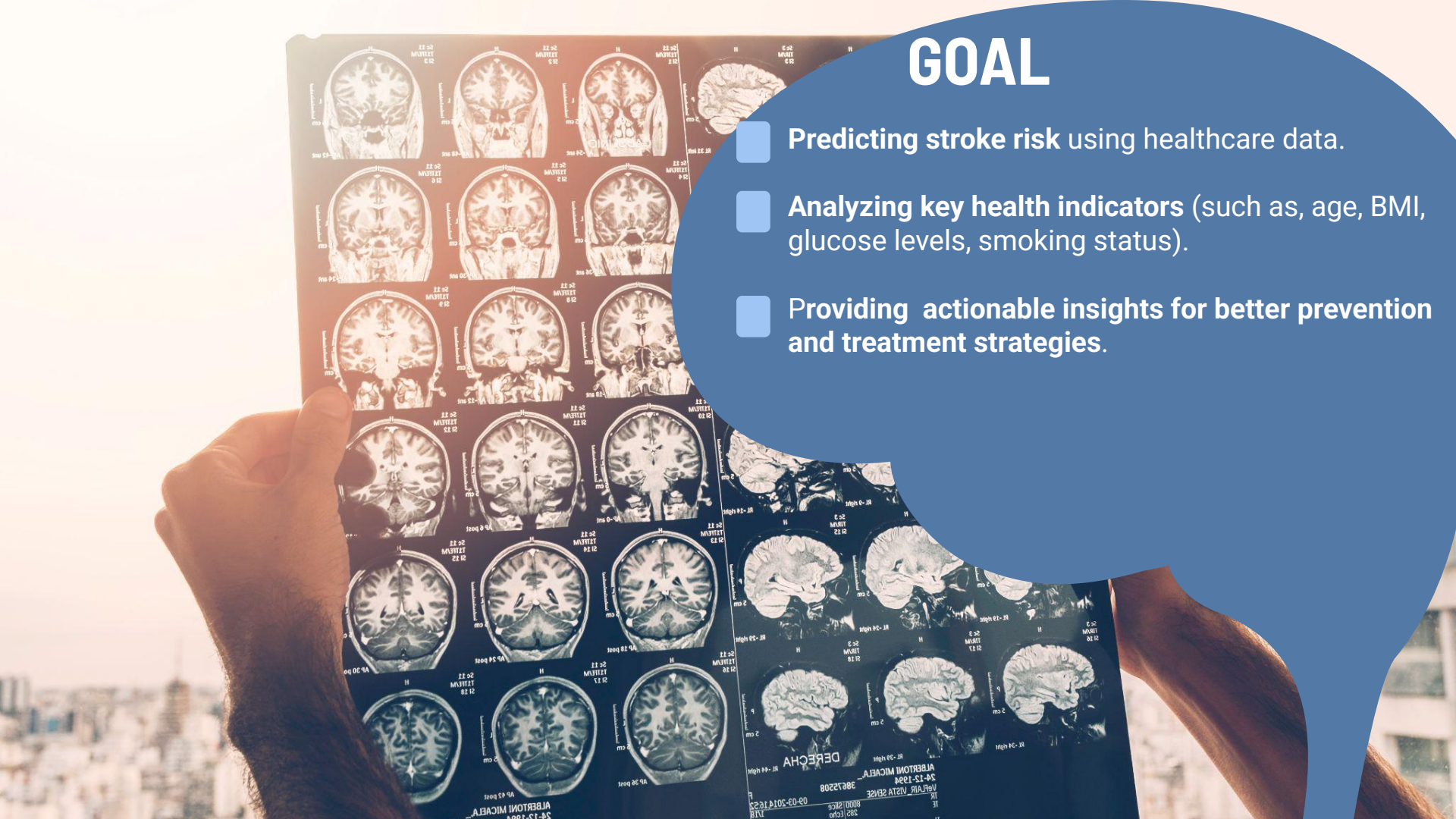
STROKE RISK FACTORS

Fundamentals of Data Science in Business and Engineering Project

Merve Pakcan Tufenk

TABLE OF CONTENTS



GOAL

- **Predicting stroke risk** using healthcare data.
- **Analyzing key health indicators** (such as, age, BMI, glucose levels, smoking status).
- P**roviding actionable insights for better prevention and treatment strategies**.

QUESTIONS

KEY RESEARCH QUESTIONS

1. Which factors have the greatest impact on stroke risk?
2. What are the differences in stroke risk and contributing factors across demographic groups (age,gender, residence type, work type)?
3. How do health metrics (e.g., BMI, glucose levels) correlate with stroke risk, and are there thresholds for increased risk?

5110 ROWS, 12 COLUMNS

4908 ROWS

DATA

Stroke Prediction Dataset from Kaggle

Data source: A clean subset of the original Electronic Health Record (EHR) dataset managed by McKinsey & Company

9 categorical variables, 3 numeric variables

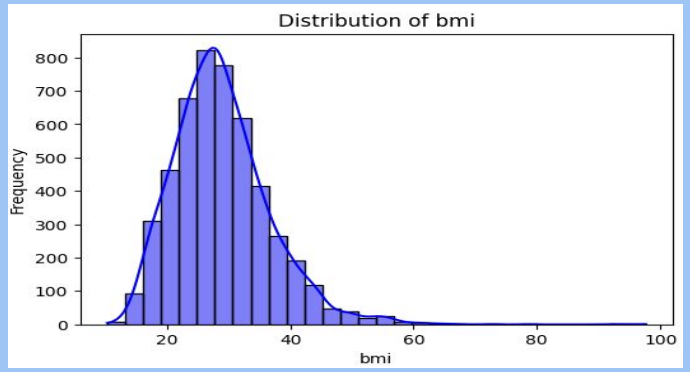
id	Unique identifier for each individual
gender	Gender of the individual (Male/Female/Other)
age	Age of the individual in years.
hypertension	Whether the individual has hypertension (0 = No, 1 = Yes)
heart_disease	Whether the individual has heart disease (0 = No, 1 = Yes)
ever_married	Whether the individual has ever been married (Yes/No)
work_type	Type of employment (Private, Self-employed, Government Job, Never Worked)
residence_type	Type of residence (Urban/Rural)
avg_glucose_level	Average glucose level in the individual's blood
bmi	Body Mass Index (weight-to-height ratio)
smoking_status	Smoking habits (formerly smoked, never smoked, smokes)
stroke	Outcome variable indicating whether the individual experienced a stroke (0 = No, 1 = Yes)

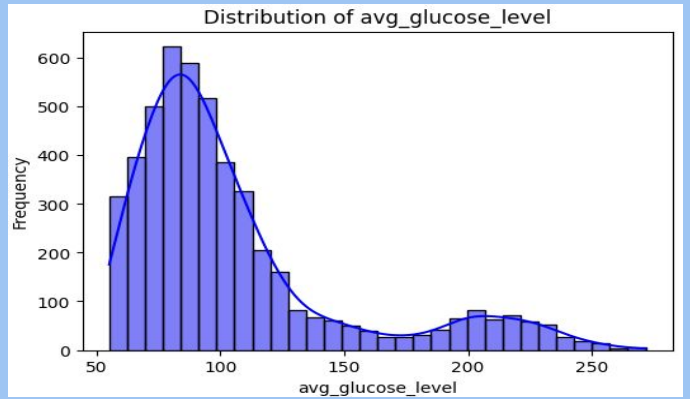
EXPLORATORY DATA ANALYSIS

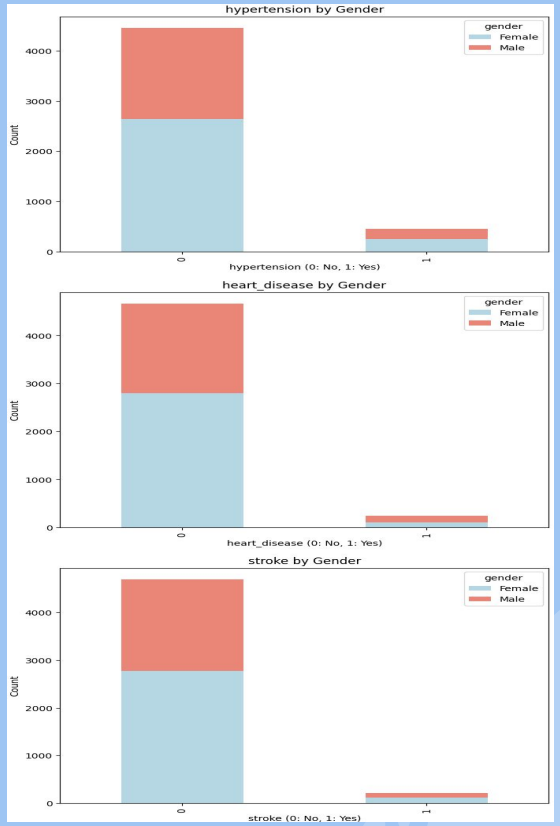
```
Column: age
Mean: 42.87
Median: 44.00
Standard Deviation: 22.56

Column: avg_glucose_level
Mean: 105.30
Median: 91.68
Standard Deviation: 44.43

Column: bmi
Mean: 28.89
Median: 28.10
Standard Deviation: 7.85
```

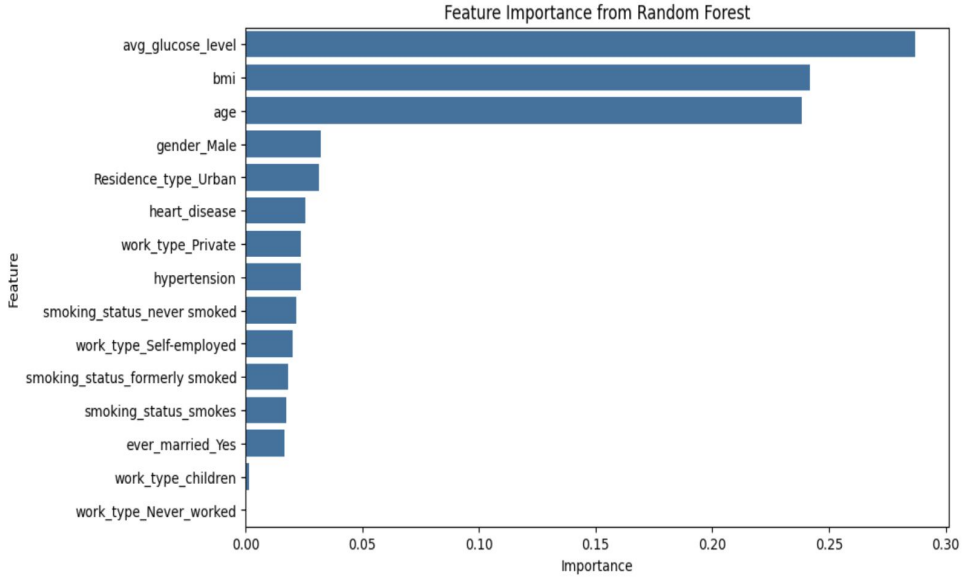






DATA ANALYSIS

Q1 **Which factors have the greatest impact on stroke risk?**

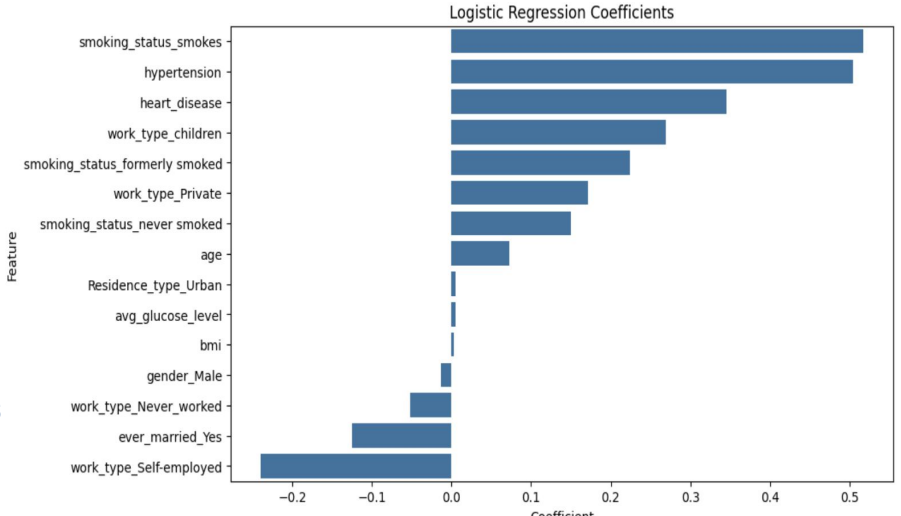


Which Methods Were Used and WHY

- **Random Forest:** Identifies complex and nonlinear relationships.
- Effective for large datasets, provides variable importance rankings.
- **Logistic Regression:** Delivers interpretable results, clearly shows the contribution of variables.
- Suitable for binary classification problems.

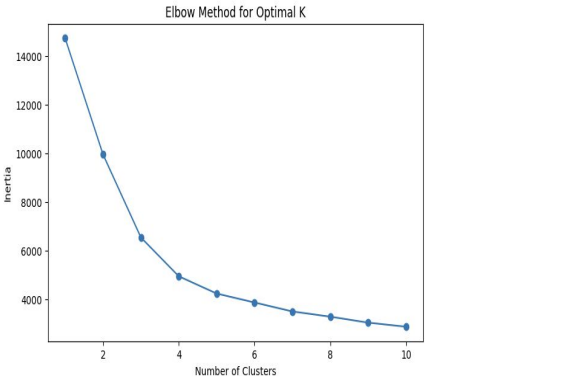
Random Forest model: The most critical predictors were identified as **average glucose level, BMI, and age**

