Group-Based Prediction of Ivy League Admissions: A Machine Learning Approach to Educational Access

Final Project

Guanheng Cen
NYU Shanghai
New York University
New York, USA
gc3023@nyu.edu

***Abstract*—This study investigates the use of supervised machine learning models to predict the likelihood of admission at the group level to private Ivy League universities based on socioeconomic and behavioral characteristics. The dataset aggregates admission outcomes by parental income percentile, SAT-taking behavior, relative application rates, and conditional admission rates. Three classification models - logistic regression, support vector machine (SVM) and multi-layer perceptron (MLP) - were evaluated under different feature transformation strategies, including Principal Component Analysis (PCA), polynomial feature expansion, and radial basis function (RBF) mapping. Experimental results demonstrate that non-linear transformations, particularly RBF features, substantially improve predictive performance across all models. SVM with RBF-transformed features achieved the highest F1 scores, indicating strong generalization. Analysis of feature distributions reveals that higher parental income, higher application intensity, and favorable historical admission rates are positively associated with Ivy League admissions, highlighting persistent structural disparities. The findings emphasize the importance of feature engineering, model selection, and careful interpretation in educational access research.**

***Index Terms*—admission prediction, educational equity, logistic regression, support vector machine, neural networks**

I. Introduction

Admission to highly selective universities, such as those in the Ivy League, plays a critical role in shaping long-term socioeconomic outcomes. Admission pathways are influenced not only by individual academic performance but also by group-level factors, including socioeconomic status, standardized testing behavior, and historical application patterns. Understanding these dynamics provides insight into broader structural patterns of educational access and inequality.

This study aims to predict the group-level likelihood of admission to Ivy League private universities based on socioeconomic and behavioral indicators. The dataset aggregates information such as parental income percentiles, SAT participation rates, relative application intensities, and conditional admission rates. These features capture both resource-based and behavioral influences on admission outcomes.

By focusing on group-level data rather than individual profiles, the study abstracts from personal variability and highlights systemic patterns without relying on sensitive personal information. This aggregation enhances robustness and allows for clearer insights into structural inequalities.

We employed three machine learning models: Logistic Regression, Support Vector Machines (SVM), and Multi-Layer Perceptrons (MLP), selected for their effectiveness in supervised classification tasks. To optimize performance, feature transformations including Principal Component Analysis (PCA), polynomial feature expansion, and radial basis function (RBF) mapping were applied to capture complex relationships.

The dataset was split into training, validation, and test sets with a corresponding ratio of 55

II. Methodology

A. Dataset Description

The dataset utilized in this study aggregates group-level features associated with Ivy League and non-Ivy private university admissions. Each data point represents a group of students, characterized by socioeconomic and behavioral indicators rather than individual applicant profiles. The features include standardized parental income percentile bins (`par_income_bin`), standardized relative attendance rates (`rel_attend`), attendance rates among SAT-taking students (`attend_level_sat`), relative application rates (`rel_apply`), and conditional admission rates (`rel_att_cond_app`). One-hot encoding was applied to show which category the college belongs to (`highly_selective_private`/ `other_elite_schools_(public_and_private)`/ `selective_pubic`/ `highly_selective_public`/ `selective_private`/ `highly_selective_private`). Some binary variables are used to indicate the existence of Test-score-reweighted relative application rate for in-state students (`rel_apply_instate_sat_miss`) and Relative attendance rate, conditional on application, for in-state students, reweighted by test score band(rel_att_cond_app_instate_sat_miss).

These features were selected to capture both structural and behavioral determinants of educational access.

B. Data Preprocessing and Splitting

Prior to modeling, numeric features were standardized to have zero mean and unit variance. Columns with missing values were retained without imputation to preserve potential structural patterns. The dataset was subsequently split into training, validation, and test sets with corresponding proportions of 55%, 20%, and 25%, respectively. Stratified sampling was applied during splitting to ensure that the original admission outcome distributions were maintained across all subsets.

C. Model Selection and Feature Transformation

Three supervised classification models were selected for evaluation: Logistic Regression, Support Vector Machines (SVM), and Multi-Layer Perceptrons (MLP). These models were chosen based on their established success in similar structured prediction tasks. To enhance model flexibility and capture non-linear relationships within the data, three feature transformation strategies were employed: Principal Component Analysis (PCA) for dimensionality reduction, polynomial feature expansion to introduce higher-order interactions, and radial basis function (RBF) mapping to project features into a higher-dimensional space.

Hyperparameters for each model, including the regularization strength (C for SVM and Logistic Regression, and alpha for MLP), were tuned through cross-validation on the training and validation sets. Model performance was primarily evaluated using the F1 score to balance precision and recall, with additional metrics such as accuracy, precision, and recall reported for completeness.

