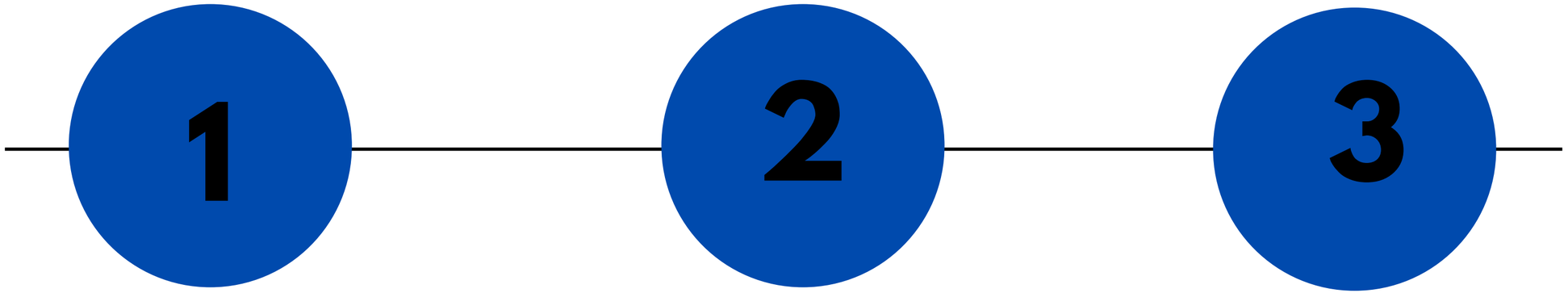
DIABETES
PREDICTION

Project submission

Presented by: Group 6

MACHINE LEARNING TUNING

3 CRITICAL ASPECTS TO SOLVE



MISSING BLOOD PRESSURES

UNBALANCED TARGET VARIABLES

DISEASE PREDICTION

MODEL INTERPRETABILITY AND EMPERICAL TUNING

BLOOD PRESSURE IMPUTATION - FEMALE ONLY

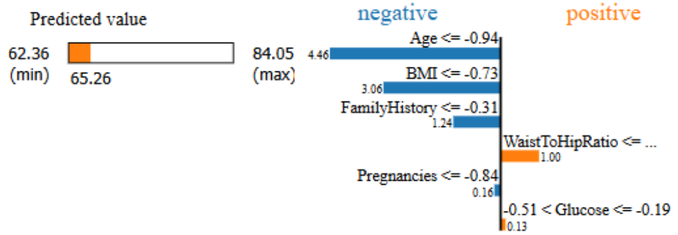
VISUALIZE RELATIONSHIPS

Features and their description

1. Pregnancies - Number of times the person conceived
2. Fasting Glucose - 8 hrs of fasting
3. Age - In years -> 21 and up
4. BMI - weight in kg / Height in meters^2
5. Family History - 0 for no Family history in diabetes, 1 if there is family history
6. Waist to hip ratio = Waist / Hips in cm
7. Bloodpressure - Diastolic blood pressure (mm Hg)
8. Outcome - 0 for not diabetic and 1 if diabetic (Target Variable)

```
BloodPressure        1.000000
Age                  0.281644
BMI                  0.227042
WaistToHipRatio      0.205303
Pregnancies          0.186144
Glucose              0.111212
FamilyHistory        0.071707
Outcome              0.043079
Name: BloodPressure, dtype: float64
```

```
LIME Feature Importance Matrix for B
            Feature  LIME Importance
5               Age         2.464679
2               BMI         1.783704
4     FamilyHistory         1.215964
3   WaistToHipRatio         0.608275
1           Glucose         0.287391
0       Pregnancies         0.191928
```



