Results

Aligning with the project's focus on ET2D, participants aged 40 or older were excluded from the Pima Indians dataset. In addition, the dataset required preprocessing steps, including handling missing values, removing outliers, checking for multicollinearity, and addressing class imbalance, resulting in a total of 802 data points.

To evaluate which algorithm best predicted early-onset type II diabetes (ET2D), Logistic Regression, Random Forest, and Decision Tree models were compared using five evaluation metrics: accuracy, precision, recall, F-1 Score, and Area Under Curve (AUC). The researcher decided to implement a 5-fold cross-validation system to ensure multiple parts of the dataset are equally trained and tested upon. Additionally, a feature importance analysis was performed to identify the variables that had the greatest impact on ET2D prediction, which was run once.

In the context of this research, accuracy refers to the percentage of diabetics and non-diabetics that the algorithm correctly classified from the total number of datapoints. Precision refers to the rate of correctly identified diabetics out of all those predicted to be diabetic datapoints. Recall refers to the percentage of actual diabetics that the model correctly identified. The AUC measure comes from the ROC graph and reflects how well a model can distinguish between non-diabetics and diabetics. Although this study uses all five evaluation metrics, recall is the primary one used to determine which algorithm performs best. In a healthcare context, especially when detecting early-onset type II diabetes, it's more concerning to miss someone who has diabetes (a false negative) than to incorrectly identify someone as diabetic (a false positive). A high recall score means the model successfully identifies most true diabetic cases, which is essential for early diagnosis and timely treatment.

305	96
100	301

Table 1: **Summed Confusion Matrix for Logistic Regression**. The confusion matrix summarizes the performance of Logistic Regression across the 5 folds through default hyperparameters.

318	83
54	347

Table 2: **Summarized Confusion Matrix for Random Forest**. The confusion matrix summarizes the performance of Random Forest across the 5 folds using default hyperparameters.

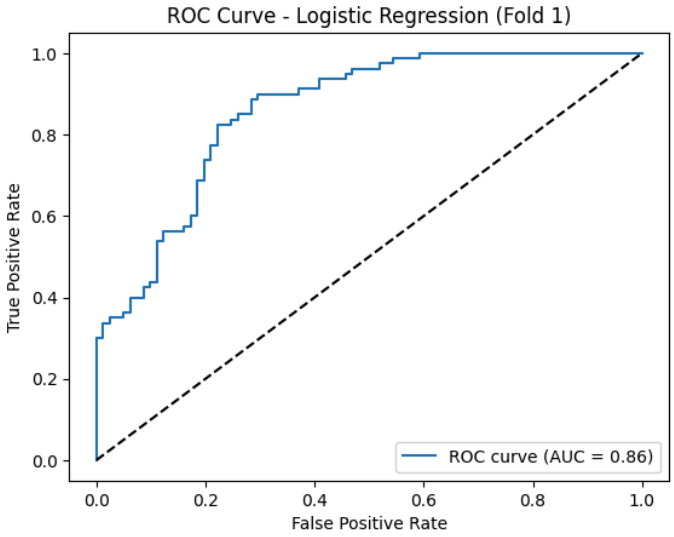
303	98
96	305

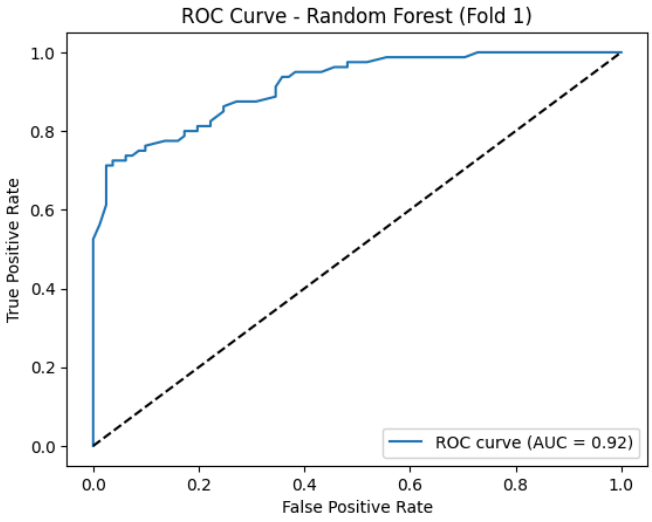
Table 3: **Summed Confusion Matrix for Decision Trees.** The confusion matrix summarizes the performance of Decision Trees across the 5 folds using default hyperparameters.