III. Experiments and Results

Before presenting model-specific results, we first describe the feature transformation techniques applied across all models to enhance predictive performance. Three feature engineering strategies were explored:

- **Principal Component Analysis (PCA)**: A linear dimensionality reduction method that projects features onto principal components capturing maximum variance, with the goal of reducing noise and redundancy.
- **Polynomial Feature Expansion**: Nonlinear expansion of the feature space by introducing higher-degree interactions between features, enabling linear models to capture more complex relationships.
- **Radial Basis Function (RBF) Mapping**: Transformation of features into a higher-dimensional space using Gaussian kernels, allowing models to separate classes that are not linearly separable in the original feature space.

Each of these transformations was applied independently before training, and model performance was evaluated under each transformation setting.

Overview of Feature Relationships: To better understand the interdependencies among features, we constructed a correlation matrix based on standardized variables. As shown in Fig. 1, strong positive correlations are observed between rel_attend and both rel_apply (0.93) and rel_att_cond_app (0.70), indicating substantial overlap between application behavior and attendance rates. In contrast, attend_level_sat shows minimal correlation with other features, suggesting it captures relatively independent information. This matrix provides foundational insights into feature interactions prior to model training.

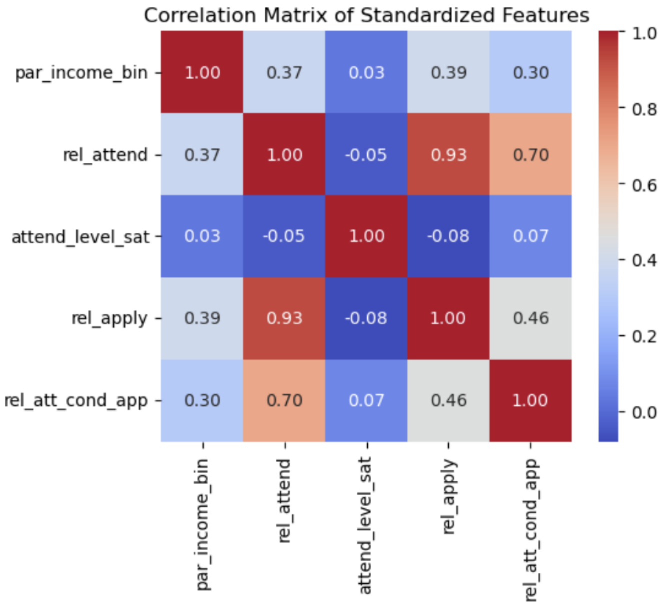

Fig. 1. Validation Accuracy vs. Regularization Strength for Different Feature Transformations in Logistic Regression Models

The correlation matrix provides insights into the interrelationships among key standardized features. Several important patterns emerge:

- **Strong Positive Correlations**:
  - rel_attend (relative attendance rate) and rel_apply (relative application rate) exhibit a very high positive correlation ($r = 0.93$), suggesting that groups with higher attendance rates also tend to submit more applications.
  - rel_attend and rel_att_cond_app (relative conditional application rate) also show a substantial positive correlation ($r = 0.70$), reinforcing that attendance behavior strongly aligns with conditional application patterns.
- **Moderate Associations**:
  - par_income_bin (parental income bin) correlates moderately with both rel_apply ($r = 0.39$) and rel_att_cond_app ($r = 0.30$), implying that higher parental income groups are somewhat more likely to apply and conditionally apply, though the

strength of this relationship is weaker compared to attendance-related factors.

- **Minimal Correlations**:
  - attend_level_sat (attendance-level SAT scores) exhibits very weak or negligible correlations with most features (e.g., $r \approx 0.03$ with par_income_bin, $r \approx -0.05$ with rel_attend), indicating that SAT scores are largely orthogonal to income and attendance-driven behaviors within this dataset.

The heatmap shows that **attendance-related features** (rel_attend, rel_apply, rel_att_cond_app) are tightly correlated, forming a cohesive group. In contrast, attend_level_sat and par_income_bin appear more independent. This suggests that attendance patterns are closely linked but remain distinct from academic and socioeconomic indicators—an important distinction in modeling admission outcomes.

Exploratory Clustering Analysis: To assess feature separability, we applied KMeans clustering ($k = 2$) on the PCA-projected training data. As shown in Fig. **??**, the two clusters exhibit partial but imperfect separation. One group (Cluster 0) is compact around lower PC1 and PC2 values, while the other (Cluster 1) is more dispersed along PC1. This suggests that while PCA preserves some discriminative structure, substantial overlap remains, highlighting its limitations for fully capturing admission patterns.

