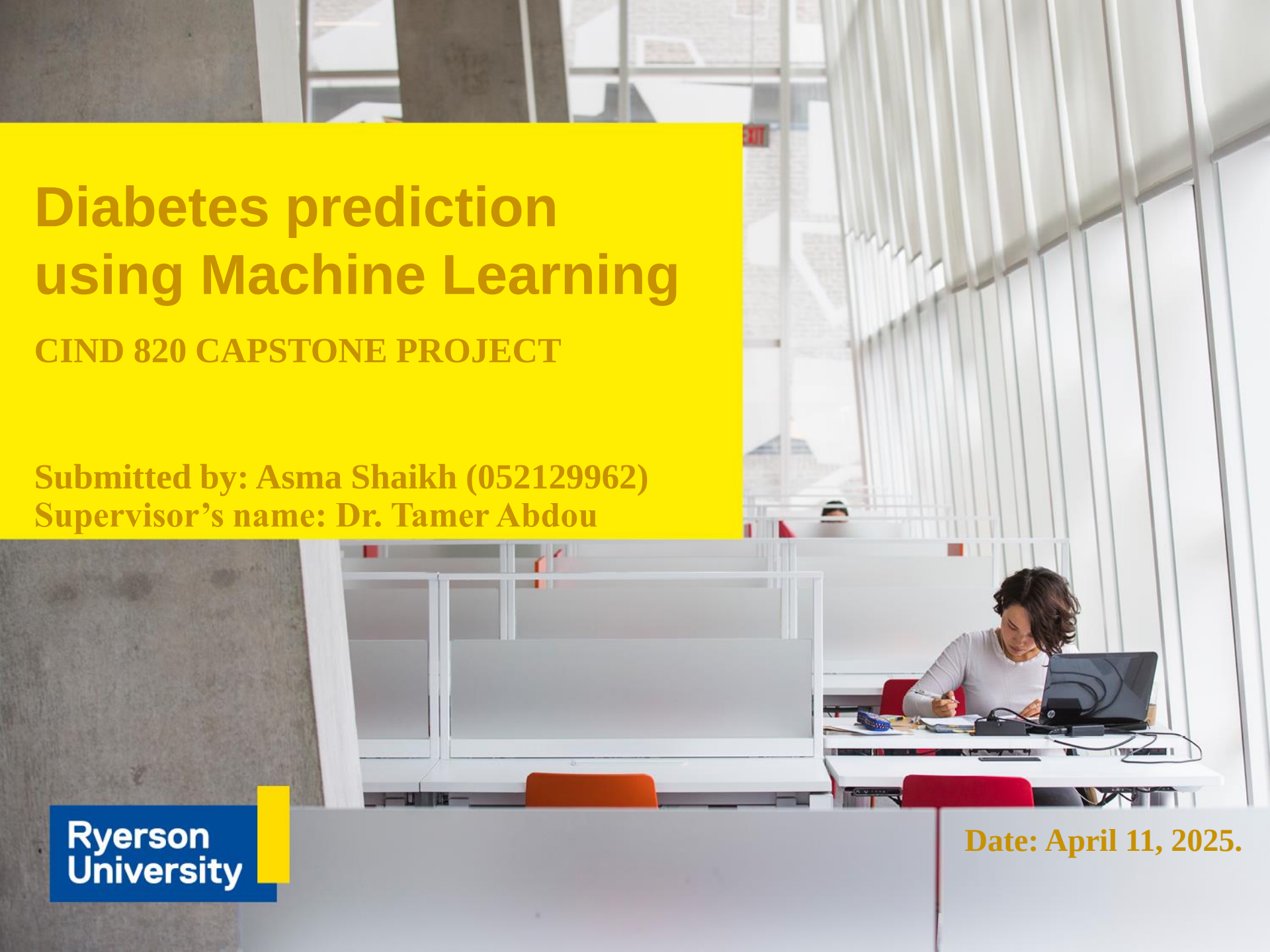


Diabetes prediction using Machine Learning

CIND 820 CAPSTONE PROJECT

Submitted by: Asma Shaikh (052129962)
Supervisor's name: Dr. Tamer Abdou

Ryerson University

Date: April 11, 2025.

BACKGROUND

- As of 2021, ~10.5% of adults aged 20–79 globally have diabetes (IDF).
- Nearly **50% of diabetic individuals are undiagnosed**, living without preventive measures.
- By **2045**, it's projected **1 in 8 adults** will have diabetes.
- Diabetes is associated with severe complications: **blindness, stroke, heart attack, and kidney failure** (WHO, Nov 2024).
- U.S. data (2021–2023) shows a **15.8% diabetes prevalence**, including **4.5% undiagnosed cases** (NHANES).

