Logistic Regression using BMI, age, gender, and hbA1C to predict the presence of heart disease

...

Kristina Rivera

The Analyst

Kristie Rivera, M.Ed., MSDA

———

I am an analyst with over a decade of experience gathering, processing, and presenting data. I am a research analyst specializing in optimizing employee health to drive business profitability and reduce healthcare costs. My focus is on studying and implementing strategies that promote employee well-being, enhance productivity, and mitigate the financial burden associated with healthcare expenses. By conducting thorough analyses and identifying effective interventions, I aim to contribute to the development of evidence-based solutions that benefit both employees and the overall success of the business.

Overview-The Problem

- Impacts of heart disease on businesses include:
    - Increased healthcare costs
        - Increased insurance premiums
        - Increased claims to be paid
    - Lost productivity
    - Employee absenteeism
        - Paying benefits to employees out on medical leave
        - Paying temporary employees to cover the absent
- Estimated to cost $600 billion USD in 2023

Hypothesis:
Gender, BMI, HbA1C, and age will statistically significantly predict the presence of heart disease

Understanding the data

Logistic Regression Model

- Provides the PROBABILITY of an outcome (0-1)
  - An employee's probability of developing heart disease at some point
  - Uses historical data to make a prediction of heart disease
  - Will be able to identify those in need of wellness initiatives to prevent heart disease
  - A cut-off value will need to be established to determine the eligible employees

```
df.head(25)
```

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
5	Female	20.0	0	0	never	27.32	6.6	85	0
6	Female	44.0	0	0	never	19.31	6.5	200	1
7	Female	79.0	0	0	No Info	23.86	5.7	85	0
8	Male	42.0	0	0	never	33.64	4.8	145	0
9	Female	32.0	0	0	never	27.32	5.0	100	0

Understanding the variables

Gender (AAB)	Age	Smoking History
Men are at higher risk of heart disease, important to know the sex of the employee for accurate predictions	- Heart disease risk increases with age - Avg 41.9 - Range 0-80	Current smokers have a greater risk than former smokers have a greater risk than non smokers

Understanding the variables

BMI

- Ratio of height and body weight but can be misleading in highly muscular people
- Range 10-95.69
- Not Useful for children

HbA1C

- Gives overall view of last three months of blood sugar levels, levels above 7 can point towards a heart disease diagnosis, present or future
- How much sugar is bound to blood cells

Data Processing

Encode all variables to numerical

- Logistic regression requires all variables to be numeric

Normalize data

- Puts all data in same scale, usually 0-1
- BMI larger values than HbA1C can skew the results
- American Cancer Society says normal BMI is 18.5-24.9, avg here is 27.3
  - Max is 95.69

Findings:
Men are at highest risk
Age is the most important variable
HbA1C least important
Stopping smoking reduces risk

Limitations of the Techniques Used

Logistic Regression	- Assumes linear relationships between data points and outcome - Ex: The older the employee the higher the risk
Normalizing the Data	Assumes each variable is as equally important as the others
Data Quality	- Accurate model requires accurate data - Representative Sample

What Now?

Identify candidate pool	• Run the model on employee data, setting a threshold to identify the candidate pool
Create Wellness Initiatives	• Work with medical professionals to increase physical activity, decrease smoking and BMI
Follow Up	• Regularly follow up with targeted employees to ensure compliance • Run data yearly to identify new candidates

Expected Outcomes

Save Money

- $53 per employee, per year the employee is in the lowest risk category
- 26% reduction in health care costs

Absenteeism

- 28% reduction in sick leave days
- 30% reduction in workers compensation and disability claims

Recruitment

- Recruit and retain high-quality employees

Expected Outcomes

Employee Benefit

- Happier, healthier people are more productive and have higher energy levels

Employee/Employer Benefit

- Foster employee engagement
- Workplace cooperation