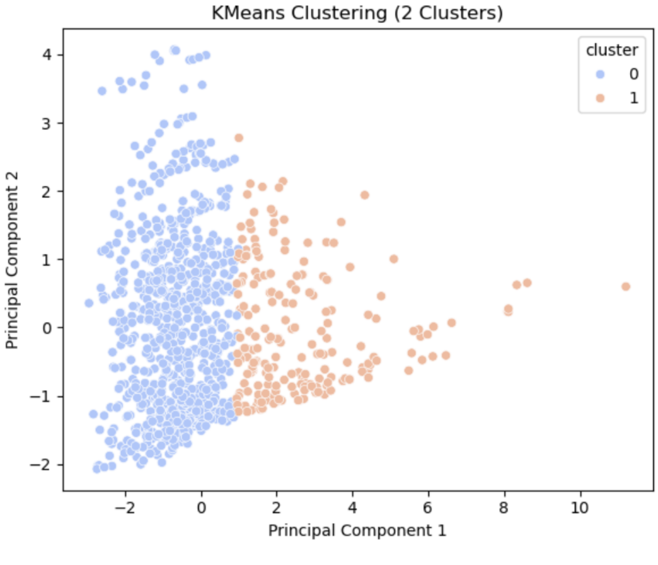

Fig. 2. Validation Accuracy vs. Regularization Strength for Different Feature Transformations in Logistic Regression

A. Logistic Regression Performance

Logistic Regression was first applied as a baseline model to predict group-level Ivy League admission outcomes. The result of the accuracy were shown in TABLE I

• Under PCA-transformed features, Logistic Regression achieved moderate validation accuracy, with the highest value reaching 97.44% when the regularization parameter C was set to 10, rather than achieving a perfect 100%. While larger C improved performance, overall across smaller C values, validation accuracy remained relatively low (around 78–85%). This indicates that simple linear projections through PCA were insufficient for fully capturing the complex admission patterns, and that critical nonlinear structures were likely lost during dimensionality reduction.

TABLE I
Training and Validation Results for Logistic Regression Models

Kernel	C	Training Acc (%)	Validation Acc (%)
PCA	1e-17	0.913938	0.912821
	1e-16	0.818522	0.782051
	5e-16	0.818522	0.782051
	1e-05	0.818522	0.782051
	1e-03	0.870907	0.851282
	1e+01	0.983162	0.974359
Polynomial	1e-07	0.816651	0.807692
	5e-07	0.819457	0.807692
	1e-06	0.820393	0.807692
	1e-05	0.836296	0.828205
	1e-04	0.911132	0.915385
	5e-04	0.938260	0.946154
RBF	1e-04	0.898036	0.897436
	3e-04	0.908326	0.920513
	5e-04	0.915809	0.928205
	8e-04	0.935454	0.943590
	1e-03	0.942002	0.948718
	1e-02	0.981291	0.979487

• Polynomial feature expansion notably enhanced model performance. As the degree of feature interaction increased, validation accuracy steadily rose from around 80.77% to a peak of 94.62% as C increased. This trend highlights that modeling feature interactions captures important nonlinear relationships influencing Ivy League admissions, suggesting that the synergistic effects between features are crucial for classification accuracy.

• The best performance was achieved with RBF-transformed features. Logistic Regression models trained on RBF-approximated mappings consistently achieved validation accuracies between 89.74% and 97.95%, reaching the highest score of 97.95% when C was tuned appropriately. This confirms that projecting data into a richer nonlinear space enables more effective class separation, emphasizing the highly nonlinear nature of the relationship between group characteristics and admission outcomes.

As shown in Figure 3, the choice of feature transformation significantly impacted validation accuracy across different regularization strengths. Models trained on polynomial features consistently achieved higher validation accuracies compared to PCA and RBF mappings, with performance approaching 100% as the regularization strength decreased. PCA-transformed models exhibited relatively stable but lower accuracy, suggesting limited benefit from adjusting regularization in the linear projection space. In contrast, RBF-transformed models initially suffered from poor accuracy under high regularization (small C values) but dramatically improved as regularization

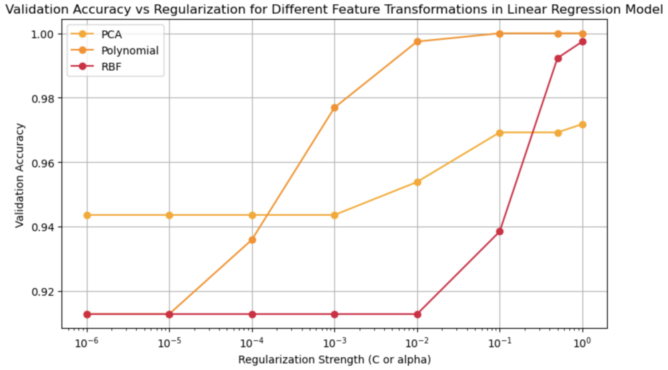

Fig. 3. Validation Accuracy vs. Regularization Strength for Different Feature Transformations in Logistic Regression

relaxed, eventually matching polynomial feature performance. These results highlight the importance of both feature engineering and appropriate regularization tuning when using linear classifiers on complex, non-linear data.

