/>
6. <https://www.pinecone.io/learn/dimensionality-reduction/>
7. <https://rpubs.com/marikan/pca_mnist>
8. <https://www.kaggle.com/code/amiribrahimtaj/mnist-9-dimensionality-reduction-techniques-knn>