Feature	Correlation Rank	LIME Importance Rank
Age	1	1
BMI	2	2
WaistToHipRatio	3	4
FamilyHistory	6	3
Glucose	5	5
Pregnancies	4	6

```
feature_sets = {
    "Set 1 (Age,BMI,WaistToHipRatio)": ['Age', 'BMI', 'WaistToHipRatio'],
    "Set 2 (Age Only)": ['Age'],
    "Set 3 (All Features Except Outcome)": ['Age', 'BMI', 'WaistToHipRatio', 'Pregnancies', 'Glucose', 'FamilyHistory']
}
```

```
    # Step 1: Perform 80/20 split
    BP_X = data[features]
    BP_y = data['BloodPressure']
    BP_X_train_80, BP_X_test_20, BP_y_train_80, BP_y_test_20 = train_test_split(BP_X, BP_y, test_size=0.2, random_state=42)

    # Step 2: Perform 60/20 split on 80% training data
    BP_X_train_60, BP_X_val_20, BP_y_train_60, BP_y_val_20 = train_test_split(BP_X_train_80, BP_y_train_80, test_size=0.25, random_state=42)

    # Step 3: Scale data
    scaler = StandardScaler()
    BP_X_train_60_scaled = scaler.fit_transform(BP_X_train_60)
    BP_X_val_20_scaled = scaler.transform(BP_X_val_20)
    BP_X_test_20_scaled = scaler.transform(BP_X_test_20)

    # Step 4: Train model on 60% training data (only rows with non-missing y)
    BP_train_complete = BP_X_train_60_scaled[BP_y_train_60.notna()]
    BP_y_train_complete = BP_y_train_60[BP_y_train_60.notna()]
```

```
Missing values successfully filled!

Chosen Feature Set for Imputation of BloodPressure:
['Age', 'BMI', 'WaistToHipRatio']
```

Feature Set	R-squared (Validation)	MSE (Validation)	R-squared (Test)	MSE (Test)
Set 1 (Age,BMI,WaistToHipRatio)	0.094433	100.519790	0.086290	102.787343
Set 2 (Age Only)	0.073139	102.883468	0.066154	105.052557
Set 3 (All Features Except Outcome)	0.094287	100.535924	0.088351	102.555416

MODEL INTERPRETABILITY AND EMPERICAL TUNING

BLOOD PRESSURE IMPUTATION - ALL GENDER

VISUALIZE RELATIONSHIPS

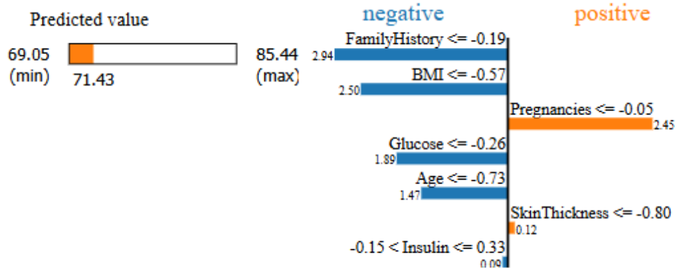
Features and Their Descriptions

- **Pregnancies** - Number of times the person conceived.
- **Glucose** - 2 hours of fasting.
- **BloodPressure** - Diastolic blood pressure (mm Hg).
- **SkinThickness** - Measurement of Triceps Skinfold Thickness (mm):
- **Insulin** - Fasting insulin levels (µU/mL). Normal range: 2–25 µU/mL.
- **BMI** - Body Mass Index (weight in kg / height in meters²):
- **Family History** - 0 for no family history in diabetes, 1 if there is family history.
- **Age** - In years, ranging from 1 to 85.
- **Outcome** - 0 for not diabetic and 1 if diabetic (Target Variable).

```
BloodPressure    1.000000
Outcome          0.335337
BMI              0.283767
Glucose          0.223614
Insulin          0.204414
Age              0.201559
SkinThickness    0.100627
FamilyHistory    0.097086
Pregnancies     -0.022024
Name: BloodPressure, dtype: float64
```

```
LIME Results All Genders:
            Feature  LIME Importance
5  FamilyHistory          2.865276
6    Pregnancies          2.272630
1        Glucose          1.854726
0            BMI          1.396009
3            Age          0.696293
2        Insulin          0.460933
4  SkinThickness          0.063152
```

```
LIME Results All Genders:
```



Feature	Value
FamilyHistory	-0.19
BMI	-1.57
Pregnancies	-0.05
Glucose	-0.26
Age	-1.27
SkinThickness	-1.48
Insulin	0.10