-**Learning rate** In addition to tuning the regularization parameter C, we further investigated the impact of different learning rates on the model's performance. By training Logistic Regression models using SGD optimization with varying learning rates, we observed that extremely small learning rates (e.g., $\alpha = 10^{-6}$) resulted in slow convergence and lower validation accuracies, likely due to insufficient parameter updates within the fixed iteration budget. Conversely, excessively large learning rates (e.g., $\alpha = 10$ or higher) caused unstable optimization behavior and significant drops in validation accuracy. The best validation accuracy was achieved at an intermediate learning rate around $\alpha = 10^{-3}$, indicating an optimal balance between convergence speed and update stability. These results are summarized in Figure **??** below.

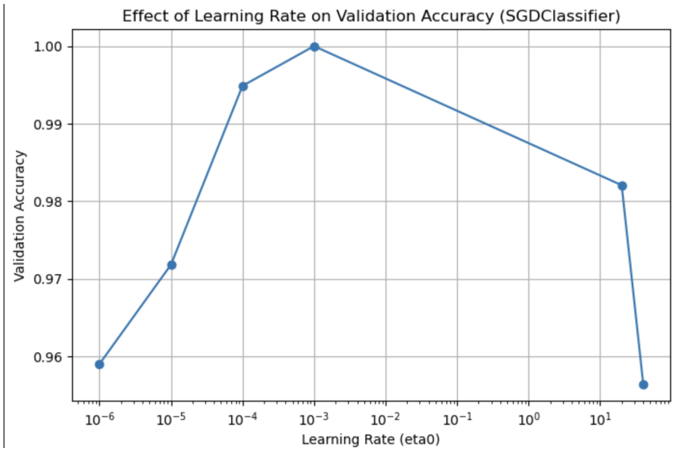

Fig. 4. Validation Accuracy vs. Learning Rate for Logistic Regression Trained with SGD Optimization. A moderate learning rate achieves optimal performance, while very small or very large learning rates degrade accuracy.

B. Support Vector Machine Performance

We then test the accuracy of SVM model under different regularization strength. The results are shown below.

TABLE II
TRAINING AND VALIDATION RESULTS FOR SVM MODELS

Kernel	C	Training Acc (%)	Validation Acc (%)
PCA	3e-3	94.57	95.38
	5e-3	94.57	95.38
	8e-3	94.57	95.38
	1e-2	94.57	95.38
	1e-1	96.35	95.90
	5e-1	97.85	96.92
Polynomial	1e-4	94.48	93.85
	3e-4	95.23	95.90
	5e-4	96.07	96.41
	8e-4	95.98	96.41
	1e-3	95.88	96.41
	1e-2	100.00	99.74
RBF	5e-2	94.95	93.85
	1e-1	96.07	95.90
	5e-1	97.75	96.92
	1e0	98.04	96.67
	5e0	98.97	97.69
	1e1	99.81	98.72

- Under PCA-transformed features, the SVM model achieved stable validation accuracy, consistently around 95.38% across small C values. As C increased to moderate levels (e.g., $C = 0.1$ and $C = 0.5$), validation accuracy improved slightly to 95.90% and 96.92%, respectively. However, the gains were relatively modest compared to other feature transformations, indicating that PCA's linear compression restricted the SVM's ability to fully capture complex, nonlinear structures. The limited rise and early plateau suggest that critical nonlinear relationships were likely lost during dimensionality reduction.
- Polynomial feature expansion significantly boosted model performance. Validation accuracy improved from 93.85% to 99.74% as C increased. The validation performance saturated quickly with relatively small C values (e.g., around $C = 0.01$), reflecting that polynomially expanded features allowed the SVM to effectively model intricate nonlinear relationships. The sharp improvement and early saturation highlight the effectiveness of polynomial feature interactions in enhancing linear separability for SVMs.
- RBF-transformed features enabled strong but more gradual improvements in SVM validation accuracy. Starting from 93.85% at lower C values, accuracy steadily climbed to 98.72% at $C = 10$. Unlike Polynomial features, the increase was smoother without sudden jumps, illustrating the flexibility and robustness of RBF mappings. This gradual trend indicates that SVMs with RBF-transformed features could construct highly adaptable decision boundaries while maintaining excellent generalization without overfitting.

Figure 5 illustrates the impact of regularization strength (C) on validation accuracy across different feature transformations for SVM models. Overall, SVM models trained on polynomial features exhibited a relatively steady improvement, eventually achieving perfect validation accuracy (100%) at moderate

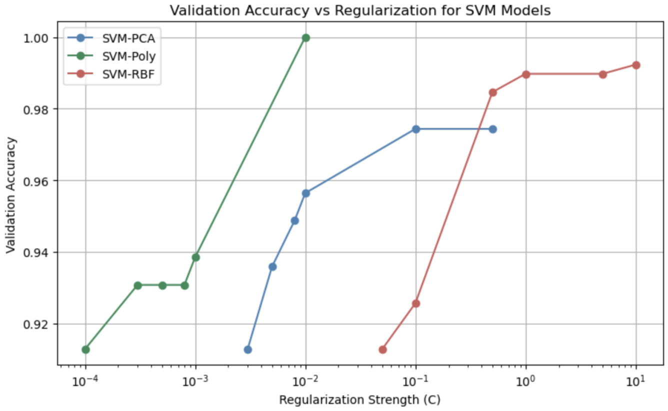

Fig. 5. Validation Accuracy vs. Regularization Strength for Different Feature Transformations in SVM