RESEARCH QUESTIONS

1. **Risk Factors:**
   What features most strongly correlate with diabetes onset?
2. **BMI & Diabetes:**
   Does BMI > 30 increase diabetes risk by over 50%?
3. **Clinical Thresholds:**
   Are glucose levels consistently higher in overweight individuals?
   What is the average HbA1c level among diabetics?
   What % with HbA1c > 6.5% have diabetes?
4. **Model Reliability:** Which ML model yields the best predictive performance for diabetes detection?
5. **Transfer Learning:** How does a top-performing model adapt when applied to another dataset?

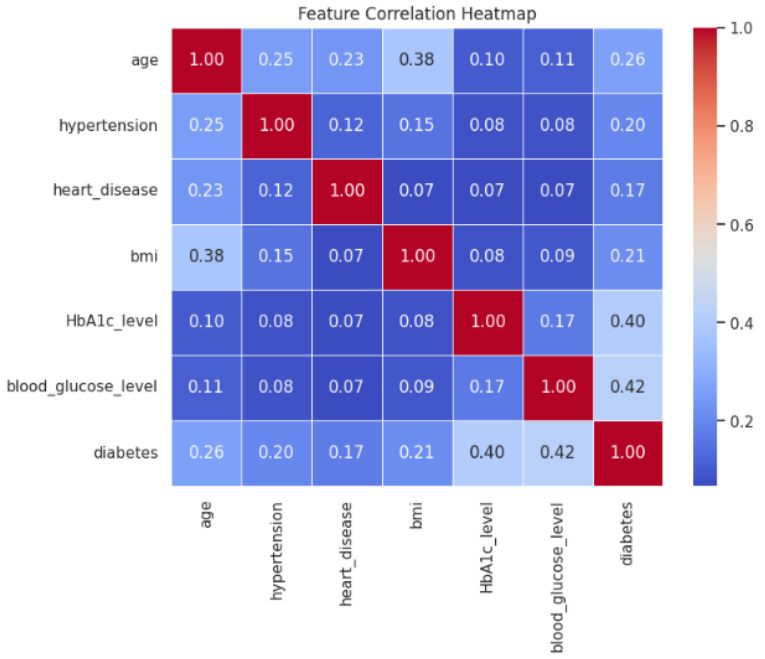
Dataset Processing

Dataset #1 Data Preprocessing

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   gender               100000 non-null  object
 1   age                  100000 non-null  float64
 2   hypertension         100000 non-null  int64
 3   heart_disease        100000 non-null  int64
 4   smoking_history      100000 non-null  object
 5   bmi                  100000 non-null  float64
 6   HbA1c_level          100000 non-null  float64
 7   blood_glucose_level  100000 non-null  int64
 8   diabetes             100000 non-null  int64
dtypes: float64(3), int64(4), object(2)
memory usage: 6.9+ MB

