Unveiling the Predictive Power of Big Data: Predicting Smoking and Drinking Behavior

IST 718 – Group 7

Vedant Patil
Shrish Vaidya
Sahil Wani
Chintan Patel

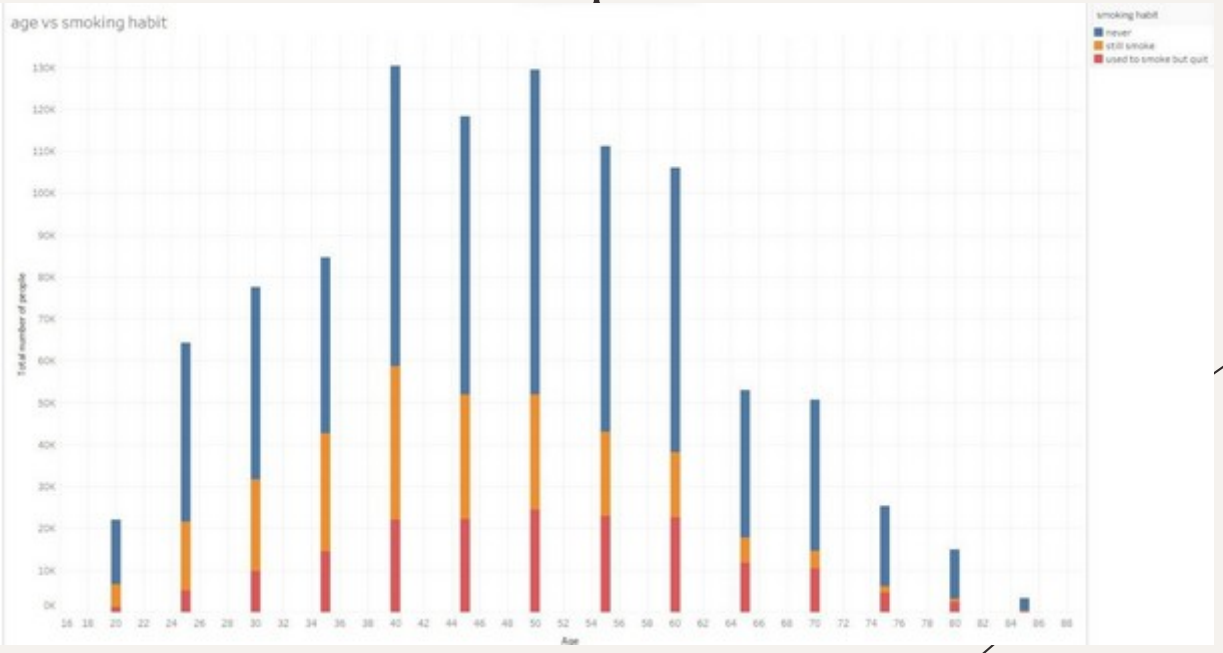
Introduction

- Investigate health metrics: Deep insights into lifestyle patterns.
- Comprehensive dataset: Essential for understanding health dynamics.
- Explore impact: Correlate lifestyle choices with overall well-being.
- Predictive resource: Informing outcomes in alcohol and tobacco research.
- Support evidence-based decisions: Vital for health professionals and policymakers.

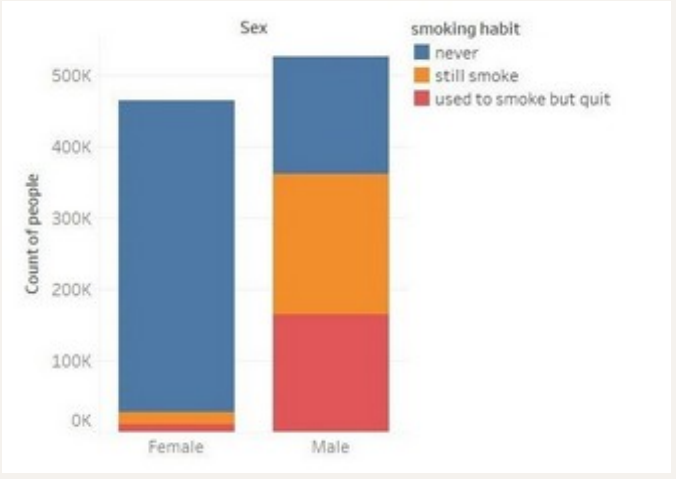
Predictive Modeling Goals

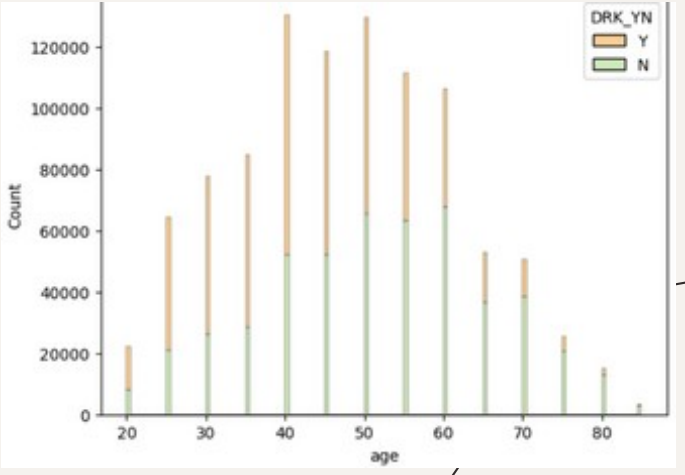
- Predict alcohol consumption.
- Discern the nuances of an individual's smoking habits.
- Predicts one of the three categories: never engaged in smoking, those who once smoked but have since ceased, and active smokers.
- Uncover underlying patterns or trends influencing drinking and smoking behaviors.
- Benefits: Insurance Companies, awareness campaigns, etc.

Data Exploration

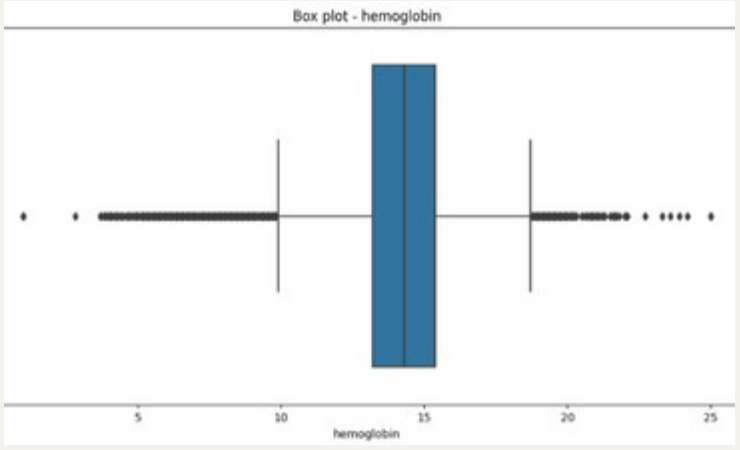


Data Exploration





Data Exploration



Before Preprocessing

	count	mean	std	min	25%	50%	75%	max
age	882865.0	47.464839	14.262972	20.0	35.0	45.0	60.0	85.0
height	882865.0	162.068017	9.268359	130.0	155.0	160.0	170.0	190.0
weight	882865.0	62.691652	12.105340	25.0	55.0	60.0	70.0	135.0
waistline	882865.0	80.649154	9.373173	48.0	74.0	81.0	87.0	116.6
sight_left	882865.0	0.954747	0.341042	0.1	0.7	1.0	1.2	2.5
sight_right	882865.0	0.952210	0.339438	0.1	0.7	1.0	1.2	2.5
hear_left	882865.0	1.030906	0.173064	1.0	1.0	1.0	1.0	2.0
hear_right	882865.0	1.029830	0.170119	1.0	1.0	1.0	1.0	2.0
SBP	882865.0	121.478702	13.654597	79.0	111.0	120.0	130.0	165.0
DBP	882865.0	75.442729	9.307915	47.0	70.0	76.0	81.0	104.0
BLDS	882865.0	97.530817	15.226252	30.0	88.0	95.0	104.0	172.0
tot_chole	882865.0	193.795205	35.743209	81.0	169.0	192.0	217.0	310.0
HDL_chole	882865.0	57.058742	14.034039	12.0	47.0	55.0	66.0	102.0
LDL_chole	882865.0	112.692138	32.693793	12.0	90.0	111.0	134.0	214.0
triglyceride	882865.0	120.165349	67.265337	1.0	72.0	103.0	150.0	409.0
hemoglobin	882865.0	14.201094	1.490764	9.5	13.1	14.2	15.3	18.9
urine_protein	882865.0	1.080163	0.392680	1.0	1.0	1.0	1.0	6.0
rum_creatinine	882865.0	0.848162	0.199967	0.1	0.7	0.8	1.0	2.2
SGOT_AST	882865.0	24.043391	8.328061	1.0	19.0	22.0	27.0	93.0
SGOT_ALT	882865.0	22.772251	12.289759	1.0	14.0	19.0	28.0	78.0

After Preprocessing

Initial Interesting Observations

- Age bracket 40-60 has most drinkers.
- Fewer Females engage in Smoking compared to Men.
- Similar to Alcohol, age 40-60 has most smokers.
- BMI, Age, and Gender, Hemoglobin somewhat influence the smoking and drinking habits the most.

Model Training Pipeline:

Basic Model Training Pipeline:

Creating a Vector using Vector Assembler

▼

Scaling the Vector Created

▼

Parameters Tuning

▼

Model Training

Machine Learning

Logistic Regression

Predicts the probability of a multinomial outcome using a linear combination of features, ideal for classification tasks.

Smoking Prediction Accuracy:
63%

Drinking Prediction Accuracy:
70%

Machine Learning

Decision Tree

Hierarchial model for predictive analysis, visually representing decisions and outcomes through a tree-like structure.

Smoking Prediction Accuracy:
68%

Drinking Prediction Accuracy:
69%

Machine Learning

Random Forest

Hierarchial model for predictive analysis, visually representing decisions and outcomes through a tree-like structure.

Smoking Prediction Accuracy:
66%

Drinking Prediction Accuracy:
69%

Problems Encountered & Future Scope

- Gradiant Boosting Algorithm just predicts binary outcomes; however, smoking column has 3 prediction outcomes.
- Biased data to some extent regarding gender.
- Removed the 'sex' variable to avoid model bias due to disproportionate gender representation.
- Focused on identifying patterns based on a broader range of equitable and informative features for improved model fairness and generalizability.

Conclusion

- **Best performing model:**

Decision TreeAccuracy: 68%

- **Drinking Prediction Model:**

Best performing model: Logistic Regression, Accuracy: 70%

- **Key Feature in Smoking**

Prediction: Hemoglobin identified as crucial feature in classifying smoking behavior.

- Assists the insurance companies and health management organization to take smart decisions to increase business.

Thanks!

Questions?