Feature	Correlation Rank	LIME Importance Rank
Age	1	1
BMI	2	2
WaistToHipRatio	3	4
FamilyHistory	6	3
Glucose	5	5
Pregnancies	4	6

```python
feature_sets = {
    "Set 1 (BMI, Glucose, Insulin, Age)": ['BMI', 'Glucose', 'Insulin', 'Age'],
    "Set 2 (BMI Only)": ['BMI'],
    "Set 3 (All Features Except Outcome)": ['Age', 'Pregnancies', 'Glucose', 'SkinThickness', 'Insulin', 'BMI', 'FamilyHistory']
}
```

```
Missing values successfully filled!

Chosen Feature Set for Imputation of BloodPressure:
['Age', 'Pregnancies', 'Glucose', 'SkinThickness', 'Insulin', 'BMI', 'FamilyHistory']
```

Feature Set	R-squared (Validation)	MSE (Validation)	R-squared (Test)	MSE (Test)
Set 1 (BMI, Glucose, Insulin, Age)	0.119334	47.845930	0.124019	48.036539
Set 2 (BMI Only)	0.076599	50.167707	0.080864	50.403078
Set 3 (All Features Except Outcome)	0.125836	47.492698	0.130732	47.668442

```python
# Define parameter grid for tuning
param_grid = {
    'n_estimators': [50, 100, 200],
    'max_depth': [3, 5, 7],
    'learning_rate': [0.01, 0.1, 0.2],
    'subsample': [0.8, 1.0],
    'colsample_bytree': [0.8, 1.0],
    'gamma': [0, 0.1, 0.5],
    'reg_alpha': [0, 0.1, 1],
    'reg_lambda': [1, 2, 5]
}
```

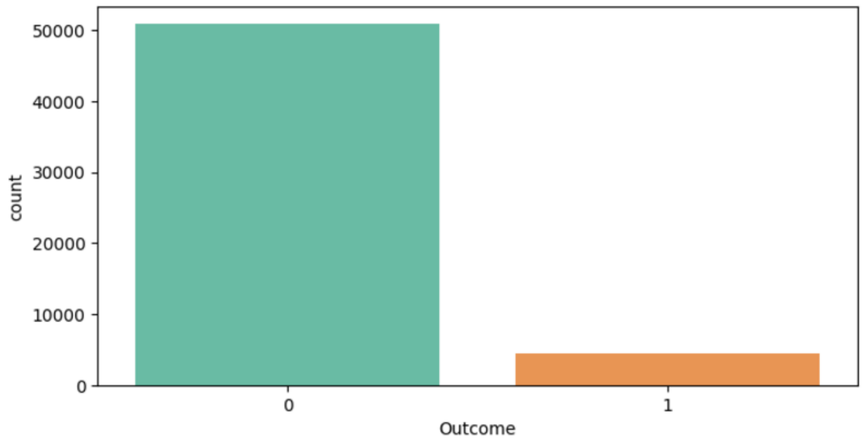
```
Performing hyperparameter tuning...
Fitting 3 folds for each of 2916 candidates, totalling 8748 fits

Best Parameters: {'colsample_bytree': 1.0, 'gamma': 0, 'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 100, 'reg_alpha': 1, 'reg_lambda': 5, 'subs
ample': 0.8}

Test Set Results for Tuned XGBoost Model:
R-squared (Test): 0.2320
MSE (Test): 36.3895
```

ORIGINAL FEATURES - FEMALE ONLY

Features and their description

1. Pregnancies - Number of times the person conceived
2. Fasting Glucose - 8 hrs of fasting
3. Age - In years -> 21 and up
4. BMI - weight in kg / Height in meters^2
5. Family History - 0 for no Family history in diabetes, 1 if there is family history
6. Waist to hip ratio = Waist / Hips in cm
7. Bloodpressure - Diastolic blood pressure (mm Hg)
8. Outcome - 0 for not diabetic and 1 if diabetic (Target Variable)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55299 entries, 0 to 55298
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Pregnancies      55299 non-null  int64  
 1   Glucose          55299 non-null  int64  
 2   BloodPressure    55299 non-null  float64
 3   BMI              55299 non-null  float64
 4   WaistToHipRatio  55299 non-null  float64
 5   FamilyHistory    55299 non-null  int64  
 6   Age              55299 non-null  int64  
 7   Outcome          55299 non-null  int64  
dtypes: float64(3), int64(5)
memory usage: 3.4 MB
```



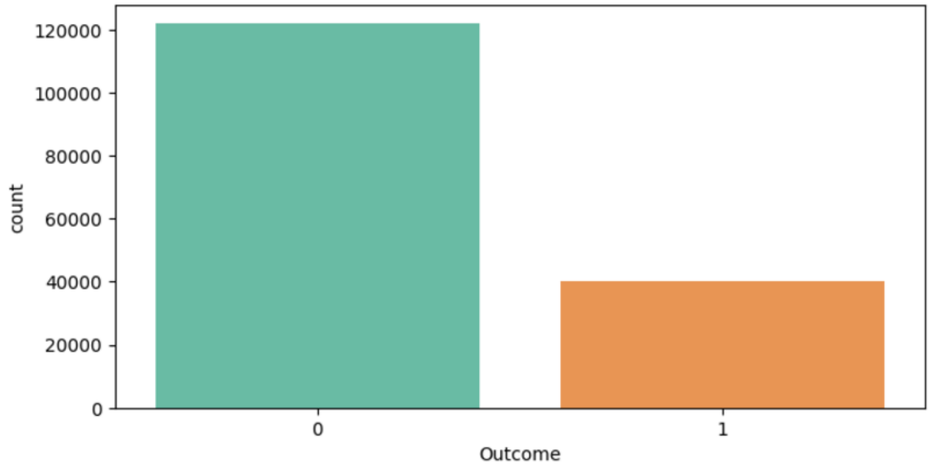

```
Class distribution after undersampling: 0    4463
1    4463
Name: Outcome, dtype: int64

Training Accuracy of Logistic Regression: 95.30812324929971
Accuracy (Test) score of Logistic Regression: 95.07278835386337
Accuracy score of Logistic Regression: 95.07278835386337
AUC: 0.9849910512569402

Training Accuracy of SVM: 95.92436974789916
Accuracy (Test) score of SVM: 95.18477043673013
Accuracy score of SVM: 95.18477043673013
AUC: 0.9837198501721406

Training Accuracy of Random Forest: 99.9859943977591
Accuracy (Test) score of Random Forest: 95.91265397536394
Accuracy score of Random Forest: 95.91265397536394
AUC: 0.9899184419828658
```

ALL GENDER RESULTS

Features and Their Descriptions

- **Pregnancies** - Number of times the person conceived.
- **Glucose** - 2 hours of fasting.
- **BloodPressure** - Diastolic blood pressure (mm Hg).
- **SkinThickness** - Measurement of Triceps Skinfold Thickness (mm):
- **Insulin** - Fasting insulin levels (µU/mL). Normal range: 2–25 µU/mL.
- **BMI** - Body Mass Index (weight in kg / height in meters²):
- **Family History** - 0 for no family history in diabetes, 1 if there is family history.
- **Age** - In years, ranging from 1 to 85.
- **Outcome** - 0 for not diabetic and 1 if diabetic (Target Variable).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 162248 entries, 0 to 162247
Data columns (total 9 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Age             162248 non-null  int64
 1   Pregnancies     162248 non-null  int64
 2   Glucose         162248 non-null  float64
 3   BloodPressure   136239 non-null  float64
 4   SkinThickness   162248 non-null  float64
 5   Insulin         162248 non-null  float64
 6   BMI             162248 non-null  float64
 7   FamilyHistory   162248 non-null  int64
 8   Outcome         162248 non-null  int64
dtypes: float64(5), int64(4)
memory usage: 11.1 MB
```



Class_weight='balanced'

```
Model: Logistic Regression
Training Accuracy: 94.64%
Test Accuracy: 94.51%
Accuracy Score: 94.51%
AUC: 0.98
```

```
Model: Support Vector Machine (SVM)
Training Accuracy: 95.14%
Test Accuracy: 95.00%
Accuracy Score: 95.00%
AUC: 0.98
```

```
Model: Random Forest
Training Accuracy: 100.00%
Test Accuracy: 97.61%
Accuracy Score: 97.61%
AUC: 0.99
```

