Parkinson's Disease Prediction with Random Forest Classification and Principal Component Analysis

Biomedical Informatics

Aura Aksha Karisma

Introduction

Parkinson's Disease: A disorder of the nervous system that affects movement and balance due to the damage or death of nerve cells in the brain.

Principal Component Analysis (PCA) : the technique of extracting relevant features from a dataset.

This project compares the performance of random forest classification models with or without PCA and with different values of test/train split (25/75, 50/50, 75/25).

Dataset

- Dataset : Parkinson's Disease Classification Dataset dari repository University of California (UCI)
- Number of instances: 756 (564 positive; 192 negative)
- Number of attributes: 755 (no missing values)
- Attributes includes id, gender, and speech features extracted from recordings used in diagnosis process.
- Attribute target: *class* (0 negative; 1 positive)

Experimental Results

- The metrics used to measure the model's performance are precision, recall/sensitivity, accuracy, dan F1 Score

Metrics	Random Forest					
	75% Train Split		50% Train Split		25% Train Split	
	No PCA	With PCA	No PCA	With PCA	No PCA	With PCA
Precision	89,542	82,036	87,338	79,420	84,348	76,015
Recall	97,163	97,163	95,390	97,163	91,726	97,400
Accuracy	89,418	82,011	86,243	79,101	81,129	75,132
F1-Score	93,197	88,961	91,186	87,400	87,882	85,389
No. of Features	755	154	755	131	755	91

- The table above lists the result of each classification model. Generally, the random forest model with 75% train split without PCA gives better result (precision, accuracy, f1-score).
- The highest recall value is achieved by the random forest model with 25% train split and PCA.

Comparative Analysis

- The highest precision and f1-score values are achieved by the model with 75% train split and without PCA.
- The highest recall value is achieved by the model with 25% train split and PCA. Therefore, in the case when the rate of positive cases that are successfully predicted is more important, this model can be implemented instead.
- The highest accuracy value is achieved by the model without PCA that was implemented on the paper.
- In comparison, the model with 75% train split generally performs better than the rest of the models with or without the implementation of PCA.

Metrics	Random Forest							
	75% Train Split		50% Train Split		25% Train Split		On Paper	
	No PCA	With PCA	No PCA	With PCA	No PCA	With PCA	No PCA	With PCA
Precision	89,542	82,036	87,338	79,420	84,348	76,015	70,175	35,087
Recall	97,163	97,163	95,390	97,163	91,726	97,400	70,175	55,555
Accuracy	89,418	82,011	86,243	79,101	81,129	75,132	89,534	76,651
F1-Score	93,197	88,961	91,186	87,400	87,882	85,389	77,669	43,010

Conclusion

This project compares the performance of random forest models with or without PCA and with different values of test/train split (25/75, 50/50, 75/25) in classifying Parkinson's Disease based on speech features dataset.

The result shows that the random forest model with 75% train split (25/75) generally gives better performance compared to the other models available. Furthermore, the model implemented without PCA performs better than the one with PCA. Regardless, to choose the model to implement, users need to adjust it according to their specific needs, one of which is based on the metric deemed as more important for the classification.

THANK YOU

PAPER

Gupta, I., Sharma, V., Kaur, S., & Singh, A. K. (2022). PCA-RF: An Efficient Parkinson's Disease Prediction Model based on Random Forest Classification. *arXiv preprint arXiv:2203.11287*.

https://arxiv.org/pdf/2203.11287

DATASET

Parkinson’s Disease Classification Dataset

https://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification