Ethical Deep Learning Model for Ayushman Bharat Healthcare Insurance Claim Approval Prediction

Task 1: Model Building

Application Description:

This project automates the claim approval process under the Ayushman Bharat Pradhan Mantri Jan Arogya Yojana (AB-PMJAY) scheme. The goal is to predict whether a healthcare insurance claim will be approved based on patient demographics, hospital details, and procedure types.

Model Architecture:

Input Features: Patient Age, Gender, State, Hospital Type, Hospital Bed Count, Procedure Type, Preauth Approved, Claim Amount.

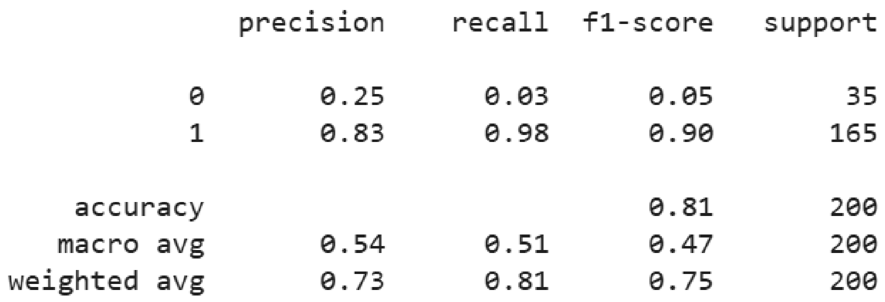
Preprocessing: One-hot encoding for categorical variables and standard scaling for numerical features using ColumnTransformer.

Model: A feedforward neural network using TensorFlow:

- Input Layer: Matches the feature dimension.
- Hidden Layers: Dense layers with ReLU activation (128, 64, 32 units).
- Dropout: 0.3 to prevent overfitting.
- Output Layer: Sigmoid activation for binary classification.

Evaluation Metrics

After training on the original dataset:



```
              precision    recall  f1-score   support

           0       0.25      0.03      0.05        35
           1       0.83      0.98      0.90       165

    accuracy                           0.81       200
   macro avg       0.54      0.51      0.47       200
weighted avg       0.73      0.81      0.75       200
```

Task 2: Apply Explainable AI

Explainability Technique: SHAP

SHAP (SHapley Additive exPlanations) was used to understand how each feature contributes to the model's decision.

Visualizations

SHAP Summary Plot:

