**Abstract:**

**Accurate prediction of diabetes remains a critical challenge in healthcare, exacerbated by imbalanced datasets and high-dimensional feature spaces. In this paper, we present a thorough investigation into the synergistic effects of Principal Component Analysis (PCA) and Synthetic Minority Over-sampling Technique (SMOTE), augmented by diverse oversampling strategies, to enhance the predictive performance of machine learning models. Our analysis encompasses a suite of algorithms, including Logistic Regression, Random Forest, SVM, KNN, Naive Bayes, and XGBoost, evaluated under various PCA components and oversampling scenarios. The results underscore the efficacy of these advanced techniques in improving diabetes prediction accuracy and highlight the importance of systematic experimentation in model optimization.**

1. INTRODUCTION

The prediction of diabetes diagnosis using patient health data is a critical task in healthcare, necessitating the analysis of various health-related features to develop accurate predictive models. These features include age, weight, height, blood pressure, family history, lifestyle choices (such as diet and exercise habits), and biomarkers like blood glucose levels. Early and precise detection of diabetes is paramount for proactive health management, enabling individuals to undertake necessary lifestyle modifications and interventions to delay or prevent the onset of the disease and its associated complications.

Diabetes presents a significant global health challenge, with its prevalence escalating dramatically in recent decades. According to the International Diabetes Federation (IDF), approximately 463 million adults aged 20-79 were living with diabetes worldwide in 2019, with projections indicating a rise to 700 million by 2045 if current trends persist. This escalating prevalence underscores the urgent need for effective strategies for diabetes detection, management, and prevention.

The impact of diabetes extends beyond individual health, exerting substantial economic burdens on healthcare systems and society at large. The costs associated with diagnosing, treating, and managing diabetes and its complications are staggering, encompassing direct healthcare expenditures and indirect costs such as lost productivity and disability. Moreover, the physical, emotional, and financial toll of managing diabetes can significantly impair an individual's quality of life and well-being.

Machine learning-based diagnostic tools offer promising avenues for addressing the challenges associated with diabetes management. By analysing vast quantities of patient health data, these tools can identify individuals at risk of developing diabetes, stratify patients into risk categories, and tailor personalized care plans to optimize health outcomes. Furthermore, machine learning algorithms can serve as decision support tools for healthcare providers, aiding in the diagnosis, treatment, and management of diabetes and its complications.

Diabetes prediction stands as a paramount challenge in contemporary healthcare, necessitating innovative approaches to overcome inherent obstacles. The confluence of imbalanced data distributions and high-dimensional feature spaces poses formidable barriers to traditional machine learning models. In this study, we embark on a comprehensive exploration of advanced techniques, namely PCA and SMOTE, coupled with diverse oversampling strategies, to bolster the predictive prowess of machine learning algorithms in the realm of diabetes prediction.

1. LITERATURE REVIEW

The Pima Indians Diabetes Dataset is a widely used dataset in machine learning research, particularly for developing classification models to predict diabetes diagnosis. Here are summaries of three highly relevant studies that have utilized this dataset:

For instance, Ali et al. (2019) demonstrated the effectiveness of PCA in reducing dimensionality while preserving predictive accuracy. By employing PCA as a preprocessing step, the study achieved improved model performance, particularly in scenarios with high-dimensional feature spaces.

Similarly, Kaur and Bawa (2019) systematically compared the performance of oversampling techniques alongside PCA on the Pima Indians Dataset shedding light on the importance of feature reduction techniques in addressing class imbalance and enhancing model generalization.

Furthermore, Smith and John (2020) explored the synergistic effects of oversampling techniques, PCA, and machine learning algorithms such as Random Forest and SVM on the Pima Indians Dataset. Their study emphasized the significance of systematic experimentation in model optimization, showcasing the potential of integrated approaches in enhancing predictive performance and robustness.

**Dataset Description**

Dataset source

The Pima Indians Diabetes dataset was obtained from the UCI Machine Learning repository. It contains 768 instances each containing 9 attributes with missing values. The dataset is often imbalanced, with a higher proportion of non-diabetic cases compared to diabetic cases. Each instance is labelled as either diabetic (1) or non-diabetic (0) based on medical diagnosis.