Logistic regression model: Positive coefficients indicate higher stroke risk, with **smoking** increasing risk significantly



Differences from approach: Logistic Regression assumes linearity, Random Forest analyses non-linear interactions

ADDITIONAL STATISTICS



Which Methods Were Used and WHY

K-Means clustering and PCA grouped individuals into low, moderate, and high-risk profiles based on similar health characteristics, focusing on **age, glucose level, and BMI**

- **K-mean clustering:Clusters individuals to uncover patterns and classify stroke risk levels.**
- **Elbow method: To determine the optimal number of clusters, which was identified as 3.**
- **PCA:Simplifies high-dimensional data into two components, enabling clear visualization of clusters.**

Cluster Profiles:

Cluster	age	avg_glucose_level	bmi
0	18.847053	92.719731	22.712373
1	54.021522	89.119093	31.766833
2	60.513428	205.042117	33.427172

Cluster Patterns:

Cluster	age mean	age std	age min	age max	avg_glucose_level mean	avg_glucose_level std
0	18.847053	12.218730	0.08	55.0	92.719731	23.949850
1	54.021522	15.454623	13.00	82.0	89.119093	19.884184
2	60.513428	14.862350	12.00	82.0	205.042117	26.794301

Cluster	bmi min	bmi max	bmi mean	bmi std	bmi min	bmi max
0	55.12	206.25	22.712373	4.485052	10.3	37.3
1	55.22	152.56	31.766833	7.173703	11.3	97.6
2	143.15	271.74	33.427172	7.632394	14.2	71.9

K-Means Clustering Results:

- **Cluster 0 (Purple):** Low-risk, younger individuals (mean age: 18.8 years), with low glucose levels (92.7 mg/dL) and BMI (22.7).
- **Cluster 1 (Teal):** Moderate-risk, middle-aged individuals (mean age: 54 years), with moderate glucose levels (89.1 mg/dL) and higher BMI (31.8).
- **Cluster 2 (Yellow):** High-risk, older individuals (mean age: 60.5 years), with significantly elevated glucose levels (205 mg/dL) and BMI (33.4).

Q2

What are the differences in stroke risk and contributing factors across demographic groups (age,gender, residence type, work type)?

```
One-way ANOVA result for age: F-statistic = 279.87841499632003, p-value = 3.8408903844855186e-61
Chi-Square Test for Gender and Stroke: Chi2 = 0.16955129804441268, p-value = 0.6805108914997836
```

Which Methods Were Used and WHY

ANOVA:One-Way ANOVA is ideal for comparing the means of two or more groups.

Chi-Square tests:Ideal for testing independence between categorical variables

ONE-WAY ANOVA TEST

- Null Hypothesis (H_0): There is no significant difference in the mean age between the stroke and non-stroke groups.
- Alternative Hypothesis (H_1): There is a significant difference in the mean age between the stroke and non-stroke groups.

RESULTS: $p < 0.05$, rejecting the null hypothesis, indicating that age is significantly associated with stroke risk

CHI-SQUARE TEST

- Null Hypothesis (H_0): There is no relationship between gender and stroke risk (they are independent).
- Alternative Hypothesis (H_1): There is a relationship between gender and stroke risk.

RESULTS: $p > 0.05$, failed to reject the null hypothesis, suggesting that gender does not have a significant effect on stroke risk.

Work type and residence type had minimal influence on stroke likelihood.

Q3

How do health metrics (BMI, glucose levels) correlate with stroke risk, and are there thresholds for increased risk?

Which Methods Were Used and WHY

T-TEST: T-Test is ideal for comparing the means of two independent groups.

```
T-Test for Average Glucose Level:
T-statistic: 9.830215360205345, P-value: 1.3476353968167712e-22

T-Test for BMI:
T-statistic: 2.968365485973203, P-value: 0.003008355955526417
The difference in average glucose levels between stroke and non-stroke groups is statistically significant.
The difference in BMI between stroke and non-stroke groups is statistically significant.
```

T-TEST For Average Glucose Level

- **Null Hypothesis (H_0):** There is no significant difference in average glucose levels between stroke and non-stroke groups.
- **Alternative Hypothesis (H_1):** There is a significant difference in average glucose levels between stroke and non-stroke groups.