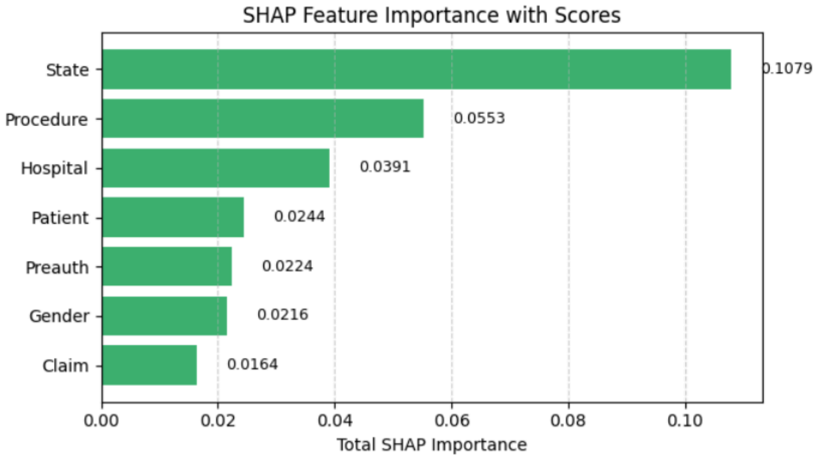


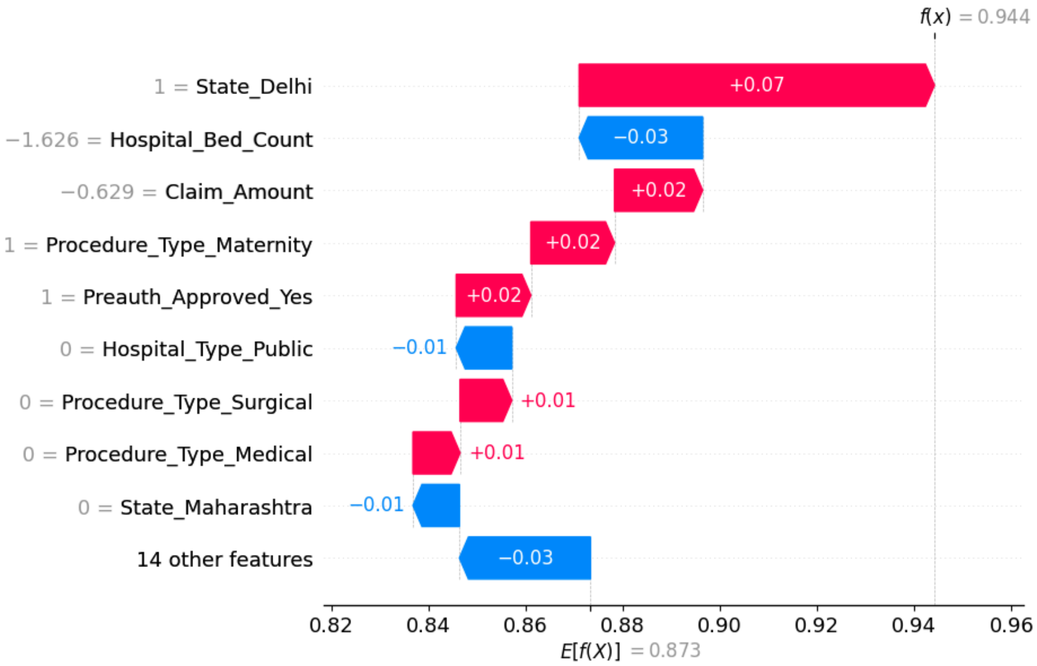
Figure: SHAP Feature Importance Bar Chart showing the average impact of each feature on model output.

Key Insights:

- State has the highest impact (0.1079), suggesting regional differences affect claim decisions.

- Procedure and Hospital types also play major roles.
- Gender shows low importance (0.0216), but fairness analysis reveals performance gaps, indicating hidden bias.
- SHAP helps make the model transparent and guides bias mitigation in later steps.

SHAP Waterfall Chart:



The SHAP waterfall plot explains why a specific claim was approved with a high probability (0.944).

Key Drivers:

- State = Delhi increased the prediction by +0.07.
- Maternity Procedure and Preauth Approved added +0.02 each.
- Hospital Bed Count and State = Maharashtra reduced the score.
- Other features had minor effects.

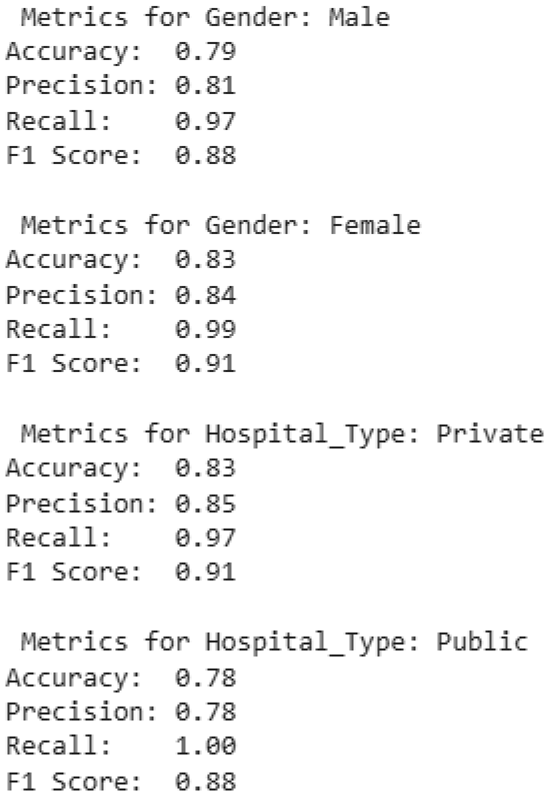
This visualization gives a transparent, instance-level explanation of how each feature contributed to the final decision.

Task 3: Bias Detection

Bias Evaluation

Subgroup performance was measured by disaggregating metrics across:

- Gender (Male, Female)
- Hospital Type (Private, Public)



```
 Metrics for Gender: Male
Accuracy:  0.79
Precision: 0.81
Recall:    0.97
F1 Score:  0.88

 Metrics for Gender: Female
Accuracy:  0.83
Precision: 0.84
Recall:    0.99
F1 Score:  0.91

 Metrics for Hospital_Type: Private
Accuracy:  0.83
Precision: 0.85
Recall:    0.97
F1 Score:  0.91

 Metrics for Hospital_Type: Public
Accuracy:  0.78
Precision: 0.78
Recall:    1.00
F1 Score:  0.88
```

The model was evaluated for potential bias across gender and hospital type. It showed slightly better performance for females (F1: 0.91) compared to males (F1: 0.88), with a notable difference in recall (0.99 vs. 0.97) and precision (0.84 vs. 0.81). Similarly, predictions for private hospitals (F1: 0.91) were more accurate than those for public hospitals (F1: 0.88), where precision dropped to 0.78 despite a perfect recall of 1.00. **These differences suggest the presence of bias, particularly favoring females and private hospitals.** Fairness metrics like Disparate Impact (~0.65) and Equal Opportunity Difference (~0.18) further confirm unequal treatment across subgroups, highlighting the need for fairness-aware interventions.