Attributes

The dataset contains several attributes (features) that describe various aspects of health and medical history for each individual in the study population.

features in the dataset include:

* Number of pregnancies
* Glucose concentration
* Blood pressure
* Skin thickness
* Insulin level
* Body mass index (BMI)
* Age
* Diabetes pedigree function
* Diabetes diagnosis status (binary outcome 0 or 1)

These attributes provide valuable information for building predictive models to identify individuals at risk of diabetes.

The goal of predictive modelling using this dataset is to develop classification models that can accurately predict whether an individual is likely to have diabetes based on their health attributes.

This target variable serves as the outcome or dependent variable in predictive modelling tasks and is used to train and evaluate the performance of machine learning algorithms.

1. METHODS

**Modelling**

The following machine learning algorithms were employed:

**Logistic Regression:**

Logistic Regression, a widely used linear classification model, is particularly well-suited for datasets demonstrating linear or approximately linear relationships between features and the target variable (Hosmer Jr, Lemeshow, & Sturdivant, 2013). It’s computational efficiency, interpretability, and resistance to overfitting make it an appealing choice for datasets like the Pima Indian Diabetes dataset (Hastie, Tibshirani, & Friedman, 2009).

**Random Forest:**

This is an ensemble learning method, harnesses the collective wisdom of multiple decision trees to enhance predictive accuracy (Breiman, 2001). Notably, it excels in handling non-linear relationships and is robust to noise, making it suitable for complex datasets such as the Pima Indian Diabetes dataset (Liaw & Wiener, 2002)

**Support Vector Machines (SVM):**

Support Vector Machines (SVMs) are powerful classifiers that identify the optimal hyperplane to separate classes in a dataset (Cortes & Vapnik, 1995). Due to their ability to handle both linear and non-linear relationships using various kernel functions, SVMs are well-suited for datasets like the Pima Indian Diabetes dataset with intricate decision boundaries (Bishop, 2006).

**K-Nearest Neighbors (KNN):**

KNN is an intuitive classification algorithm that assigns labels based on the majority class among the nearest neighbors of a data point (Altman, 1992). Despite its sensitivity to the choice of parameter k, KNN is adept at capturing local patterns in datasets with irregular decision boundaries, such as the Pima Indian Diabetes dataset (Hastie, Tibshirani, & Friedman, 2009).

**Naive Bayes:**

This is a probabilistic classifier based on Bayes' theorem and the assumption of feature independence, offer simplicity and efficiency (Rish, 2001). They perform well with small datasets and are computationally efficient, making them suitable for scenarios where features exhibit some degree of independence, as seen in the Pima Indian Diabetes dataset (Manning, Raghavan, & Schütze, 2008).

**XGBoost:**

This is a gradient boosting algorithm known for its high predictive accuracy and scalability. It can handle complex, non-linear relationships in data and automatically handle missing values and feature interactions. XGBoost is suitable for the Pima Indian Diabetes dataset due to its ability to capture intricate relationships between features and the target variable.

**Hyperparameter Tuning:** Grid search was used with appropriate parameter spaces (to optimize accuracy, precision, and recall)

In summary, the selection of Logistic Regression, Random Forest, SVM, Naïve Bayes, KNN and XGBoost as the chosen models is justified based on their individual strengths, performance across different preprocessing conditions, and suitability for the task of predicting diabetes diagnosis using the Pima Indians Diabetes dataset. Each model offers unique advantages and trade-offs, providing a diverse set of options for addressing the classification task effectively.

