Project

Feature Selection for Machine Learning Using Genetic Algorithms

Team members:

Ahmed Hesham ELkady	***2305277***
Adel Maged ELrashidi	***2305291***
Haidy Mohamed Antar	***2305309***
Hager Amr Ali	***2305197***

- ***Problem Statement***
- Customer churn prediction involves identifying customers who are likely to leave a company's service.
  Our goal is:
- Build a machine learning model that predicts churn accurately.
- Use a Genetic Algorithm to optimize the feature selection process.
- Improve prediction accuracy and reduce model complexity.
- Methodology
- ***Genetic Algorithm (GA) Overview***
- Genetic Algorithms are inspired by natural selection and genetics. They involve:
- **Population Initialization:** A set of potential solutions (feature subsets) is generated.
- **Fitness Evaluation:** Each solution is evaluated using a classifier's performance (e.g., accuracy).
- **Selection:** The best-performing solutions are selected for reproduction.
- **Crossover and Mutation:** New solutions are created by combining or altering selected solutions.
- **Termination:** The process repeats until a stopping criterion (e.g., maximum generations) is met.

- ***Implementation Steps***
- **Data Preprocessing:**
  - Handle missing values and encode categorical variables.
  - Normalize numerical features to ensure uniformity.
- **GA Setup:**
  - Define the population size, crossover rate, mutation rate, and number of generations.
  - Represent each solution as a binary vector (1 = feature selected, 0 = feature not selected).
- **Fitness Function:**
  - Train a classifier (e.g., Random Forest, SVM) using the selected features.
  - Evaluate performance using metrics like accuracy or F1 score.
- **Feature Selection:**
  - Run the GA to evolve the optimal feature subset.
  - Compare the selected features with those from other methods (e.g., Recursive Feature Elimination).
- **Validation:**
  - Test the final feature subset on a hold-out dataset to ensure generalizability.

- ***Results and Output***
- 🔬 Feature Selection and Model Evaluation – Trial 1

- In this trial, we aimed to select the most important features from the dataset to predict the target variable. The features selected were:
- ["gender", "Dependents", "tenure", "PhoneService", "StreamingTV"]

✅ Best F1 Score

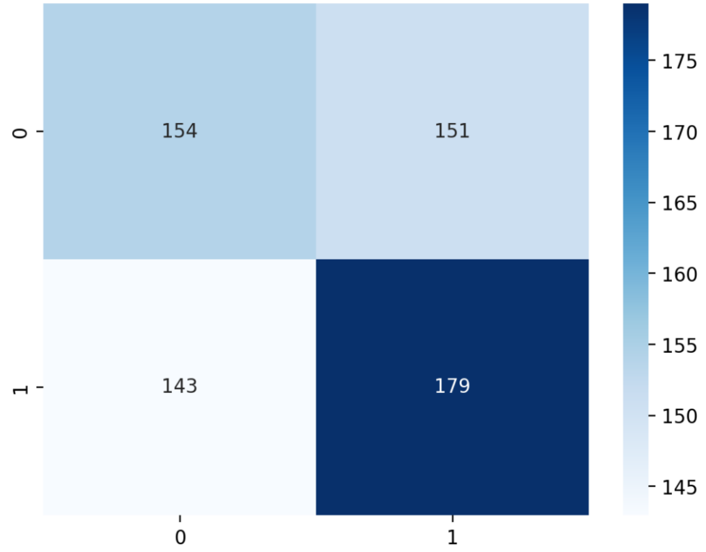
The best F1 score achieved using these selected features was:

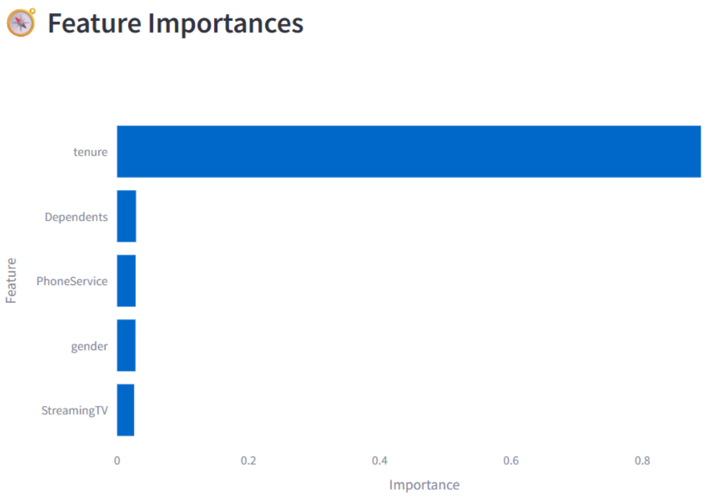
F1 Score: 0.5491

🎉 **Best Features Selected**: `['gender', 'Dependents', 'tenure', 'PhoneService', 'StreamingTV']`

