

HEALTH METRICS IN PREDICTING DIABETES

DSF1 TEAM 8
ZHI SHEN, GUO YONG, SHAWN



TABLE OF CONTENTS

01

**PRACTICAL
MOTIVATION**

02

**EXPLORATORY
DATA ANALYSIS**

+

+

03

**CORE
ANALYSIS**

04

INSIGHTS



+

01

+

PRACTICAL MOTIVATION





422 MILLION

Individuals has diabetes worldwide according to the World Health Organisation (2022).

The prevalence of diabetes in Singapore is costing our country over \$1 billion a year to manage. Learn more about how we intend to win the war against this lifestyle condition.



The Singapore government has issued a clarion call — it officially declared war on diabetes, calling the disease one of the biggest drains on the healthcare system, and one which costs the country over \$1 billion a year to manage.

Prevalence of Diabetes

During the 2016 Committee of Supply debates in Parliament, Health Minister Gan Kim Yong revealed that over 400,000 people have diabetes in Singapore. Of these, one in three is not aware he/she has the disease, and of the rest who do know, one in three has poor control of it. If left unchecked, nearly one million people in Singapore will have diabetes by 2050.

Fig. 1: *Prevalence of Diabetes in Singapore.*
(Singapore's War on Diabetes, 2021)

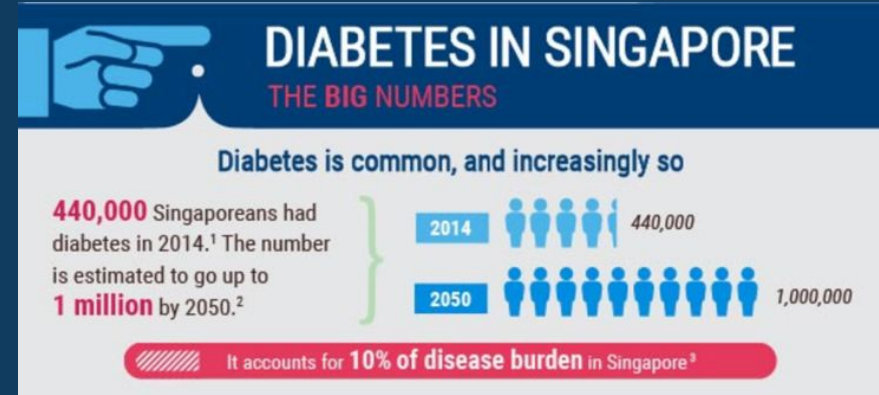


Fig. 2: *Diabetes in Singapore.*
(James, T., n.d.)



**WHAT ARE SOME OF THE
IMPORTANT HEALTH METRICS
IN DETERMINING RISK OF
DIABETES?**



THE DATASET

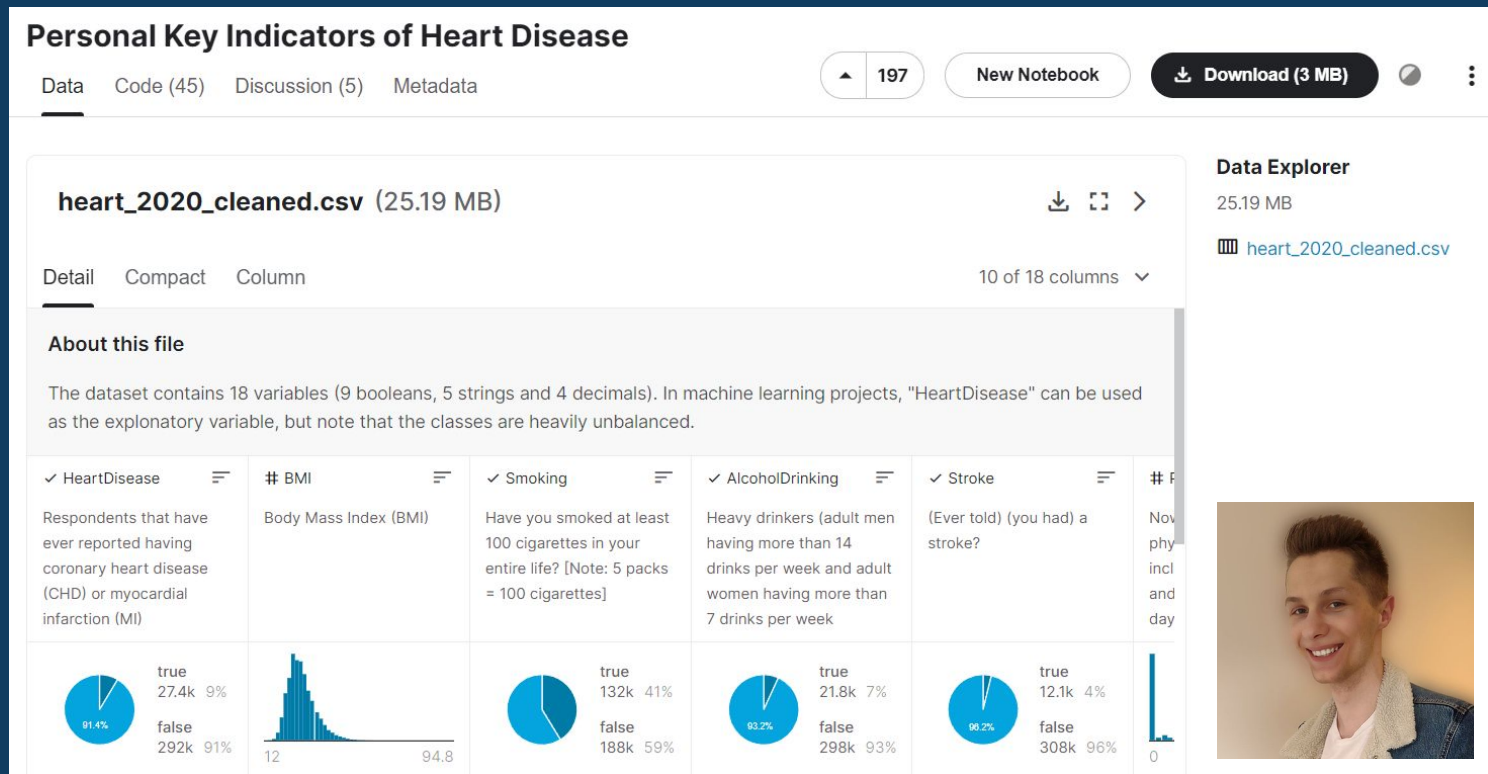
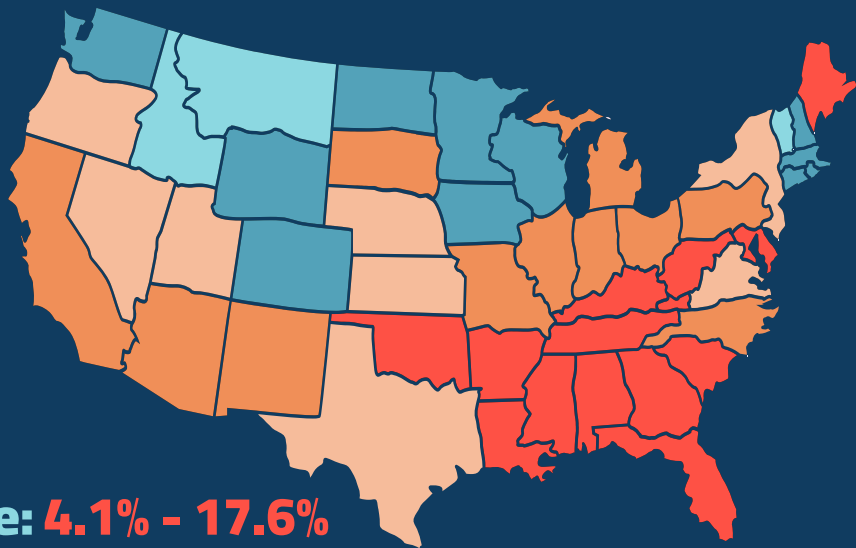


Fig. 3: Personal key indicators of heart disease dataset on Kaggle (Pytlak, K., 2022)

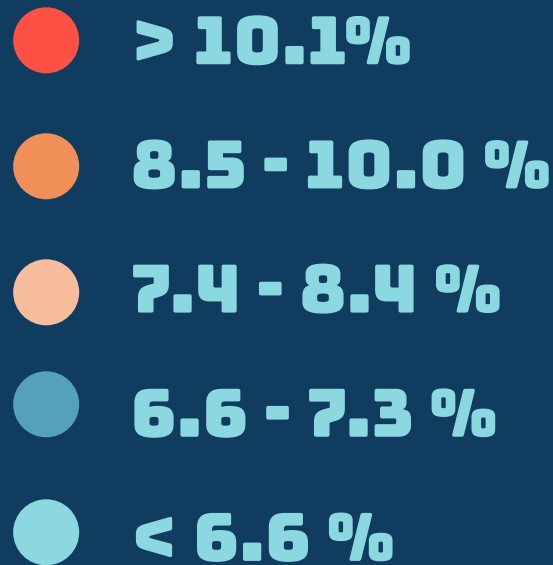
PREVALENCE IN THE U.S. (2019)

Estimates of diagnosed diabetes across US counties



Range: **4.1% - 17.6%**

Average: **8.7%**



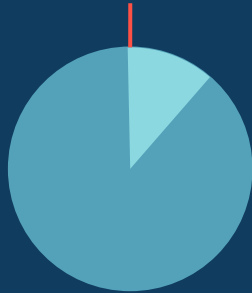
Source: National Diabetes Statistics Report (CDC, 2020)



SINGAPORE VS U.S.

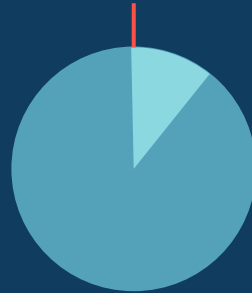
Percentage Estimates of Individuals with Diabetes (2021)

11.6%



Singapore

10.7%



United States

Source: International Diabetes Federation, 2021

02

EXPLORATORY DATA ANALYSIS





INITIAL CLEANING OF DATA



- **Dropping** duplicate entries
- **Dropping** subjective variables:
Mental health and physical health
- No missing values
- **Dropping** diabetic variable categories:
No (with borderline diabetes) and Yes
(during pregnancy) rows





VARIABLES AT A GLANCE



BMI	Sleep Time	Sex
Age Category	General Health	Race
Physical Activity	Alcohol Drinking	Smoking
Difficulty Walking	Asthma	Kidney Disease
Skin Cancer	Heart Disease	Stroke

Legend:

Numeric Categorical





BMI & DIABETES

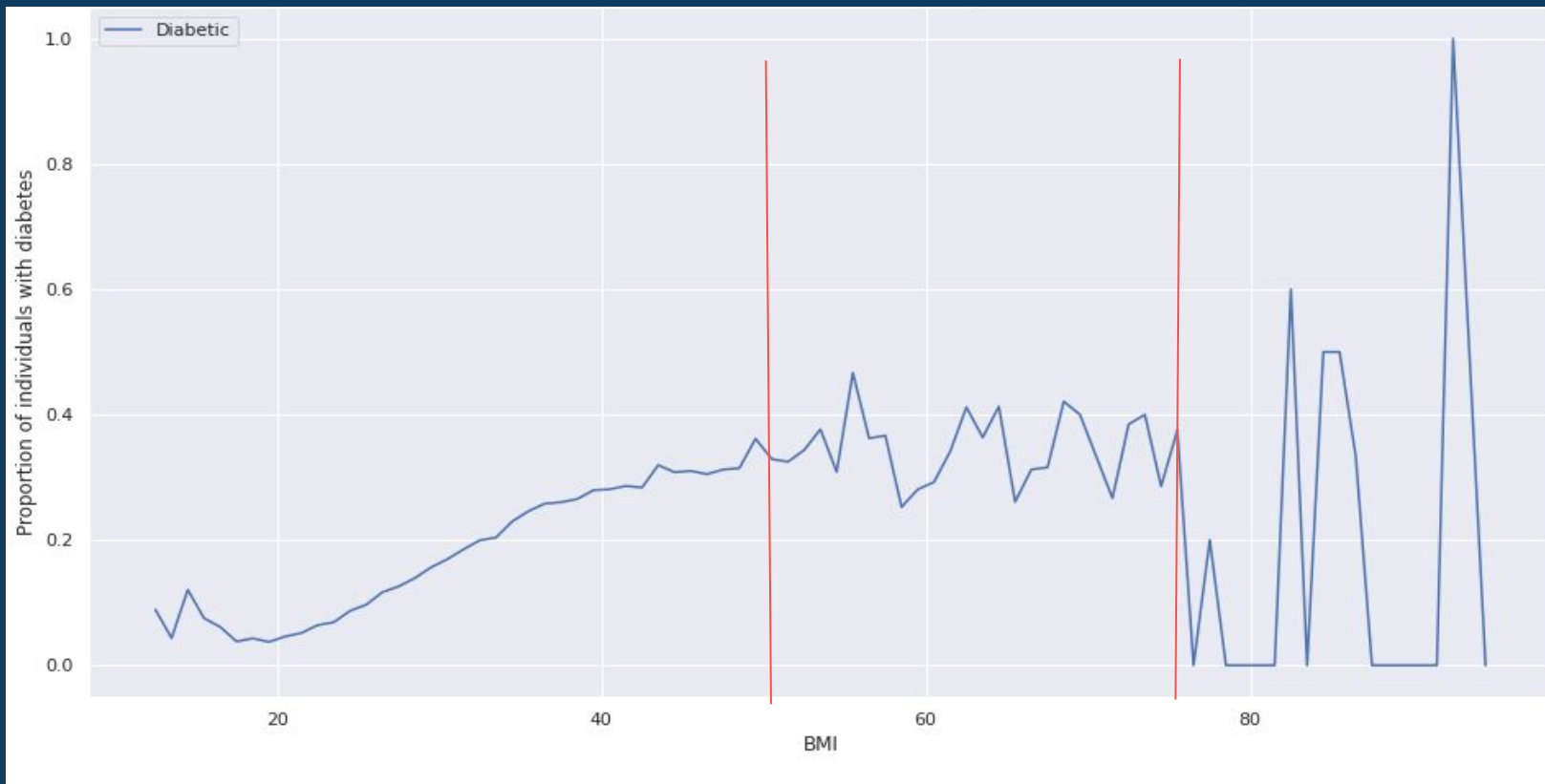


Fig. 4: *Proportion of diabetic individuals against BMI*





SLEEP TIME & DIABETES

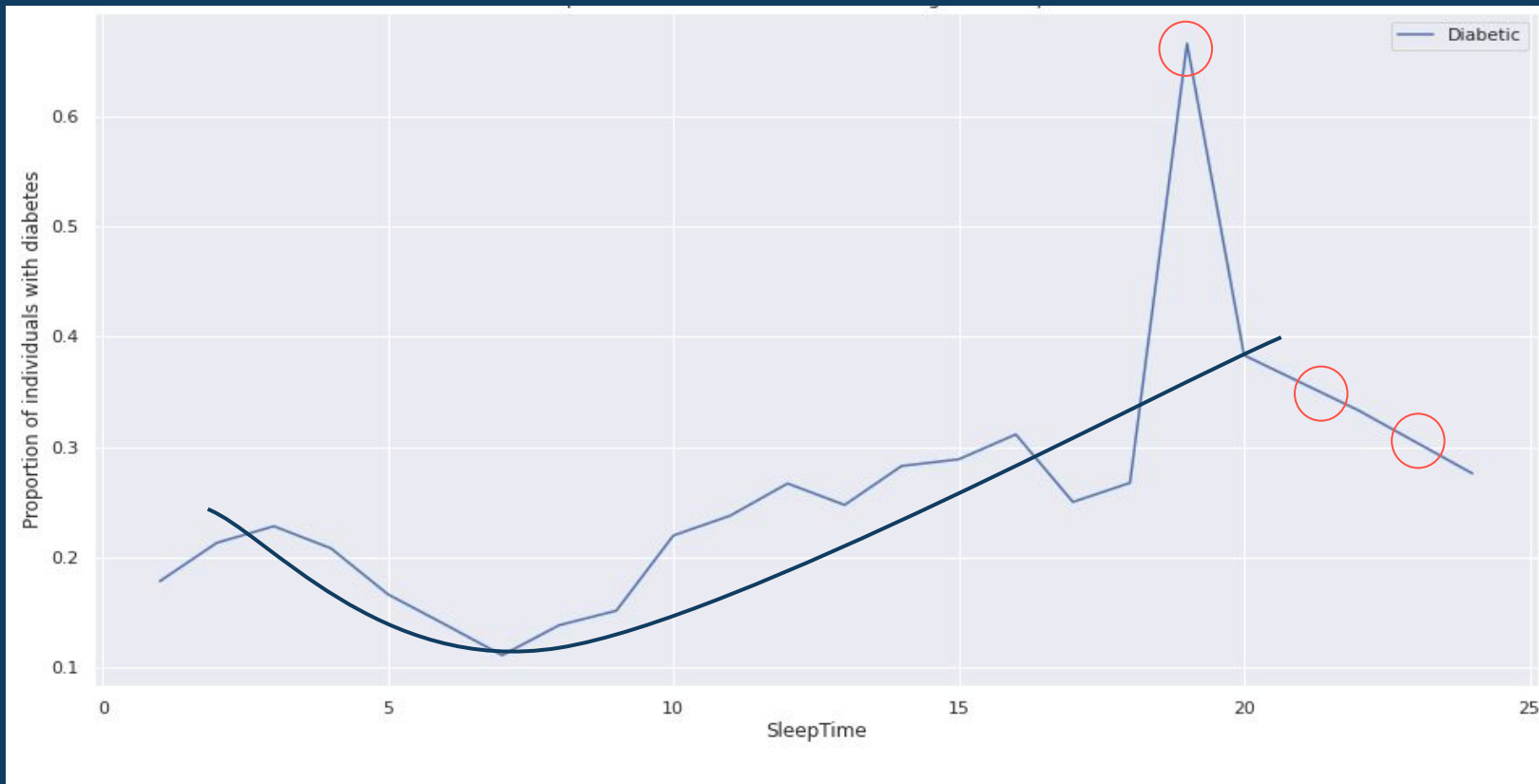


Fig. 5: Proportion of diabetic individuals against sleep time





MODIFICATIONS TO DATASET



- **Polynomial Feature** was used to model sleep time.
- **Dropping outliers** based on box plots.
- The data was transformed into a **Gaussian distribution** of **mean = 0** and **s.d. = 1**.



AGE CATEGORY & DIABETES

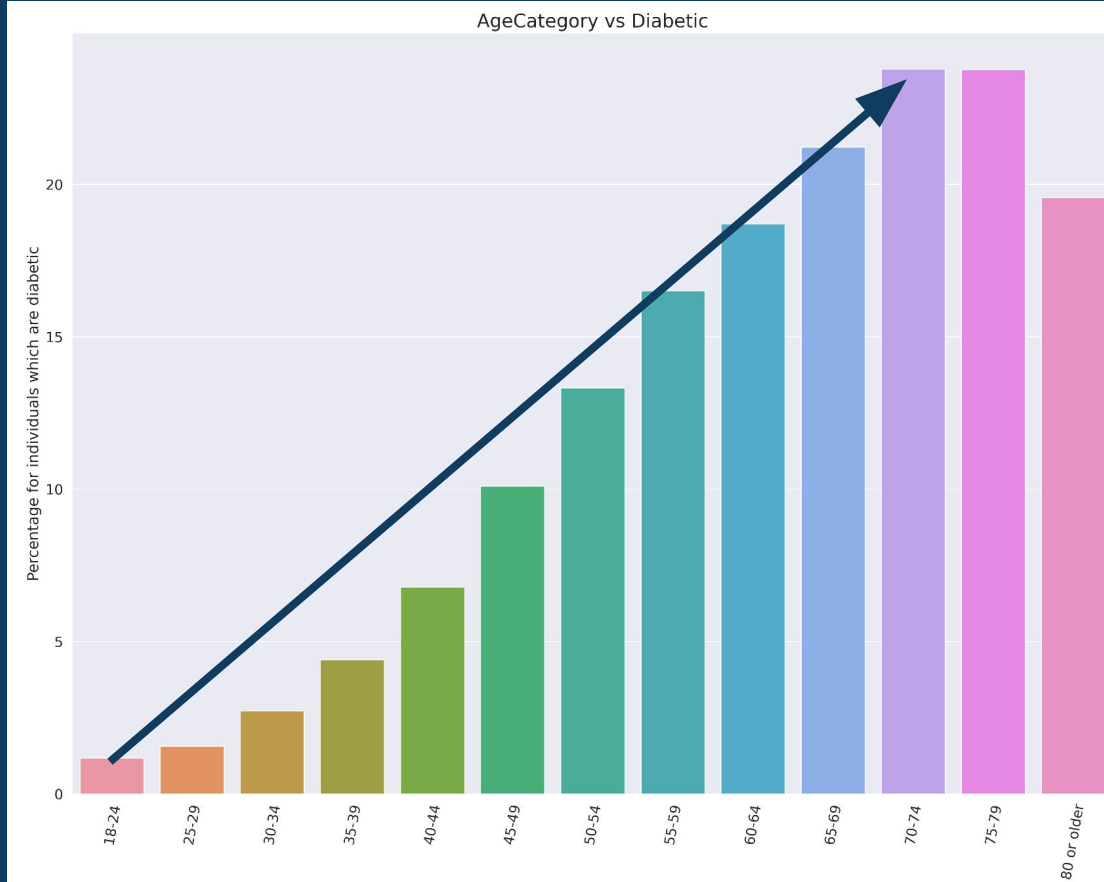


Fig. 6: *Percentage of diabetic individuals with age category*



INITIAL INSIGHTS



Categorical variable	Relationship with Diabetes (response variable)
Age Category	Proportion increases with age
Others	Proportion changes significantly
Sex, Asthma, Skin Cancer	Proportion remains relatively similar

Most categorical variables show a **relationship** with diabetes.



CHI-SQUARED TEST OF INDEPENDENCE

Null hypothesis
Independence

Alternative hypothesis
Association

Degrees of Freedom	Chi-Square (χ^2) Distribution									
	Area to the Right of Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750

Fig. 7: Chi-square distribution table (Seb, 2021)

Existence of Relationship

Strength

Chi-squared value (χ^2) \geq critical value

$\chi^2 \rightarrow$ Cramer's V



CHI-SQUARED TEST RESULTS



Categorical variable	General Health	Age Category
Chi-squared value	23453.4700	15657.2020
Critical value	9.488	21.026
Critical value \geq chi-squared value?	Yes \rightarrow Relationship exists	Yes \rightarrow Relationship exists
Cramer's V	0.2832	0.2314

General Health and Age Category show the **strongest relationship** with diabetes.



+

03

+ CORE
ANALYSES
USED



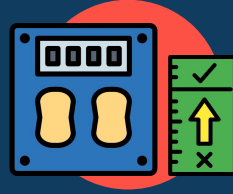


VARIABLES USED



SLEEP TIME

The approximate amount of sleep the respondents had.



BMI

Body mass index. A typical measurement for an individual's physical metrics.



GENERAL HEALTH

Respondents were asked to rate their own personal health over the past 30 days.



AGE CATEGORY

Respondents' ages in discrete categories.



HANDLING DATA IMBALANCE

- Our data was **moderately unbalanced**.
- Downsampling & Upweighting
 - **Downsampling** - Undersampling
 - **Upweighting** - Relative class weight
- We decided to **downsample** and **upweight** it by a factor of 6.6.