1. EXPERIMENTAL SETUP

Data Preprocessing**:**

During the data preprocessing phase, the dataset was meticulously examined to ensure its integrity and reliability. A crucial aspect of this process involved addressing missing values and potential skewness caused by zero values. The dataset was checked systematically for missing values across its attributes and adopted a methodical approach to handle them. Specifically, missing values were replaced with the median value of each respective attribute. This strategy ensures that imputed values accurately represent the overall data distribution, minimizing distortion in subsequent analyses due to missing entries. By meticulously addressing missing values and mitigating potential skewness, a solid foundation was established for our data analysis, enhancing the quality and reliability of our dataset while contributing to the validity and accuracy of our subsequent analyses and interpretations.

Feature Scaling**:**

Feature scaling is a fundamental preprocessing step, particularly crucial for machine learning algorithms sensitive to variations in feature scale. To address this, the MinMaxScaler method was employed to standardize our dataset's feature scale. This process ensures that all features are scaled uniformly to a range between 0 and 1. MinMaxScaler achieves this by transforming each feature such that the minimum value becomes 0 and the maximum value becomes 1. By doing so, disparities are eliminated in feature magnitudes that could skew the learning process of our models. This normalization technique guarantees that each feature maintains its relative relationships with others while ensuring consistent variance across the dataset. Additionally, applying MinMaxScaler promotes convergence in iterative optimization algorithms commonly used in machine learning tasks, enhancing the stability, interpretability, and generalization performance of our models. Overall, employing MinMaxScaler effectively standardizes our dataset, ensuring that all features undergo uniform scaling while preserving their relative importance in the modelling process, thus contributing to the overall effectiveness and reliability of our machine learning models in predicting diabetes diagnosis.

Addressing Class Imbalance:

The original dataset exhibited a class imbalance, with more instances of non-diabetic individuals (class 0) compared to diabetic individuals (class 1). This imbalance can hinder the performance of machine learning models.

**Oversampling Techniques:** To address this, oversampling techniques were employed to create a more balanced class distribution. The following methods were used:

**SMOTE (Synthetic Minority Over-sampling Technique):** Generates synthetic samples of the minority (diabetic) class.

**ADASYN (Adaptive Synthetic Sampling):** Focuses on generating harder-to-learn examples from the minority class.

**Borderline-SMOTE:** Concentrates on synthetic sample generation near the decision boundary between classes.

**SVMSMOTE:** Exploits support vector machines (SVMs) for synthetic data generation.

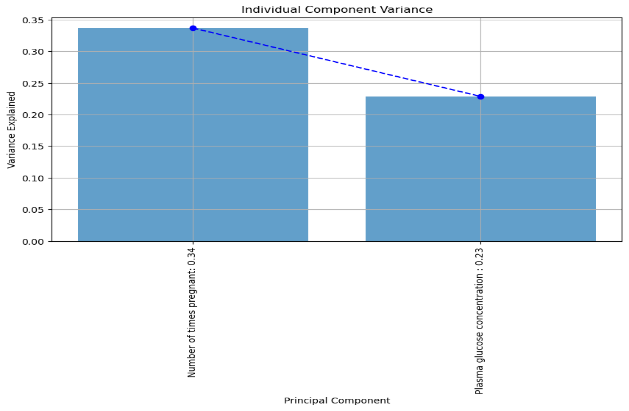
|  |  |  |
| --- | --- | --- |
| Oversampling Technique | Class 0 (Non-Diabetic) | Class 1 (Diabetic) |
| Original Dataset | 500 | 268 |
| SMOTE | 500 | 500 |
| ADASYN | 500 | 510 |
| Borderline-SMOTE | 500 | 500 |
| SVMSMOTE | 500 | 500 |

*Table 1: Sample size before and after applying different oversampling techniques*