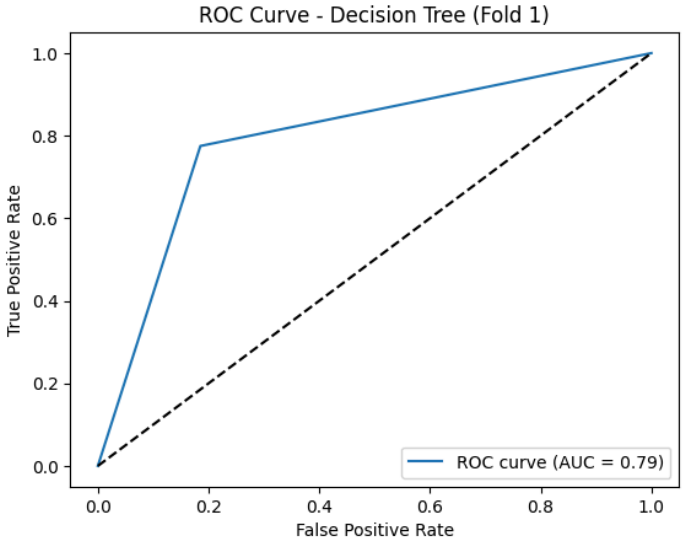
5-fold Validation	**Logistic Regression**	**Random Forest**	**Decision Tree**
Accuracy	0.7556 ± 0.0464	0.8292 ± 0.0302	0.7580 ± 0.0320
Precision	0.7565 ± 0.0463	0.8328 ± 0.0307	0.7589 ± 0.0319
Recall	0.7556 ± 0.0464	0.8292 ± 0.0302	0.7580 ± 0.0320
F-1 Score	0.7554 ± 0.0466	0.8287 ± 0.0304	0.7578 ± 0.0321
AUC	0.8457 ± 0.0257	0.9184 ± 0.0156	0.7579 ± 0.0319

Table 4: **Average Evaluation of Default Parameters of the ML Model according to Scikit Learn API.**

The ROC curves for each algorithm are representative of the first fold in the five-fold cross-validation. From the first fold, it can be seen that Random Forrest achieved the highest AUC (0.93), indicating its ability to distinguish between non-diabetic and diabetic cases. Logistic Regression follows with an AUC of 0.85, indicating strong predictability, with a curve that is mostly located in the left corner. Decision Tree achieved a moderate AUC score of 0.77, indicating a less consistent ability compared to the other models in the study.







Next, hyperparameter tuning was conducted using GridSearch CV for each algorithm. For Logistic Regression, the parameters tuned were the C-values (Inverse regularization strength) and the penalty type. The best parameters for Logistic Regression were revealed to be 'clf__C': 10, 'clf__penalty': 'l2'.

For Random Forrest, the parameters tuned were: param_clf__n_estimators, param_clf__max_depth, and param_clf__min_samples_split. Best Hyperparameters (Random Forest): {'clf__max_depth': None, 'clf__min_samples_split': 2, 'clf__n_estimators': 200}.

For Decision Trees, the parameters tuned were: param_clf__max_depth, param_clf__min_samples_split, param_clf__min_samples_leaf. Best Hyperparameters (Decision Tree): {'clf__max_depth': None, 'clf__min_samples_leaf': 4, 'clf__min_samples_split': 10}.

Using these parameters, the hypertuned models were evaluated on the dataset.

306	95
98	303

Table 5: **Summed Confusion Matrix for Hypertuned Logistic Regression Model**. The confusion matrix summarizes the performance of Logistic Regression across the 5-fold CV hypertuned model.

317	84
50	351

Table 6: **Summed Confusion Matrix for Hypertuned Random Forrest Model**. The confusion matrix summarizes the performance of Random Forrest across the 5-fold CV hypertuned model.