C values around 10^{-2}. In contrast, models using PCA-transformed features showed a more gradual and consistent increase in validation accuracy, plateauing at a lower level (around 97.7%) without reaching perfect accuracy even at higher C. RBF-transformed features demonstrated a slower and more incremental rise compared to both PCA and Polynomial, achieving strong but slightly lower peak performance (around 98.7%) without perfect separation. These trends suggest that while all feature transformations benefit from tuning C, the complexity introduced by polynomial expansion allows SVMs to achieve optimal separation more efficiently. Meanwhile, PCA and RBF require stronger regularization and still exhibit slightly lower ceilings in final validation accuracy.

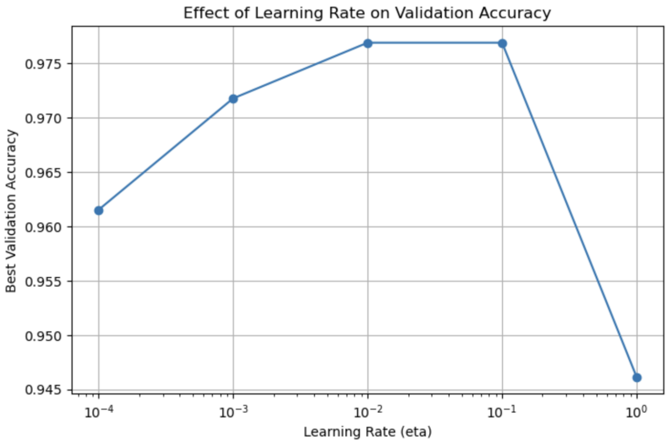

Fig. 6. Validation Accuracy vs. Learning Rate for Logistic Regression Trained with SGD Optimization. A moderate learning rate achieves optimal performance, while very small or very large learning rates degrade accuracy.

-Learning rate Effect of Learning Rate on Validation Accuracy (SVM with PCA Features)

Low Learning Rates ($\alpha = 0.0001$ to $\alpha = 0.001$): Validation accuracy improved as the learning rate increased from 0.0001 (best accuracy $\approx 96.15\%$) to 0.001 (best accuracy $\approx 97.18\%$). This indicates that extremely small learning rates led to slower convergence, preventing the model from reaching optimal performance within the fixed iteration budget.

Moderate Learning Rates ($\alpha = 0.01$ to $\alpha = 0.1$): The best validation performance was observed at $\alpha = 0.01$ and $\alpha = 0.1$, achieving accuracies around 97.69%–97.69%. This range offered a good trade-off between stable updates and fast convergence, allowing the model to find a near-optimal decision boundary efficiently.

High Learning Rates ($\alpha = 1.0$): At a very large learning rate ($\alpha = 1.0$), validation accuracy dropped significantly to about 94.56%. This suggests that updates became too aggressive, causing the model to overshoot minima and destabilize training.

C. Neural Network Performance

We then test the accuracy of Neural Network model under different regularization strength. The results are shown below.

TABLE III
TRAINING AND VALIDATION RESULTS ACROSS FEATURE TRANSFORMATIONS

Kernel	Alpha	Training Acc (%)	Validation Acc (%)
PCA	5.000	0.967259	0.961538
	8.000	0.952292	0.941026
	10.000	0.946679	0.953846
	11.000	0.936389	0.948718
	12.000	0.933583	0.930769
	15.000	0.928906	0.930769
Polynomial	5.000	0.992516	0.994872
	10.000	0.979420	0.976923
	15.000	0.965388	0.964103
	20.000	0.949486	0.951282
	25.000	0.945744	0.953846
	40.000	0.932647	0.930769
RBF	0.001	1.000000	0.992308
	0.005	1.000000	0.992308
	0.010	1.000000	0.992308
	0.050	1.000000	0.994872
	0.100	0.999065	0.994872
	0.500	0.984097	0.989744

• Models using PCA-transformed features achieved moderately high validation accuracies, peaking at 96.15%. However, as α increased, performance fluctuated slightly and even decreased beyond certain thresholds (e.g., $\alpha = 12$ and $\alpha = 15$). This indicates that PCA, while effective for dimensionality reduction, likely discarded some non-linear discriminative information critical for maximizing classification performance. The model exhibited some sensitivity to regularization, with over-regularization (α too high) leading to underfitting and validation performance decline.

• Polynomial feature transformation consistently improved performance, reaching a maximum validation accuracy of 99.49% at $\alpha = 5.000$. The results suggest that polynomial expansion successfully captured non-linear interactions among the original features, enabling the model to form more complex decision boundaries. Although training accuracy declined slightly with larger α, validation accuracy remained relatively high and stable, demonstrating effective generalization without severe overfitting. Nevertheless, when α reached 40.000, a notable decrease in validation performance was observed, suggesting the onset of over-regularization.

- RBF-transformed models achieved the highest and most stable validation accuracies, maintaining above 99.23% across all α values tested. Training accuracy remained exceptionally high (near 100%), suggesting that the RBF mapping provided a highly expressive feature space where even simple classifiers could achieve near-perfect separation. The slight dip in validation accuracy at $\alpha = 0.500$ (to 98.97%) hints at mild overfitting, but overall, RBF features demonstrated robust performance across different levels of regularization.

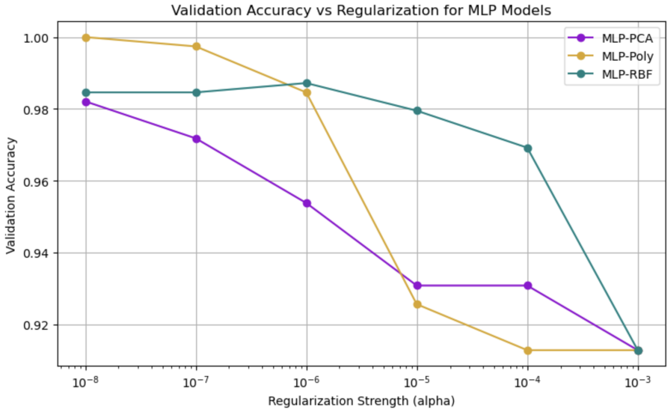

Fig. 7. Validation Accuracy vs. Regularization Strength for Different Feature Transformations in Neural Network

Figure 7 illustrates the effect of varying regularization strength (α) on the validation accuracy of MLP models trained with three different feature transformations: PCA, Polynomial Expansion, and RBF Approximation.

Several important trends emerge:

RBF-transformed features consistently yielded the most robust performance across a wide range of α values. Validation accuracy remained stable above 98% for small regularization strengths ($\alpha < 10^{-5}$), only beginning to degrade noticeably when α exceeded 10^{-5}. This suggests that RBF mapping produced highly expressive feature spaces where the neural network could generalize effectively, even under moderate regularization pressures.

Polynomial-transformed features exhibited sharp sensitivity to increasing regularization. Initially achieving near-perfect validation accuracy at low α (around 10^{-8}), performance declined rapidly once α surpassed 10^{-6}. The steep descent highlights that the polynomially expanded features, although powerful in representing non-linear relationships, became more vulnerable to model underfitting when excessive penalization was applied.

PCA-transformed features resulted in consistently lower validation accuracy across all α values compared to RBF and Polynomial features. Although the MLP model under PCA maintained a relatively smooth decline in accuracy, it started from a lower baseline (below 98.5%) and experienced steady degradation as α increased. This reflects PCA's compression effect, where crucial discriminative information may have been lost, thereby limiting the neural network's capacity to fit the data optimally.

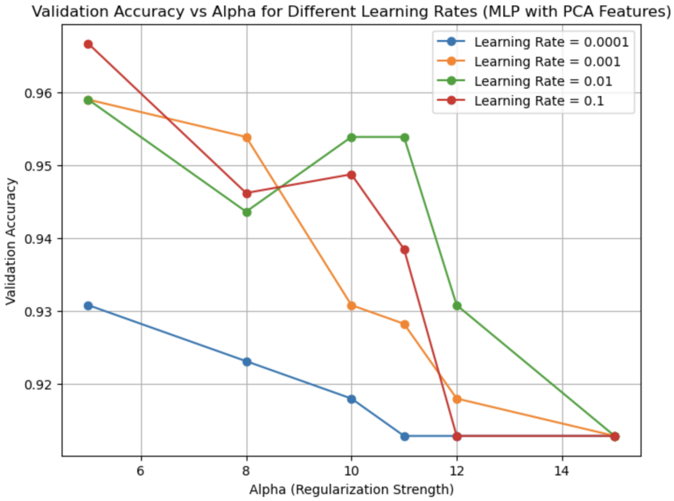

Fig. 8. Validation Accuracy vs. Learning Rate for Logistic Regression Trained with SGD Optimization. A moderate learning rate achieves optimal performance, while very small or very large learning rates degrade accuracy.

-Learning rate Effect of Learning Rate on Validation Accuracy (Neural Network with PCA Features)

The visualization reveals distinct overall patterns in how different learning rates (η) affect MLP model generalization under varying regularization strengths (α):

Low learning rates ($\eta = 0.0001$) resulted in consistently inferior validation accuracies across all regularization levels. The gradual downward slope indicates that models optimized with excessively small learning rates struggled to converge effectively within the fixed iteration budget. This insufficient update magnitude limited the network's ability to fit even moderately regularized data.

Moderate learning rates ($\eta = 0.001$ and $\eta = 0.01$) achieved substantially higher validation accuracies, especially at lower α values. However, as regularization strength increased, performance gradually deteriorated, highlighting the typical bias–variance trade-off. These moderate learning rates struck a better balance between convergence speed and stability under mild to moderate regularization.

High learning rate ($\eta = 0.1$) initially yielded the highest validation accuracy, but performance degraded more sharply as α increased. This behavior suggests that while larger learning rates accelerate initial convergence, they also make the model more sensitive to regularization-induced constraints, leading to optimization instability or premature convergence under stronger penalties.

Overall, moderate learning rates provided the most stable and robust performance across different regularization strengths, while both overly small and overly large learning rates compromised generalization — either through underfitting (slow convergence) or overfitting/instability (sensitivity to regularization).

D. Effect of Hidden Layer in Neural Netork Model

To evaluate the influence of network capacity on model generalization, we test the performance of the Neural Network Model with different nodes. Specificly, we take MLP classifiers with polinomial transformation with two different hidden layer sizes: (8,4) and (4,2) as example and compare the effect.

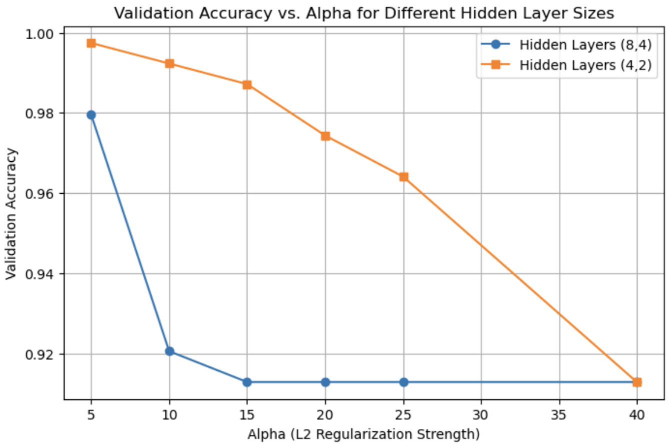

Fig. 9. Validation Accuracy vs. Learning Rate for Logistic Regression Trained with SGD Optimization. A moderate learning rate achieves optimal performance, while very small or very large learning rates degrade accuracy.

As shown in Figure 9, the smaller network consistently outperformed the larger one in terms of validation accuracy across various L2 regularization strengths. This pattern can be attributed to the smaller model's reduced capacity to overfit the training data, especially under conditions of limited sample size and expanded feature dimensionality due to polynomial transformation. In contrast, the larger network, despite regularization, exhibited greater susceptibility to overfitting, resulting in lower validation performance. These results highlight the importance of appropriately matching model complexity to data characteristics to achieve optimal generalization.

E. Effect of Model Complexity

To further explore how model complexity and training duration influence predictive performance, we systematically varied the neural network architectures and the number of training epochs. Specifically, we evaluated Multi-Layer Perceptron (MLP) models with different hidden layer configurations—(4,2), (8,4), and (16,8)—across multiple epoch settings (100, 200, and 300). The objective was to understand the interplay between network capacity, optimization time, and model generalization when trained on PCA-transformed features. Validation accuracy was used as the primary evaluation metric, and results were visualized to highlight the trends across different settings.

Table IV and Figure 10 summarize the validation accuracy of MLP models across different hidden layer configurations and epoch settings. Validation accuracy generally improved as the number of epochs increased from 100 to 300, indicating that extended training enhances convergence and generalization.

TABLE IV
VALIDATION RESULTS FOR DIFFERENT MLP ARCHITECTURES WITH PCA FEATURES

Model	Layers	Hidden Units	Epochs	Validation Acc (%)
1	2	(4, 2)	100	91.28
			200	92.05
			300	93.08
2	2	(8, 4)	100	91.28
			200	93.08
			300	95.13
3	2	(16, 8)	100	93.08
			200	94.62
			300	94.87

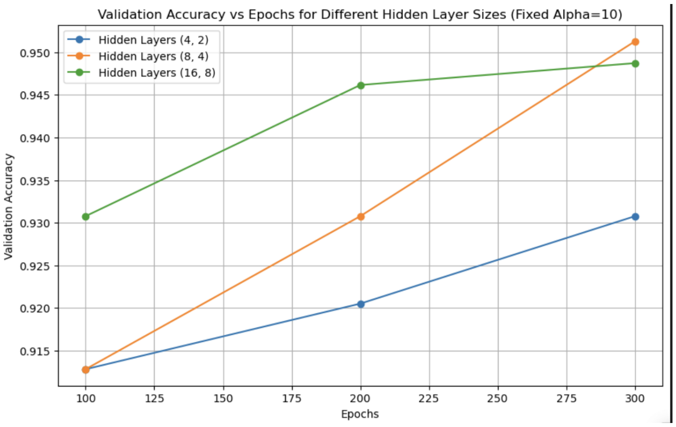

Fig. 10. Validation Accuracy vs. Complexity for Neural Network Trained with different epoch.