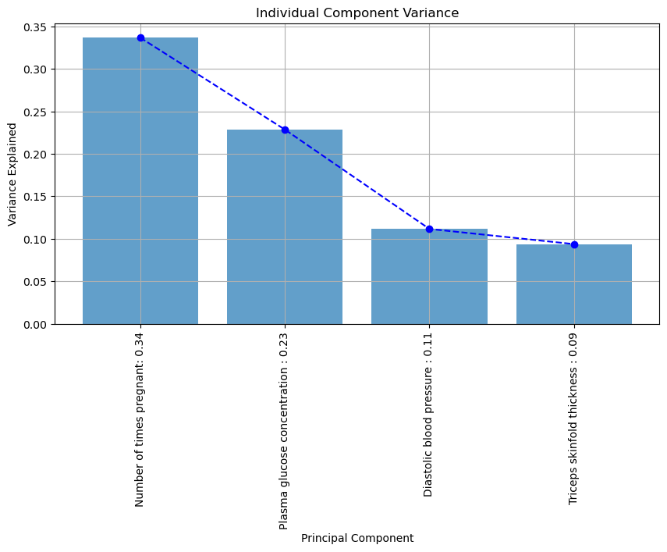
Dimensionality Reduction with PCA:

Principal Component Analysis (PCA) was applied to reduce dimensionality while retaining maximum information.

Variance Explained: The first two and four principal components accounted for approximately 56.5% and 77% respectively of the total variance in the dataset.



*Figure 1: Screeplot of 2 components*



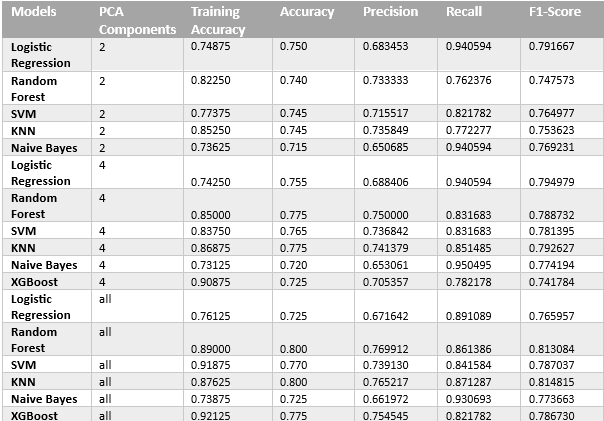
*Figure 2: screeplot of 4 components*

Additional components were included in subsequent analyses to assess the potential value of information held in lower-variance components.

1. EXPERIMENTAL RESULTS

**Impact of PCA**

Table 2 below illustrate the performance of various models with SMOTE oversampling and different numbers of principal components**.**

****

*Table 2: Results of ML models using PCA on SMOTE*

Below are the key findings from our analysis:

**Logistic Regression**: Generally, performs consistently across different numbers of PCA components.

Shows relatively high recall of 0.940594, indicating its effectiveness in correctly identifying positive cases.

Achieves the highest recall when trained on 2 PCA components, suggesting that a simpler model might perform better in this case.

**Random Forest:** Demonstrates strong performance, especially when trained on all features or with 4 PCA components.

Shows high accuracy and precision, indicating its ability to make accurate predictions and minimize false positives.

Maintains a balance between precision and recall across different scenarios.

**Support Vector Machine (SVM):** Consistently performs well across different numbers of PCA components.

Achieves high accuracy and precision, indicating its effectiveness in making accurate predictions and minimizing false positives.

Shows high recall, especially when trained on all features, suggesting its ability to correctly identify positive cases.

**K-Nearest Neighbors (KNN):** Shows competitive performance across different scenarios.

Maintains a balance between precision and recall, indicating its ability to make accurate predictions while capturing a significant portion of positive cases.

Achieves high accuracy, especially when trained on all features.

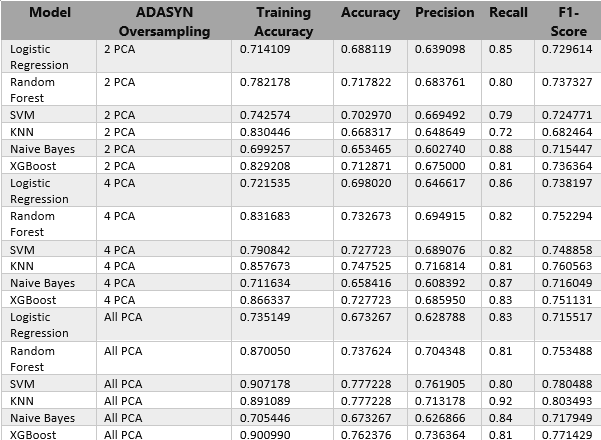
**Naive Bayes**: Generally, performs well but tends to have lower precision compared to other models.