SUBSEQUENT MODEL BUILDING

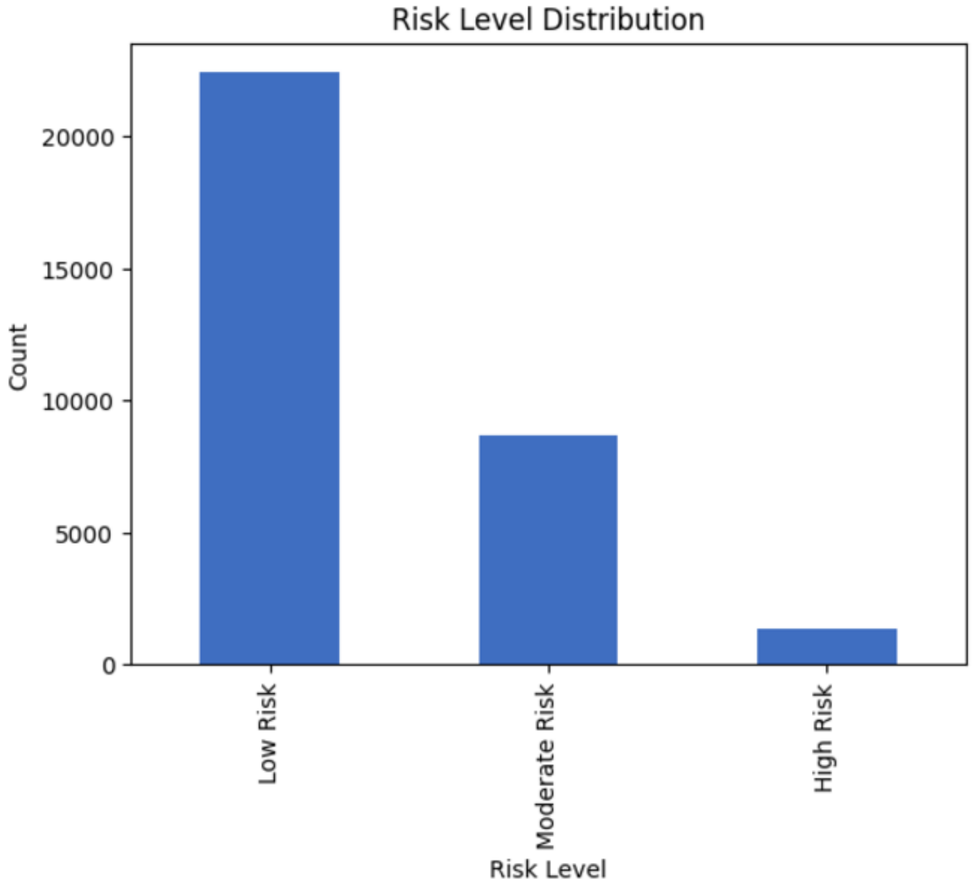
RISK LEVEL CLASSIFICATION:

High Risk:
Glucose > 180 or BloodPressure > 140

Moderate Risk:
BMI > 30

Low Risk:
All other cases

Classification applied to Outcome == 1




```
Model: Random Forest
Training Accuracy: 100.00%
Test Accuracy: 97.61%
Accuracy Score: 97.61%
AUC: 0.99

Confusion Matrix:
[[24143   121]
 [  656  7530]]

Definitions:
True Positives (TP): 7530 - Correctly predicted positive cases.
False Negatives (FN): 656 - Positive cases incorrectly predicted as negative.
False Positives (FP): 121 - Negative cases incorrectly predicted as positive.
True Negatives (TN): 24143 - Correctly predicted negative cases.

Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.98     24264
           1       0.98      0.92      0.95      8186

    accuracy                           0.98     32450
   macro avg       0.98      0.96      0.97     32450
weighted avg       0.98      0.98      0.98     32450
```

GIT REPOSITORY

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights

diabetes-prediction Public

Unwatch 2 | Fork 0 | Star 0

main | 3 Branches | 0 Tags | Go to file t | Add file | Code

sunny-aiden flask — 855c5cf · 4 days ago — 58 Commits

.github/workflows	cicd	4 days ago
app	flask	4 days ago
config	connection	4 days ago
reaserch	connection	4 days ago
.env	setup	4 days ago
.gitignore	setup	4 days ago
Dockerfile	docker	4 days ago
README.md	setup	4 days ago
docker-compose.yml	docker	4 days ago
init_setup.sh	start files	4 days ago
requirement.txt	setup	4 days ago
run.py	start files	4 days ago
test_app.py	start files	4 days ago

README

About

No description, website, or topics provided.

