



# **DATA ANALYTICS**

# **PROJECT PRESENTATION**

# **FEATURE EXTRACTION**

# **USING GENETIC ALGORITHM**

Shreyas Battula    PES1UG20CS405  
Kiran H Kademani    PES1UG20CS654  
Vinay Kumar S    PES1UG20CS701



# PROJECT PRESENTATION

- Feature selection has become a fundamental step of data processing for training a machine learning model due to the increasing amount of high dimensional data.
- There are a huge number of state-of-the-art algorithms that aim to optimize feature selection including genetic algorithms. While not the best performing technique, genetic algorithms provide an exciting solution based on evolution and are widely used in fields such as robotics, marketing or medicine.
- It is worth mentioning that the selection of individuals for reproduction is stochastic, so the best individuals will not always be selected, helping the algorithm not to fall into local minimums.

# Implementation Details



## Genetic Algorithm Steps

Modules Used:

1. Initialise Random Population
2. Fitness Evaluation
3. Selection of Best Parents
4. CrossOver
5. Mutation
6. Generation

# OUTPUTS

## Performance of Models when All Features are Used

### 2. Accuracies Of All Models

```
score1 = acc_score(data_bc, label_bc)  
score1
```

	Classifier	Accuracy
0	RandomForest	0.972028
1	LinearSVM	0.958042
2	Logistic	0.958042
3	GradientBoosting	0.958042
4	RadialSVM	0.951049
5	AdaBoost	0.951049
6	DecisionTree	0.930070

Performance of Random Forest Classifier When GA is used for Features Selection.

- An improvement of 2%

```
logmodel = RandomForestClassifier(n_estimators=200, random_state=0)
X_train,X_test, Y_train, Y_test = split(data_bc,label_bc)
chromo_df_bc,score_bc=generations(data_bc,label_bc,size=80,n_feat=data_bc.shape[1],n_parents=64,mutation_rate=0.20)
💡 | | | | | X_train = X_train,X_test = X_test,Y_train = Y_train,Y_test = Y_test)
```

[17]

```
... Best score in generation 1 : [0.986013986013986]
      Best score in generation 2 : [0.993006993006993]
      Best score in generation 3 : [0.993006993006993]
      Best score in generation 4 : [0.993006993006993]
      Best score in generation 5 : [0.993006993006993]
```

# List of Selected Features For Every Generation

The features selected for generation 1 are
['radius_mean', 'texture_mean', 'smoothness_mean', 'compactness_mean', 'symmetry_mean', 'radius_se', 'perimeter_se', 'area_se', 'smoothness_se', 'concavity_se', 'concave points_se', 'symmetry_se', 'radius_worst', 'area_worst', 'smoothness_worst', 'concavity_worst', 'symmetry_worst']
The features selected for generation 2 are
['radius_mean', 'texture_mean', 'perimeter_mean', 'smoothness_mean', 'compactness_mean', 'fractal_dimension_mean', 'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se', 'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se', 'fractal_dimension_se', 'radius_worst', 'area_worst', 'smoothness_worst']
The features selected for generation 3 are
['radius_mean', 'texture_mean', 'smoothness_mean', 'compactness_mean', 'fractal_dimension_mean', 'radius_se', 'perimeter_se', 'area_se', 'smoothness_se', 'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se', 'fractal_dimension_se', 'radius_worst', 'smoothness_worst', 'concave points_worst']
The features selected for generation 4 are
['concavity_mean', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se', 'compactness_se', 'concave points_se', 'perimeter_worst', 'area_worst', 'smoothness_worst', 'compactness_worst', 'concavity_worst', 'fractal_dimension_worst']
The features selected for generation 5 are
['perimeter_mean', 'concavity_mean', 'texture_se', 'perimeter_se', 'area_se', 'compactness_se', 'concavity_se', 'concave points_se', 'smoothness_worst', 'concavity_worst', 'fractal_dimension_worst']

# Conclusion

Feature extraction is an important data preprocessing step that allows us to improve the accuracy of our results.

---

Genetic Algorithms are stochastic thus not always falling into the local minimum making it one of the best algorithms to optimise feature selection





# Future Work

- Compare Model Metrics Against various Selection Techniques
- Try to reduce computational time and achieve the required accuracy with minimal number of generations.
- Try to avoid the risk of overfitting of feature selection when observations is insufficient .
- Try A Hybrid GA for larger datasets



# References

<https://towardsdatascience.com/feature-selection-using-genetic-algorithms-d3f5fc7bbef1>



<https://medium.com/analytics-vidhya/feature-selection-using-genetic-algorithm-20078be41d16>



[https://www.neuraldesigner.com/blog/genetic\\_algorithms\\_for\\_feature\\_selection](https://www.neuraldesigner.com/blog/genetic_algorithms_for_feature_selection)



*Thank  
you!*