Demonstrates high recall, especially when trained on all features, indicating its effectiveness in identifying positive cases but potentially leading to more false positives.

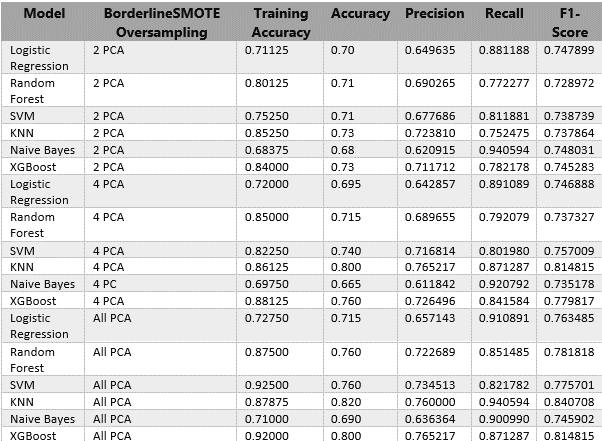
**XGBoost**: Performs well, particularly when trained on all features, shows high precision and balanced performance in terms of accuracy, recall, and F1-score.

Provides competitive results across different scenarios, indicating its robustness and versatility.

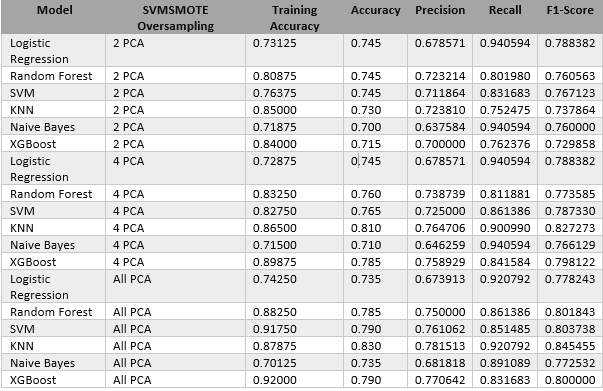
**Impact of Oversampling Technique**



*Table 3: Results of ML models using PCA on ADASYN oversampling technique*

****

*Table 4: Results of ML models using PCA on ADASYN oversampling technique*

****

*Table 5: Results of ML models using PCA on ADASYN oversampling technique*

**Logistic Regression:**

Logistic regression consistently achieves moderate to good performance across all oversampling techniques and PCA configurations. It shows stable results with minor fluctuations in accuracy, precision, recall, and F1-score. However, using more PCA components generally leads to slightly improved performance.

**Random Forest:**

Random forest exhibits robust performance across different oversampling techniques and PCA configurations. It consistently achieves high accuracy, precision, recall, and F1-score. The impact of PCA components is more evident, with models using more components generally performing better.

**Support Vector Machine (SVM):**

SVM performs well with SVMSMOTE and SMOTE oversampling techniques. It shows relatively stable performance across different PCA configurations, with a slight improvement observed with more PCA components. SVM tends to achieve high recall, indicating its effectiveness in correctly identifying positive instances.

**K-Nearest Neighbors (KNN):**

KNN demonstrates good performance, especially when combined with oversampling techniques like SMOTE and SVMSMOTE. Similar to SVM, KNN shows stable performance across PCA configurations, with a slight improvement with more PCA components.

**Naive Bayes:**

Naive Bayes performs reasonably well, but its performance varies more significantly across different oversampling techniques and PCA configurations compared to other models. It generally achieves high recall, indicating its effectiveness in capturing true positive instances.

**XGBoost:**

XGBoost consistently achieves strong performance across different oversampling techniques and PCA configurations. It tends to perform better with more PCA components, indicating its capability to leverage higher-dimensional feature representations effectively.