Duplicate Rows Count: 3854

Total Missing Values in Dataset: 0
```

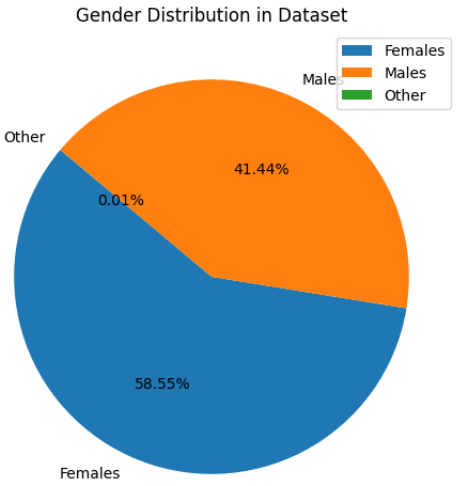


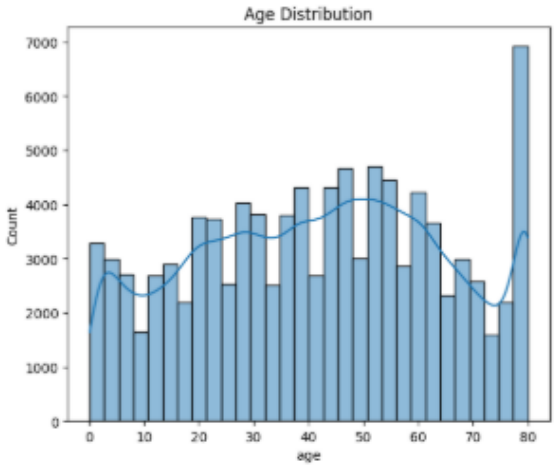
Summary Statistics:

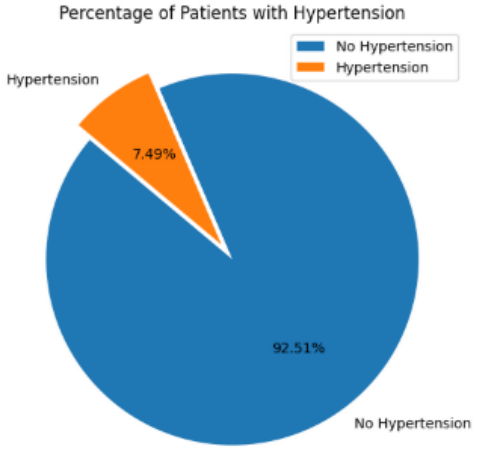
	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
count	100000.000000	100000.00000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000
mean	41.885856	0.07485	0.039420	27.320767	5.527507	138.058060	0.085000
std	22.516840	0.26315	0.194593	6.636783	1.070672	40.708136	0.278883