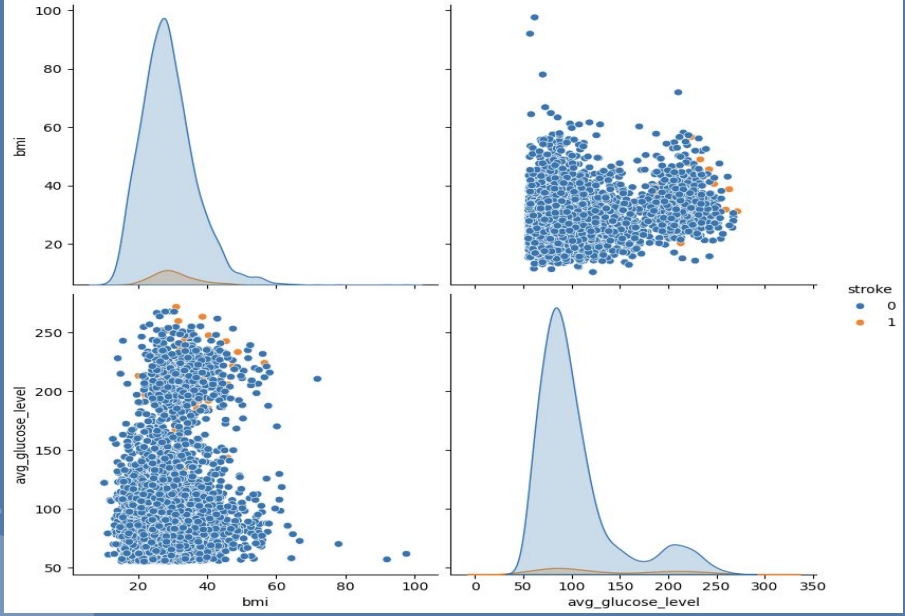
RESULTS: $p < 0.05$, rejecting the null hypothesis, indicating that there is a statistically significant difference in the average glucose levels between stroke and non-stroke groups.

T-TEST For BMI

- **Null Hypothesis (H_0):** There is no significant difference in BMI between stroke and non-stroke groups.
- **Alternative Hypothesis (H_1):** There is a significant difference in BMI between stroke and non-stroke groups.

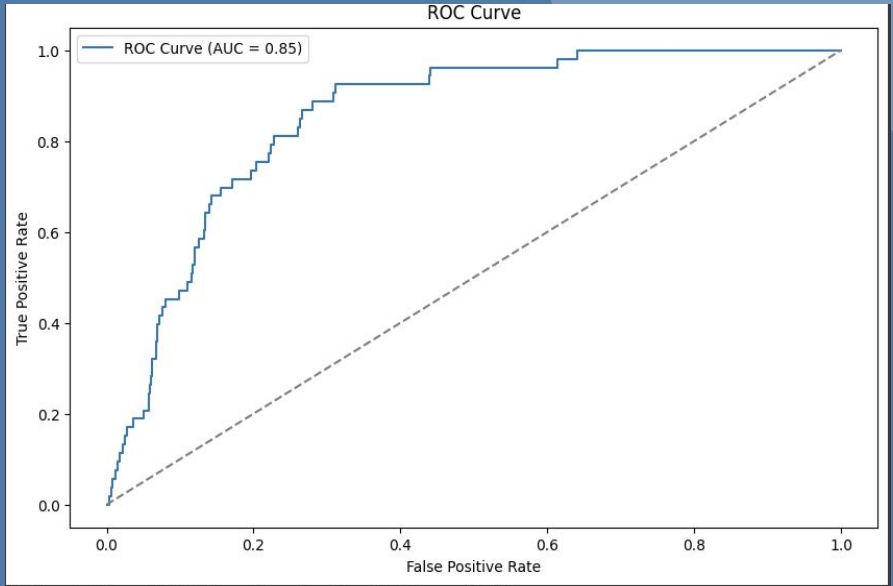
RESULTS: $p < 0.05$, rejecting the null hypothesis, BMI values are significantly different between stroke and non-stroke groups.

PAIRPLOT: suitable for visualizing relationships between two or more variables



Stroke cases are linked to higher BMI (30-40 range) and glucose levels (over 150 mg/dL), highlighting obesity and elevated glucose as potential risk factors.

ROC CURVE ANALYSIS: To evaluate the model's ability to distinguish between stroke and non-stroke cases.



RESULT: Achieved a high AUC value of **0.85**, indicating strong predictive performance.

Optimal Threshold: Determined as **0.03**, balancing true positive and false positive rates for optimized classification.

CONCLUSION

Key Findings

- **Average glucose level, BMI, and age are the strongest predictors of stroke.**
- **Age is critical, while gender and work type have minimal impact.**
- **Machine learning models achieved high accuracy (AUC = 0.85).**
- **K-Means clustering identified high-risk groups for targeted interventions.**

THANKS!