Readme
Activity
0 stars
2 watching
0 forks
Report repository

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 8

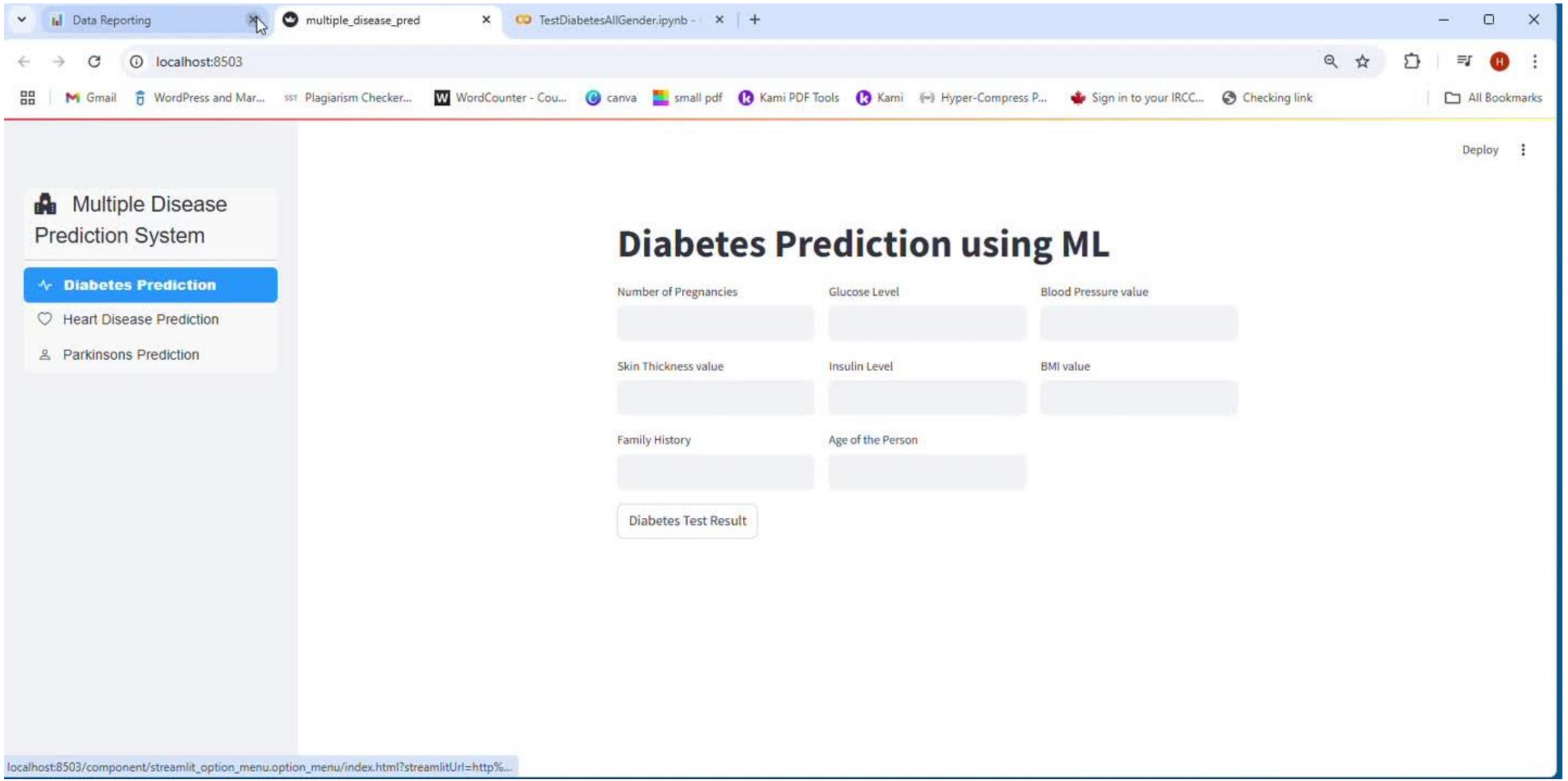
Languages

Jupyter Notebook 49.7% | Python 39.3% | HTML 7.0% | CSS 2.0%