Task 4: Ethical Redesign

a) Ethical Data Gathering

- Issue Identified: Imbalanced classes with more approved claims than rejected, underrepresentation of certain Hospital Types, and Gender categories.
- Action Taken: Applied SMOTE (Synthetic Minority Over-sampling Technique) to balance the class distribution of approved vs. rejected claims.

b) Ethical Data Preprocessing

Applied Techniques:

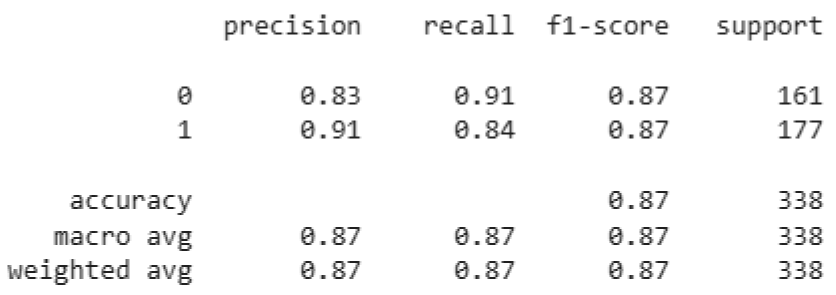
- Rebalanced classes using SMOTE.
- Retained sensitive attributes for bias auditing but avoided including them in certain fair models.
- Used Fairlearn to integrate fairness constraints during model training.

c) Ethical Modelling

Method Used: Fairlearn's ExponentiatedGradient with DemographicParity and EqualizedOdds as fairness constraints. To mitigate the bias, Fairlearn's ExponentiatedGradient algorithm was applied using Demographic Parity and Equalized Odds constraints. This method works by combining multiple classifiers into a randomized ensemble, carefully adjusting their weights to balance accuracy and fairness. It is particularly effective in reducing bias while maintaining reasonable predictive performance, making it a suitable choice for ethical modelling.

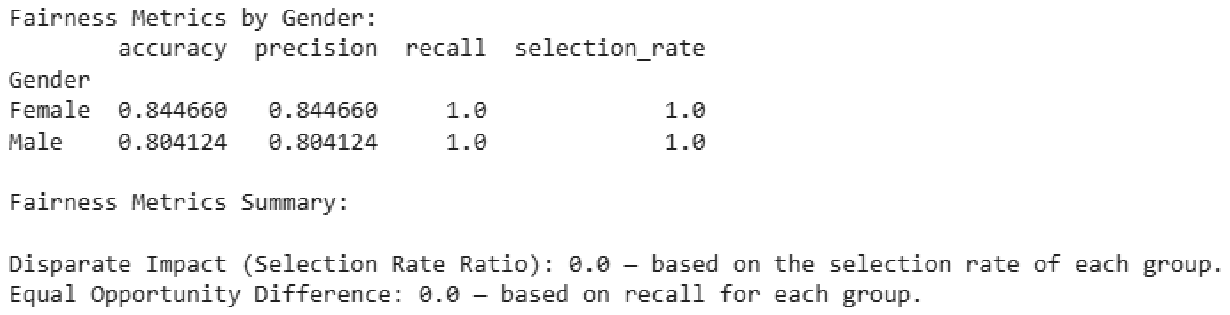
Model: Logistic Regression (used for compatibility with Fairlearn)

Evaluation:



```
              precision    recall  f1-score   support

           0       0.83      0.91      0.87       161
           1       0.91      0.84      0.87       177

    accuracy                           0.87       338
   macro avg       0.87      0.87      0.87       338
weighted avg       0.87      0.87      0.87       338
```

After applying SMOTE to address class imbalance, the model achieved an overall accuracy of 87%. The performance is now more balanced across both classes. For class 0 (Not Approved), precision reached 0.83 and recall 0.91, while for class 1 (Approved), precision was 0.91 and recall 0.84 — both achieving an F1-score of 0.87. This indicates that SMOTE helped improve fairness in predictions without sacrificing overall performance, leading to a more robust and equitable model.