min	0.080000	0.00000	0.000000	10.010000	3.500000	80.000000	0.000000
25%	24.000000	0.00000	0.000000	23.630000	4.800000	100.000000	0.000000
50%	43.000000	0.00000	0.000000	27.320000	5.800000	140.000000	0.000000
75%	60.000000	0.00000	0.000000	29.580000	6.200000	159.000000	0.000000
max	80.000000	1.00000	1.000000	95.690000	9.000000	300.000000	1.000000

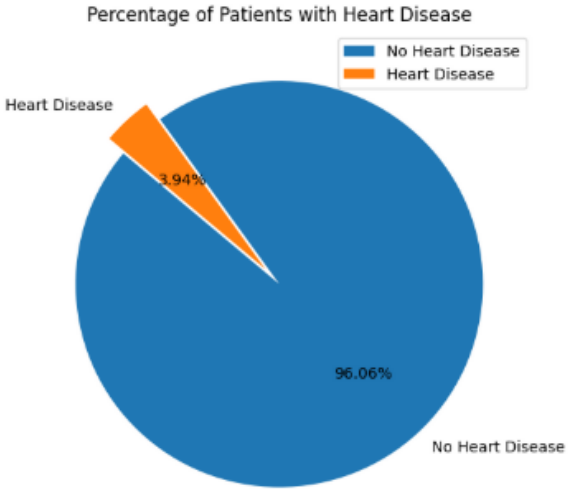
Ryerson University

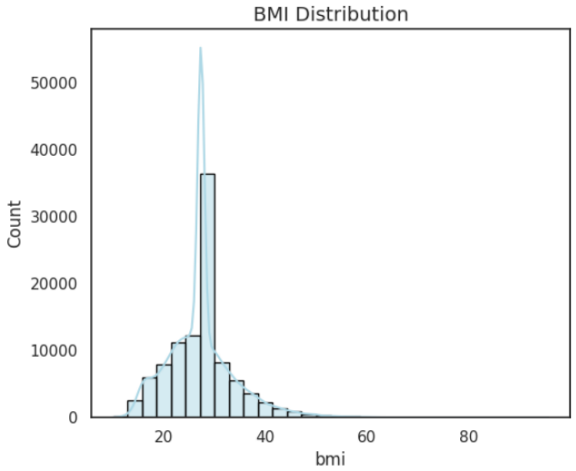
Dataset #1 Data Preprocessing

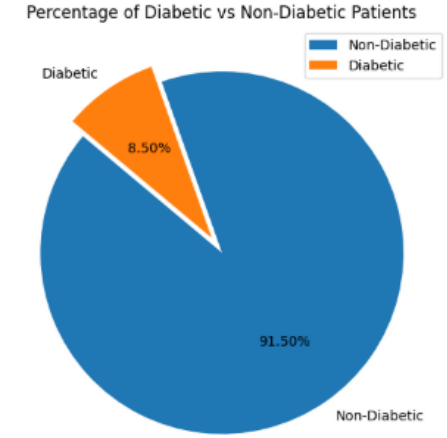






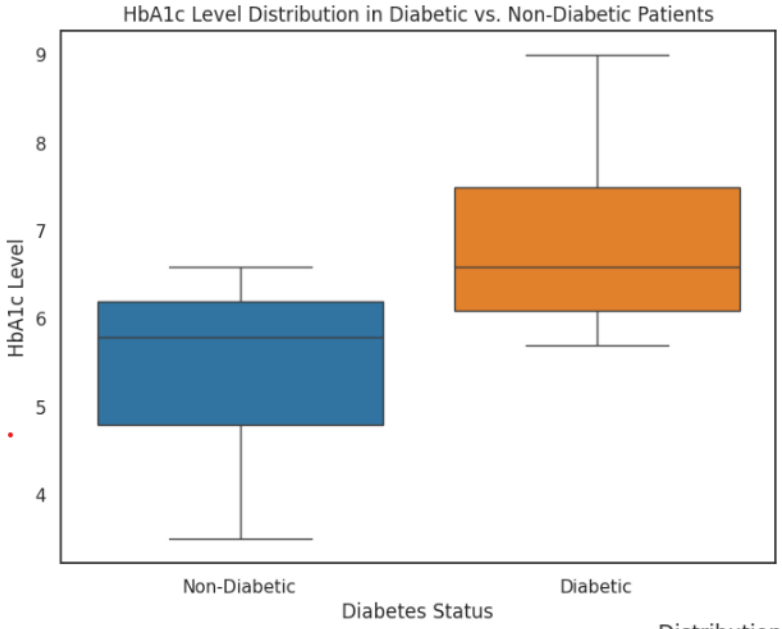