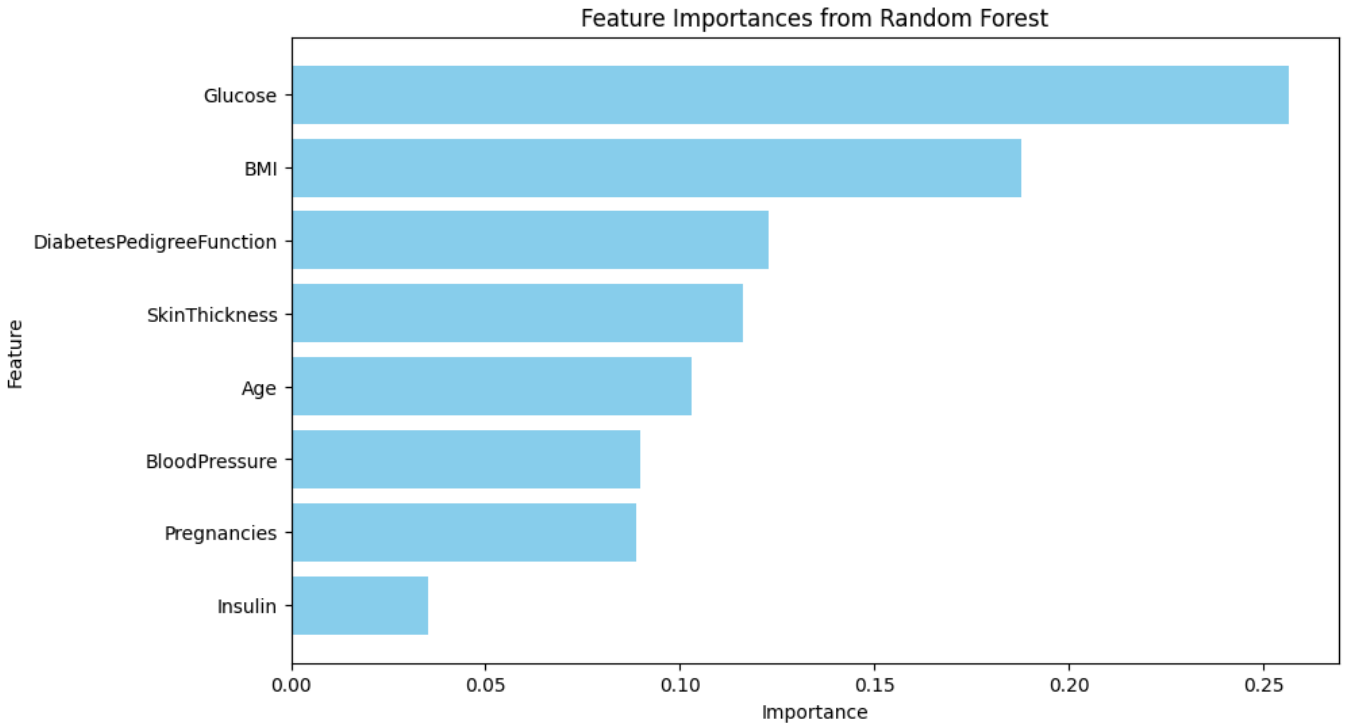
303	98
95	306

Table 7: **Summed Confusion Matrix for Hypertuned Decision Tree Model**. The confusion matrix summarizes the performance of the Decision Tree across the 5-fold CV hypertuned model.

5-fold Validation	**Logistic Regression**	**Random Forest**	**Decision Tree**
Accuracy	0.7594 ± 0.0307	0.8329 ± 0.0265	0.7593 ± 0.0323
Precision	0.7604 ± 0.0310	0.8385 ± 0.0287	0.7601 ± 0.0322
Recall	0.7594 ± 0.0307	0.8329 ± 0.0265	0.7593 ± 0.0323
F-1 Score	0.7591 ± 0.0307	0.8287 ± 0.0304	0.7590 ± 0.0324
AUC	0.8427 ± 0.0213	0.9184 ± 0.0156	0.7592 ± 0.0322

Table 8: **Average Evaluation of Hypertuned Parameters of the ML Model according to the GridSearch CV conducted.**

The Random Forrest Feature Importance:



Feature:	**Score:**
Pregnancies	0.08863
Glucose	0.25644
Blood Pressure	0.08985
SkinThickness	0.11606
Insulin	0.03540
BMI	0.18788
Diabetes Pedigree Function	0.12264
Age	0.10310