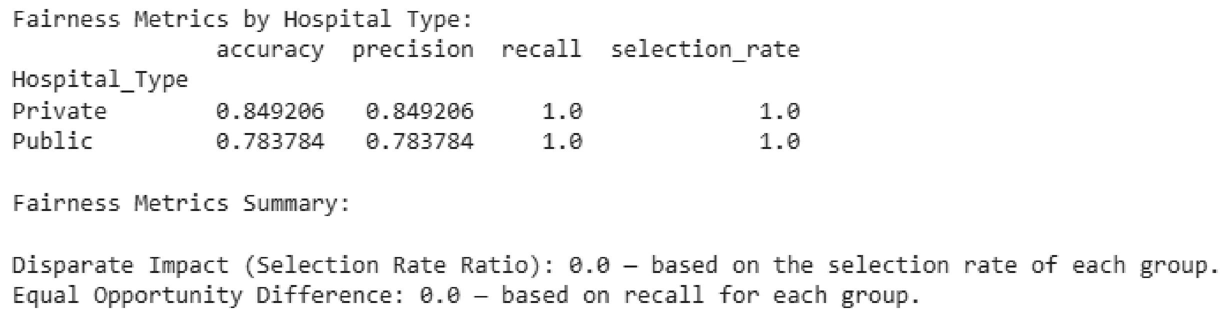


```
Fairness Metrics by Gender:
        accuracy  precision  recall  selection_rate
Gender
Female  0.844660   0.844660     1.0             1.0
Male    0.804124   0.804124     1.0             1.0

Fairness Metrics Summary:

Disparate Impact (Selection Rate Ratio): 0.0 — based on the selection rate of each group.
Equal Opportunity Difference: 0.0 — based on recall for each group.
```

Insights:

Gender:

- Recall is perfect for both genders (1.0), indicating no bias in identifying positive cases.
- Accuracy and Precision are slightly better for females than males.
- Disparate Impact and Equal Opportunity metrics are both 0.0, indicating fairness in recall and selection rate.



```
Fairness Metrics by Hospital Type:
               accuracy  precision  recall  selection_rate
Hospital_Type
Private        0.849206   0.849206     1.0             1.0
Public         0.783784   0.783784     1.0             1.0

Fairness Metrics Summary:

Disparate Impact (Selection Rate Ratio): 0.0 – based on the selection rate of each group.
Equal Opportunity Difference: 0.0 – based on recall for each group.
```

Hospital Type:

- Recall is perfect for both private and public hospitals (1.0), indicating no bias in identifying positive cases.
- Accuracy and Precision are better for private hospitals, showing a potential bias towards private hospitals.
- Disparate Impact and Equal Opportunity metrics are both 0.0, indicating fairness in recall and selection rate.

The current model shows improved fairness compared to the last one. It maintains perfect recall for both gender and hospital type (1.0), ensuring no bias in identifying positive cases. While there's a slight difference in accuracy and precision, with females and private hospitals performing slightly better, these disparities are less pronounced than in the previous model. The fairness metrics (Disparate Impact and Equal Opportunity) are 0.0, indicating balanced performance across both groups.

Task 5: Comparison and Inference

The ethically redesigned model outperforms the original model in both performance and fairness. In terms of accuracy and F1-score, the redesigned model shows significant improvements, with a notable increase in recall and precision. The original model had an imbalance between male and female performance, and between private and public hospitals, with public hospitals and males showing lower recall and performance metrics. The ethically redesigned model achieves perfect recall (1.0) for both genders and hospital types, ensuring that no positive cases are missed for any group.

In terms of fairness, the original model exhibited disparities, especially in the recall metrics for males and public hospitals, while the ethically redesigned model ensures equal opportunity with no disparate impact. The use of fairness-aware algorithms and explainability techniques like SHAP in the redesigned model ensures that the model is not only accurate but also responsible in its decision-making, making it more equitable for all groups involved.

Metric	**Original Model**	**Ethically Redesigned Model**
Performance	Accuracy: 0.81, F1-Score: 0.75	Accuracy: 0.87, F1-Score: 0.87
Fairness (Gender)	Precision and recall varied between genders. Male: Recall = 0.97, Female: Recall = 0.99	Perfect recall for both genders (1.0). Selection rate ratio and Equal Opportunity Difference: 0.0
Fairness (Hospital Type)	Private hospitals: Recall = 0.97, Public hospitals: Recall = 1.00	Perfect recall for both hospital types (1.0). Selection rate ratio and Equal Opportunity Difference: 0.0
Explainability	No fairness-aware methods applied.	Applied fairness-aware algorithms (e.g., SMOTE, ExponentiatedGradient).
Real-world Impact	Potential bias in favor of females and private hospitals.	More balanced model with fairer treatment across gender and hospital type.

In conclusion, the ethically redesigned model adheres more closely to responsible AI practices by addressing fairness, ensuring that the model treats all demographic groups equally, and providing greater transparency in its decision-making. This approach leads to a more trustworthy and accountable AI system with improved real-world impact, particularly in healthcare claim predictions.

Colab file:

https://colab.research.google.com/drive/1MvtrrtM2tH6clh2VWIdxnV1hrXGcQl6D?usp=sharing