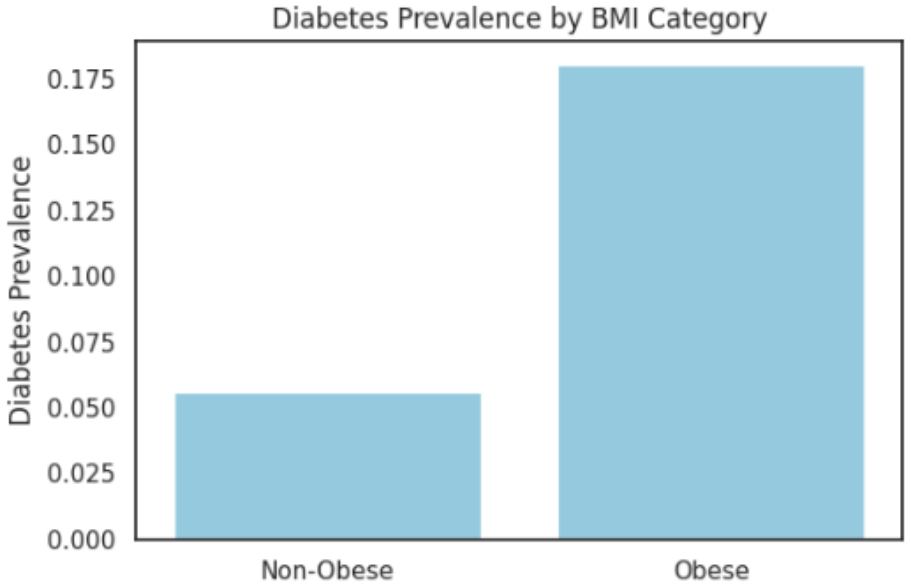


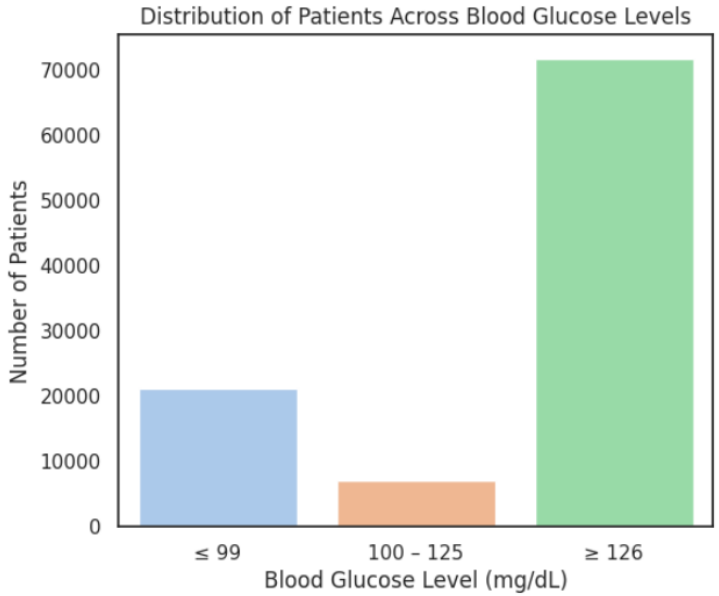


Ryerson University

Dataset #1 Data Preprocessing



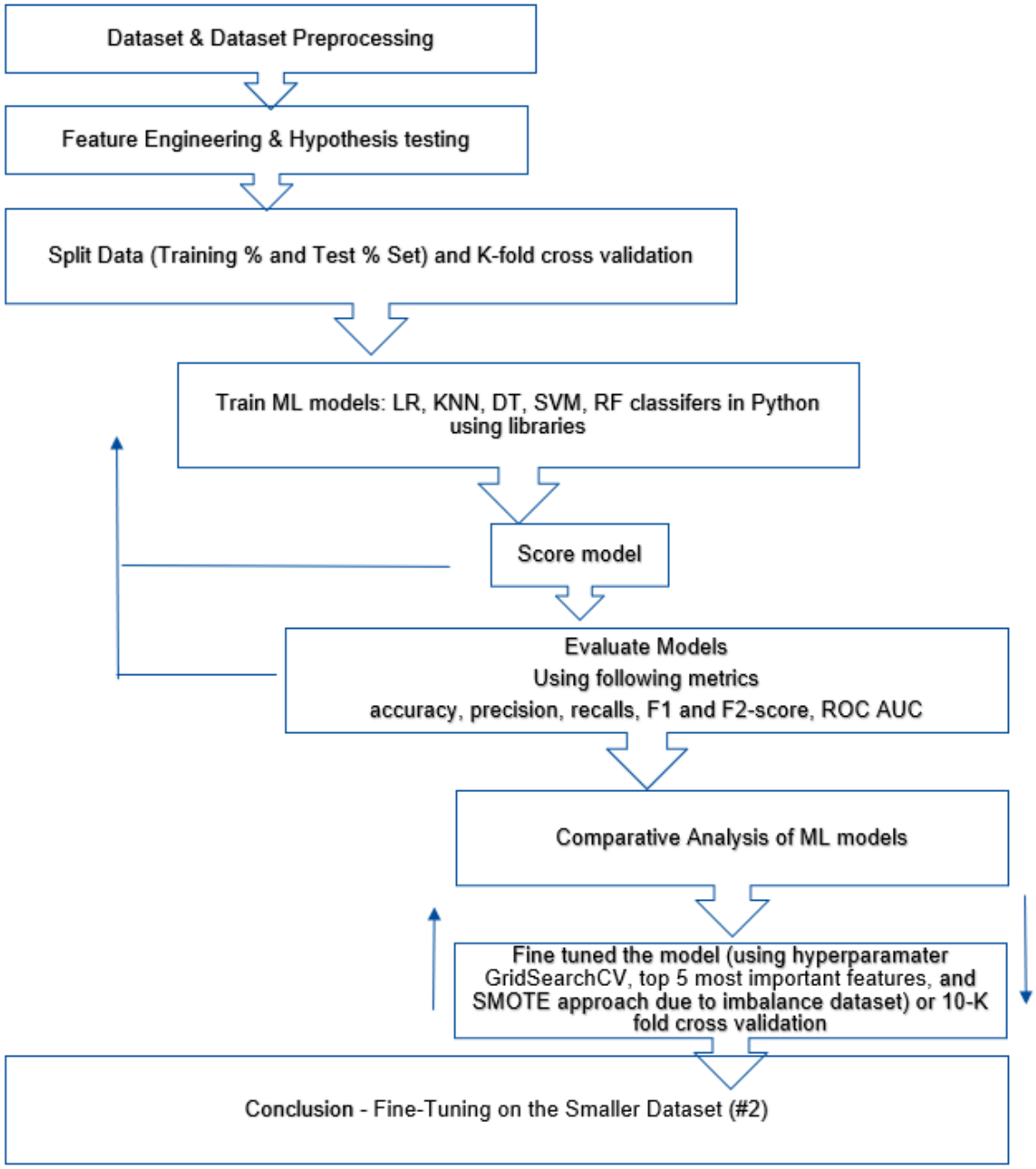




Ryerson University

Modeling

METHODOLOGY



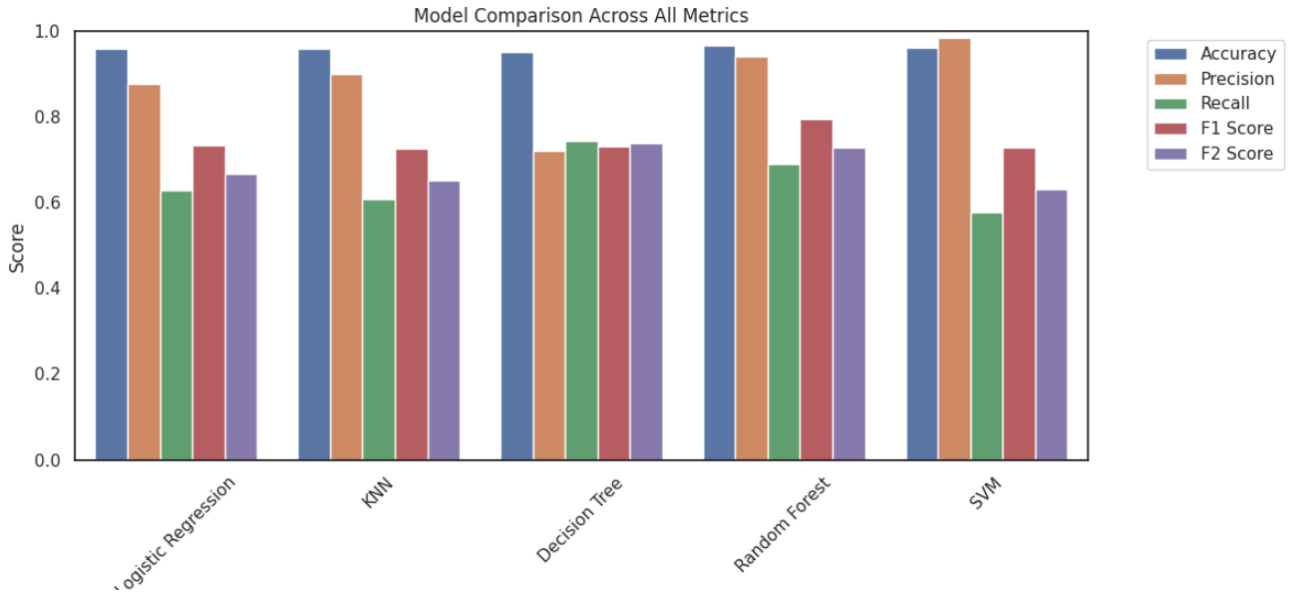
Ryerson University

Comparison of Baseline Models

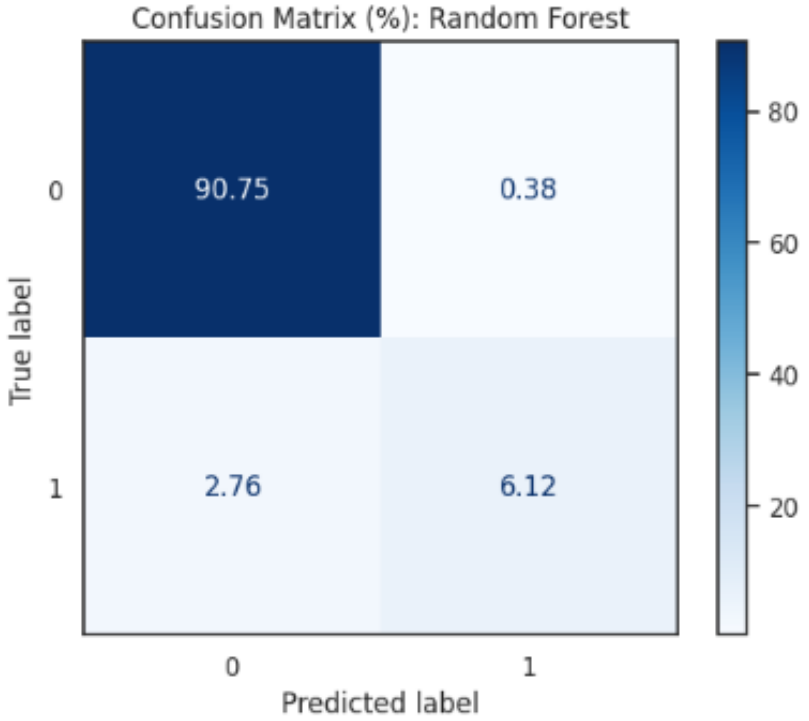
Results Table:

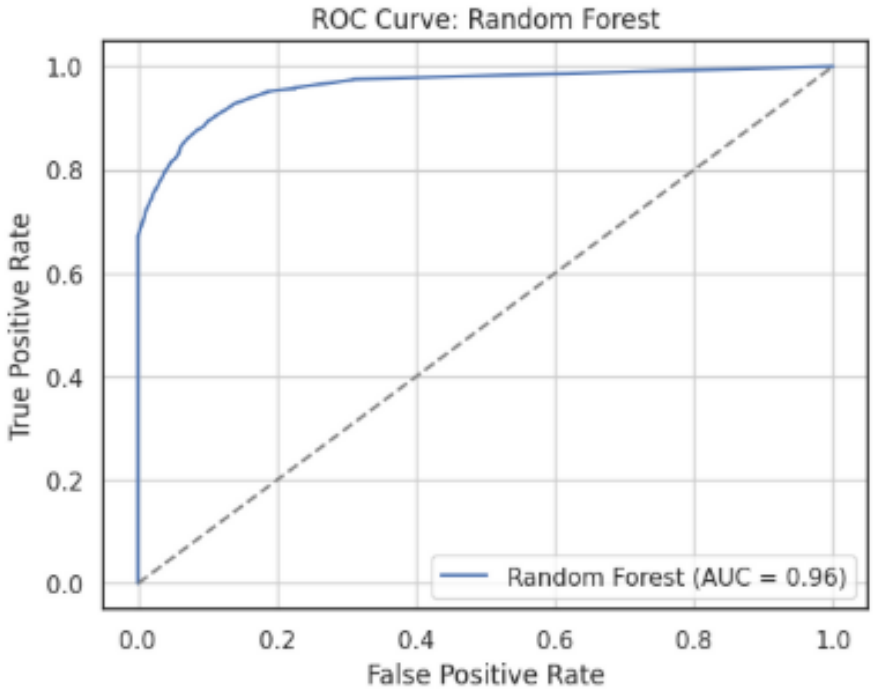
Model	Mean CV Score	Test Score	Accuracy	Precision	Recall	F1 Score	F2 Score
Logistic Regression	0.960805	0.959494	0.959494	0.879026	0.630423	0.734252	0.668219
KNN	0.960318	0.959444	0.959444	0.901667	0.609577	0.727395	0.651807
Decision Tree	0.951941	0.951843	0.951843	0.722101	0.743662	0.732723	0.739247
Random Forest	0.969732	0.968695	0.968695	0.942263	0.689577	0.796357	0.728658
SVM	0.963631	0.961794	0.961794	0.984660	0.578592	0.728886	0.630603

Table-5: Evaluation metrics of the five baseline models.



Baseline RF Models





Top Tuned RF Models - focus on F2 score.

SMOTE + GridSearchCV 5Folds + Top-5 Model (F2 Scoring):