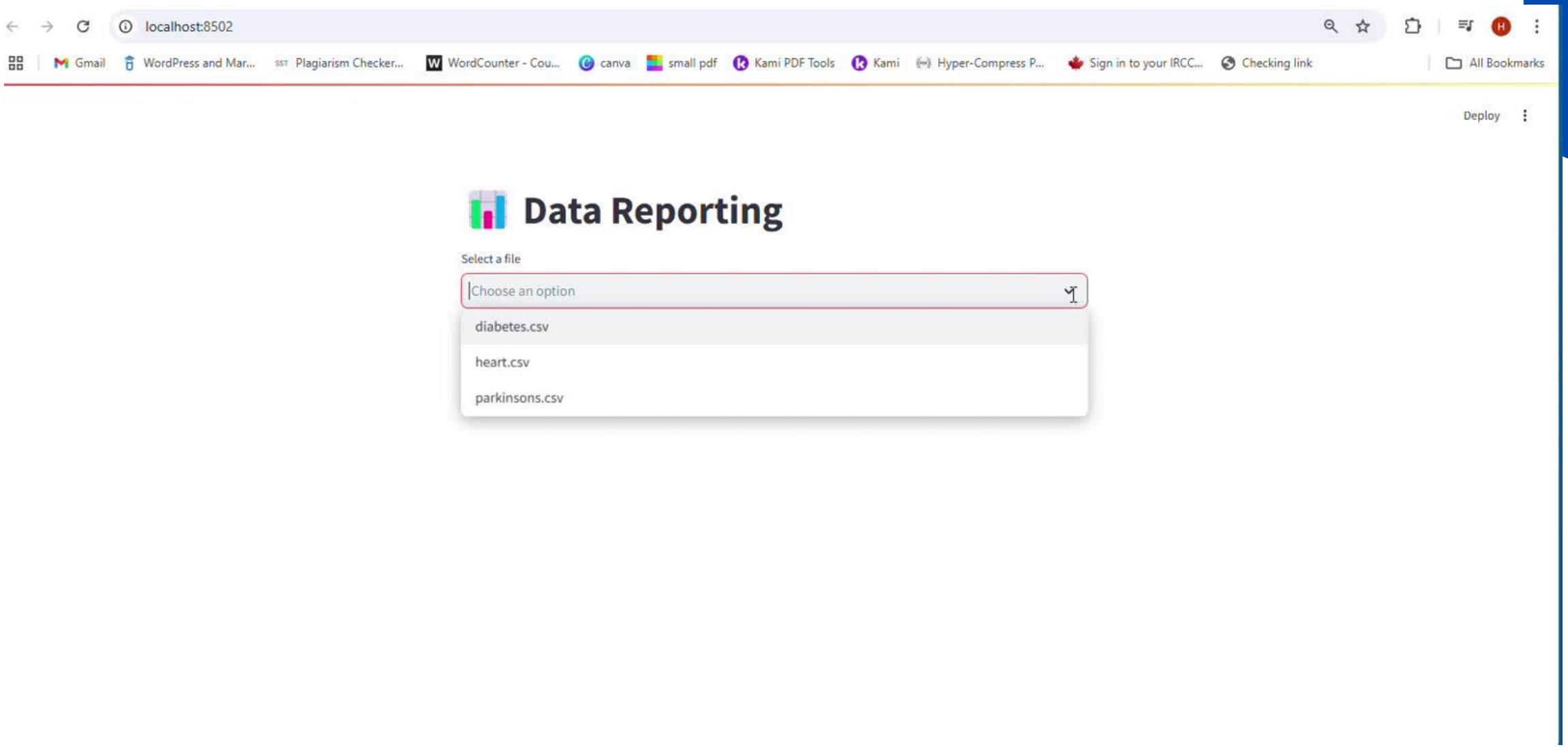
https://github.com/krunalpatel355/diabetes-prediction/graphs/contributors

PITCH

UI APPLICATION



UI APPLICATION - DATA REPORTING



THANK YOU

FROM: GROUP 6