The model with hidden layers (8,4) achieved the highest validation accuracy (95.13%) at 300 epochs, outperforming both the smaller (4,2) and larger (16,8) architectures. While the (16,8) model showed strong initial performance, its gains plateaued with more training, suggesting a need for additional regularization. In contrast, the (4,2) model improved steadily but was limited by its smaller capacity.

These findings underscore the importance of balancing network complexity and training duration. A moderate network size, combined with sufficient epochs, yielded the best validation performance under PCA-transformed features.

F. Sample Confusion Matrix Analysis

To illustrate model prediction outcomes more intuitively, a sample confusion matrix was generated based on the SVM model trained with PCA-transformed features and a regularization parameter of $C = 0.01$. This example provides a detailed breakdown of true and false predictions without presenting confusion matrices for all models.

Based on the sample confusion matrix, the model achieved a precision of 1.000, indicating that all instances predicted as positive were indeed true positives with no false positives. The recall was calculated as 0.949, reflecting the model's ability to correctly identify approximately 95% of actual positive cases

TABLE V

SAMPLE CONFUSION MATRIX FOR SVM MODEL (PCA FEATURES, $C = 0.01$)

	Predicted Positive	**Predicted Negative**
Actual Positive	338	18
Actual Negative	0	34

while missing a small fraction. The resulting F1-score of 0.973 demonstrates a strong overall balance between precision and recall, suggesting that the model effectively minimizes both false positives and false negatives under the given feature transformation and regularization setting.

G. Performance Comparison Across Models and Feature Transformations

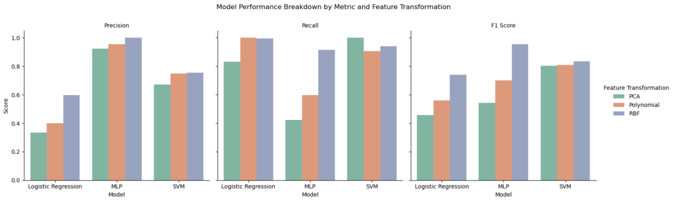

Fig. 11. Model performance comparison across different feature transformations evaluated by precision, recall, and F1 score.

Figure 11 presents a detailed comparison of model performance across different feature transformations. Multi-Layer Perceptron (MLP) models consistently achieved the highest scores across all three metrics, particularly when combined with RBF-transformed features. Logistic Regression models showed competitive precision and recall with polynomial features but exhibited lower F1 scores, suggesting a slight imbalance. Support Vector Machines (SVM) maintained relatively stable performance across different feature transformations.

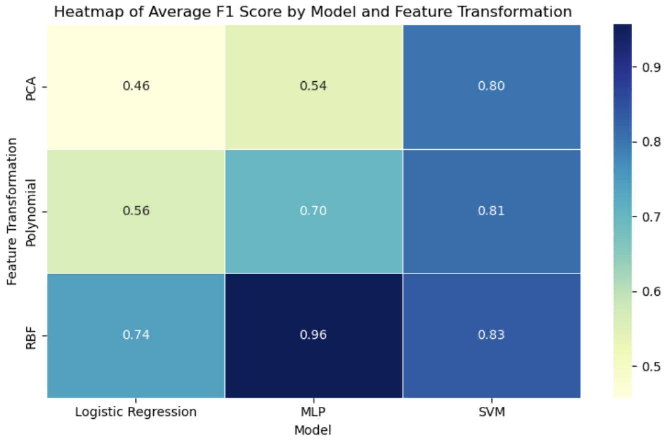

Fig. 12. Heatmap of average F1 score across different models and feature transformations.

Figure 12 further highlights that MLP models with RBF features achieved the highest F1 score (0.96), significantly outperforming other configurations. SVM models demonstrated consistent and strong F1 performance across transformations, while Logistic Regression models benefited most from RBF feature expansion, improving substantially from 0.46 (PCA) to 0.74 (RBF). These results underscore the critical role of non-linear feature mappings in boosting model generalization, particularly for simpler classifiers.

IV. CONCLUSION AND ANALYTICAL DISCUSSION

This study demonstrates the viability of predicting group-level admission outcomes using supervised learning models informed by socioeconomic and behavioral indicators. By focusing on aggregated group characteristics rather than individual profiles, the approach offers valuable insights into structural patterns of educational access and inequality, while mitigating privacy concerns.

Beyond model performance optimization, the findings underscore the critical importance of aligning model complexity with feature expressiveness to ensure meaningful generalization. The observed sensitivity of different models to feature transformations highlights the nuanced interplay between algorithm design and data structure.

Looking forward, future research could extend this work by incorporating longitudinal data to capture temporal shifts in access patterns, exploring fairness-aware modeling strategies, and integrating additional contextual variables such as geographic mobility or institutional characteristics. These efforts would further enhance the understanding of systemic barriers and inform more equitable education policy interventions.