```
Fitting 5 folds for each of 9 candidates, totalling 45 fits

✅ Best Hyperparameters: {'randomforestclassifier__max_depth': 10, 'randomforestclassifier__n_estimators': 200}

📊 Final Model Comparison:
                    Model  Accuracy        F1        F2
0  Top-5 Tuned + SMOTE (F2)  0.905786  0.630298  0.770559
```

SMOTE + GridSearchCV (10 Folds) – Top-5 Features (F2 Scoring):

```
Fitting 10 folds for each of 9 candidates, totalling 90 fits

✅ Best Hyperparameters: {'classifier__max_depth': 5, 'classifier__n_estimators': 50}

📊 Final Model Comparison:
                    Model  Accuracy        F1        F2
0  Top-5 Tuned + SMOTE (F2)  0.914187  0.643983  0.764909
```


Ryerson University

April 11, 2025 | 12

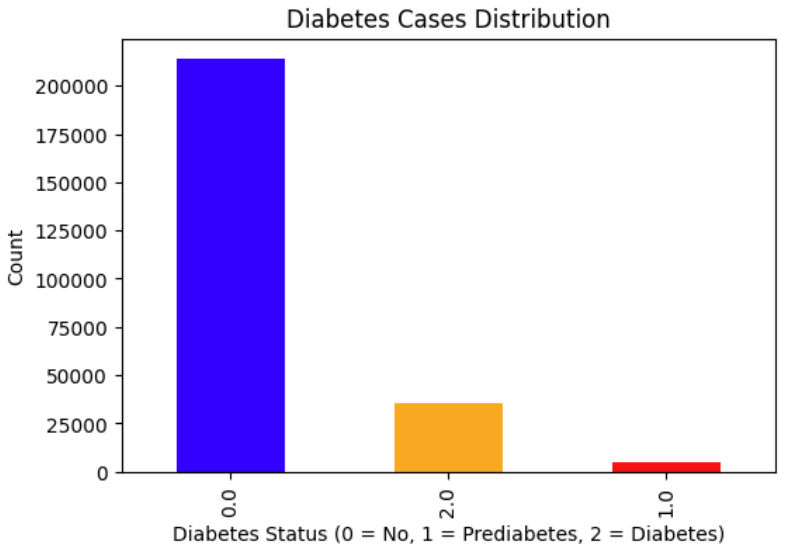
Transfer Learning

Dataset #2 Data Preprocessing

```
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_012          253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
memory usage: 42.6 MB
```

Diabetes_012	count
0.0	213703
2.0	35346
1.0	4631

dtype: int64