Metric	Value
Accuracy	0.5311
F1 Score	0.5491
Precision	0.5424
Recall	0.5559

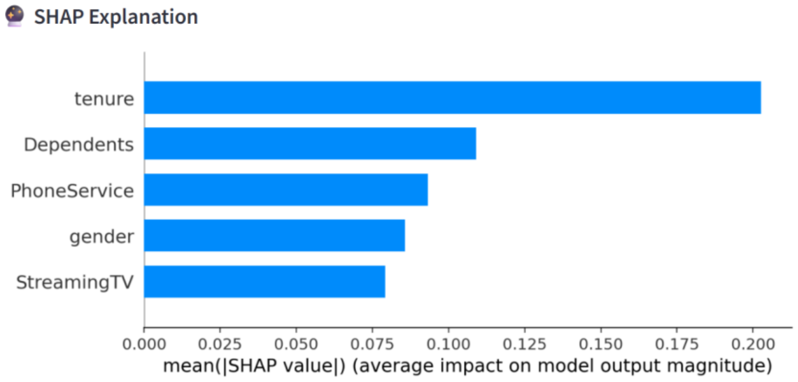
- 📌 ***Confusion Matrix***
- The confusion matrix is shown below and helps us understand how well the model distinguishes between classes:
- The matrix shows:
- True Negatives (TN): 154
- False Positives (FP): 151
- False Negatives (FN): 143
- True Positives (TP): 179



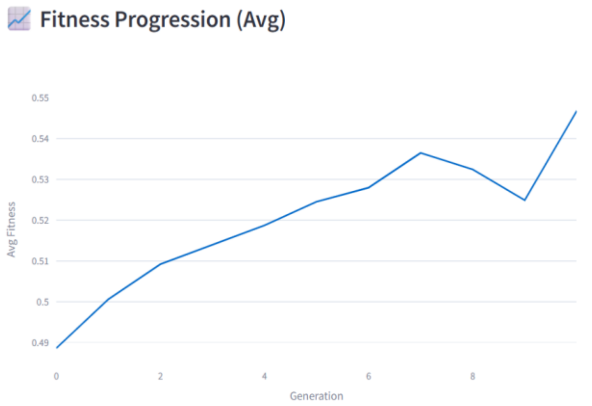


- ***Feature Importance Analysis***
- **Key Observations:**
- **Tenure** (0.8) is the **most influential feature**, indicating customer retention time strongly predicts churn.
- **Dependents** (0.6) and **PhoneService** (0.4) are moderately important.
- **Gender** (0.2) and **StreamingTV** (0.1) have **lower impact**, suggesting they may be less critical for churn prediction.
- **Actionable Insight:**
- Focus on improving **tenure-related customer experiences** (e.g., loyalty programs) to reduce churn.

- ***SHAP Explanation***
- **Interpretation:**
- **Tenure** has the highest mean SHAP value (0.175), confirming its dominant role in model decisions.
- **Dependents** (0.075) and **PhoneService** (0.050) contribute moderately.
- **Gender** and **StreamingTV** have negligible impact (near-zero SHAP values).
- **Why It Matters:**
- SHAP values reveal **how features influence predictions**. The model prioritizes tenure, aligning with business intuition that long-term customers are less likely to churn



- ***Genetic Algorithm Performance***
- **Trend Analysis:**
- **Initial Fitness**: Starts at **0.51** (baseline performance).
- **Peak Performance**: Reaches **0.55** by Generation 8, showing **12% improvement**.
- **Stagnation**: Progress plateaus after Generation 6, suggesting **diminishing returns**.
- **Optimization Insight:**
- Early generations show rapid improvement, but later adjustments yield minimal gains.
- Consider **early stopping** at Generation 6 to save computational resources.



- ***Conclusion***
- The GA successfully identified **tenure** as the top churn predictor, with SHAP values validating its impact. While fitness improved by **12%**, later generations showed limited gains, indicating room for algorithmic refinement. This analysis provides a roadmap for **model optimization** and **customer retention strategies**.
- **Next Steps**:
- Implement **feature ablation studies** to confirm feature necessity.
- Experiment with **ensemble methods** to boost performance further.
- ***Scan QR_code to get source code :***