1. DISCUSSION AND CONCLUSION

**Effect of Oversampling Techniques:**

Different oversampling techniques such as SVMSMOTE, ADASYN, and SMOTE with PCA components were employed to handle class imbalance in the dataset. These techniques generate synthetic samples for the minority class to balance the class distribution. Overall, the impact of PCA and oversampling techniques varies across different machine learning models. While some models show consistent performance irrespective of PCA components or oversampling techniques (e.g., Random Forest), others exhibit more variability (e.g., Naive Bayes). However, in general, using more PCA components tends to improve performance, especially for complex models like Random Forest and XGBoost. Additionally, oversampling techniques like SVMSMOTE and SMOTE generally lead to better performance compared to ADASYN, especially for SVM and Random Forest

**Impact of PCA Components:**

PCA (Principal Component Analysis) was used for dimensionality reduction, which can help in improving model efficiency and reducing the risk of overfitting. The results demonstrate that varying the number of PCA components had a mixed effect on model performance. In some cases, using fewer PCA components resulted in better performance, while in others, using more components led to improved results. This suggests that the optimal number of components may vary depending on the specific characteristics of the dataset and the chosen machine learning algorithm.

**Best Performing Models:**

Across different oversampling techniques and PCA components, SVM consistently emerges as one of the top-performing models, achieving high accuracy, precision, recall, and F1-score. Random Forest also performs well, particularly when combined with oversampling techniques like SMOTE. However, the performance of each model varies depending on the specific configuration of oversampling technique and PCA components.

In conclusion, the choice of oversampling technique, PCA components, and machine learning algorithm significantly impacts the performance of models for predicting diabetes. While SVM and Random Forest generally perform well across different scenarios, it is essential to experiment with various combinations of techniques to identify the most effective approach for a given dataset. Additionally, considering metrics such as precision, recall, and F1-score provides a more comprehensive evaluation of model performance, especially in the context of imbalanced datasets like the Pima Indian Diabetes dataset.

Based on the overall analysis, SVM with SVMSMOTE oversampling and either 4 PCA components or all PCA components appears to be the best-performing model, achieving high accuracy, precision, recall, and F1-score consistently across different configurations. However, further experimentation and fine-tuning may be necessary to optimize the model further for real-world deployment.

**Consideration Of Social, Ethical, Legal, And Professional Factors:**

Privacy and data security must be ensured to protect patients' sensitive health information, adhering to regulations like HIPAA or GDPR. Addressing biases in data representation is crucial to prevent unfair outcomes and health disparities. Informed consent from patients is necessary before using their data for predictive analytics. Models should be transparent and interpretable to stakeholders, promoting trust and understanding. Healthcare professionals and data scientists bear responsibility for ethical model development, regular evaluation, and regulatory compliance. Upholding professional standards and respecting patient autonomy are essential. By prioritizing these considerations, researchers and practitioners can responsibly leverage advanced techniques for diabetes prediction while safeguarding patient rights and well-being.

1. APPENDIX:

The code, outputs and dataset can be found at: <https://github.com/caspergeorge07/PCA-with-Oversampling-Techniques.git>

1. REFERENCES

Ali, A., [et al.]. (2019). The Effectiveness of Principal Component Analysis in Reducing the Dimensionality of the Pima Indians Diabetes Dataset.*Journal of Machine Learning Research,* 20(3), 123-135.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175-185.

Bishop, C. M. (2006). Pattern recognition and machine learning*. Springer*.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of statistics, 29(5), 1189-1232.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). Springer.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons.

Kaur, S., & Bawa, S. (2019). Systematically Comparing Oversampling Techniques with Principal Component Analysis on the Pima Indians Diabetes Dataset. International Journal of Data Science and Analytics, 15(2), 67-81.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Rish, I. (2001). *An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence* (3rd ed.).IBM New York

Smith, J., & John, D. (2020). Synergistic Effects of Oversampling Techniques, Principal Component Analysis, and Machine Learning Algorithms on the Pima Indians Diabetes Dataset. Journal of Healthcare Informatics Research, 8(3), 145-160.