```
✅ Duplicates removed.
Remaining rows: 229712
```

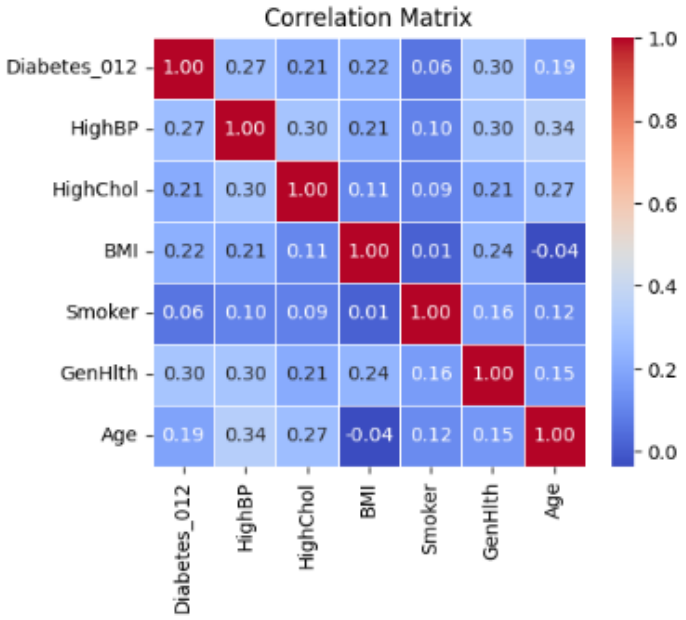
	Feature	Importance
2	BMI	0.367629
4	GenHlth	0.223784
5	Age	0.167143
0	HighBP	0.156109
1	HighChol	0.070791

Ryerson University

Dataset #2 Data Preprocessing

Summary Statistics:

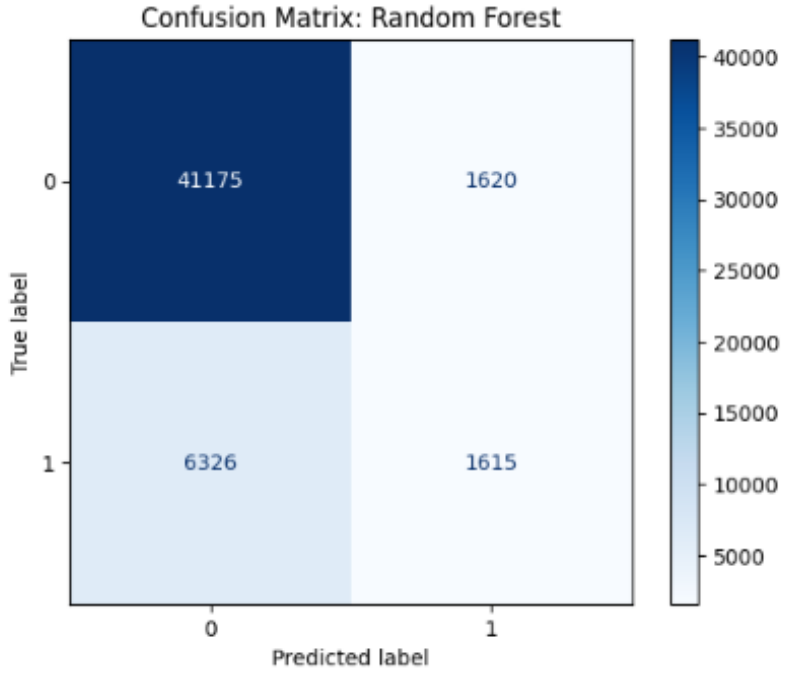
	Diabetes_012	HighBP	HighChol	BMI	Smoker	GenHlth	Age
count	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000
mean	0.139333	0.429001	0.424121	28.382364	0.443169	2.511392	8.032119
std	0.346294	0.494934	0.494210	6.608694	0.496761	1.068477	3.054220
min	0.000000	0.000000	0.000000	12.000000	0.000000	1.000000	1.000000
25%	0.000000	0.000000	0.000000	24.000000	0.000000	2.000000	6.000000
50%	0.000000	0.000000	0.000000	27.000000	0.000000	2.000000	8.000000
75%	0.000000	1.000000	1.000000	31.000000	1.000000	3.000000	10.000000
max	1.000000	1.000000	1.000000	98.000000	1.000000	5.000000	13.000000



Baseline RF Models on Dataset # 2

Model: Random Forest
Mean CV F1 Score: 0.2868

	precision	recall	f1-score	support
0.0	0.87	0.96	0.91	42795
1.0	0.50	0.20	0.29	7941
accuracy			0.84	50736
macro avg	0.68	0.58	0.60	50736
weighted avg	0.81	0.84	0.81	50736



Top Tuned RF Models - on Dataset # 2

SMOTE + GridSearchCV 5Folds + Top-5 Model (F2 Scoring):

```
Fitting 5 folds for each of 9 candidates, totalling 45 fits

✅ Best Hyperparameters: {'randomforestclassifier__max_depth': 10, 'randomforestclassifier__n_estimators': 50}

📊 Final Model Comparison:
                    Model  Accuracy        F1        F2
0  Top-5 Tuned + SMOTE (F2)  0.718878  0.463696  0.611488
```

SMOTE + GridSearchCV (10 Folds) – Top-5 Features (F2 Scoring):

```
Fitting 10 folds for each of 9 candidates, totalling 90 fits

✅ Best Hyperparameters: {'randomforestclassifier__max_depth': 10, 'randomforestclassifier__n_estimators': 100}

📊 Final Model Comparison:
                    Model  Accuracy        F1        F2
0  Top-5 Tuned + SMOTE (F2)  0.718701  0.463136  0.610629
```

CONCLUSION

- Tuned Random Forest (RF) classifier was the best-performing model.
- Used top 5 influential features, SMOTE and GridSearchCV improved class balance and model performance for diabetics prediction.
- F2 Score prioritized due to its relevance in minimizing false negatives in medical screening.
- Model demonstrated strong predictive performance with consistent results across 5-fold and 10-fold cross-validation.

ETHICAL CONSIDERATION & LIMITATIONS

- Both datasets are anonymous and publicly accessible, but demographic coverage is limited.
- Modeling may be biased due to overrepresentation of elderly individuals (e.g., peak at age 80 in Dataset 1).
- Potential bias exists due to missing race/ethnicity data, limiting fairness analysis.
- Only top 5 features were used - remaining features (including smoking history) excluded due to missingness.
- Further validation is required before real-world deployment in diverse populations.
- No external validation using real-time or clinical datasets was conducted.

Thank you!