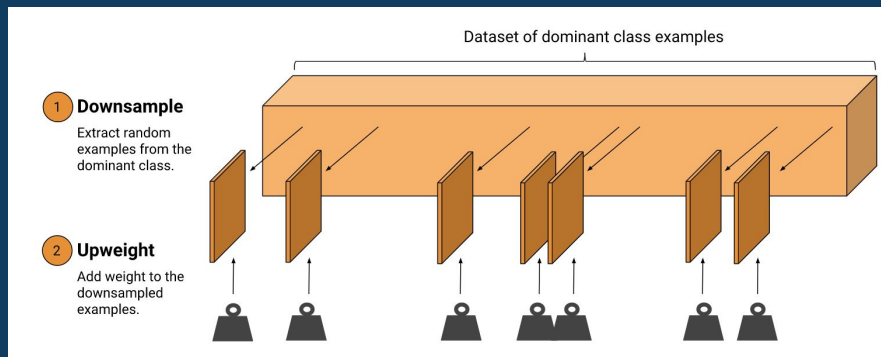


Fig. 8: Downsample and upweight
(Imbalanced Data, 2021)

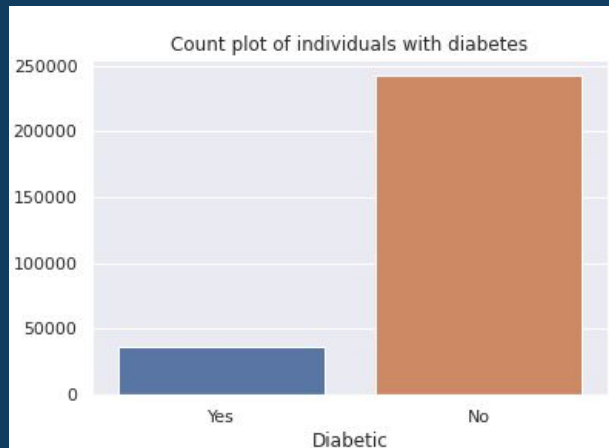


Fig. 9:
*Count plot of
diabetic
individuals*



MACHINE LEARNING TOOLS

Logistic Regression

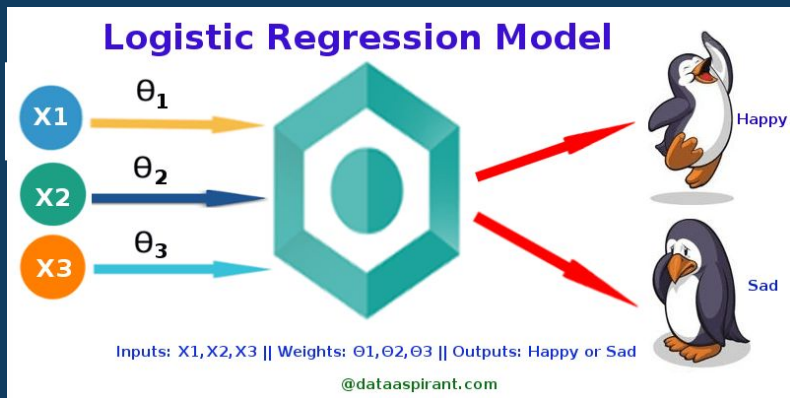


Fig. 10: *Logistic Regression Model*
(Polamuri, S., 2017)

Random Forest Classifier

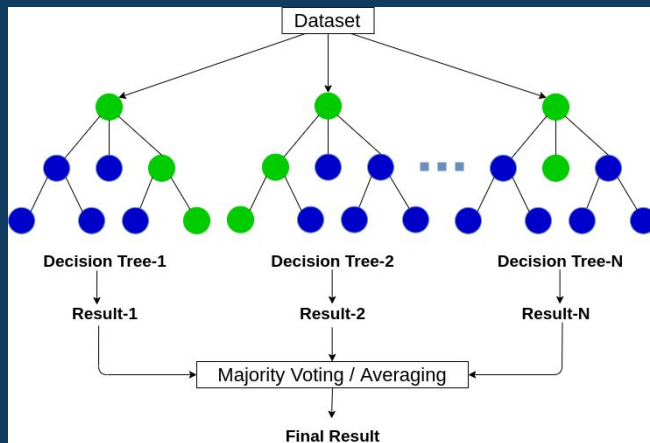


Fig. 11: *Random Forest Classifier*
(Sharma, A., 2020)





04

OUR INSIGHTS

PERFORMANCE OF EACH MODEL

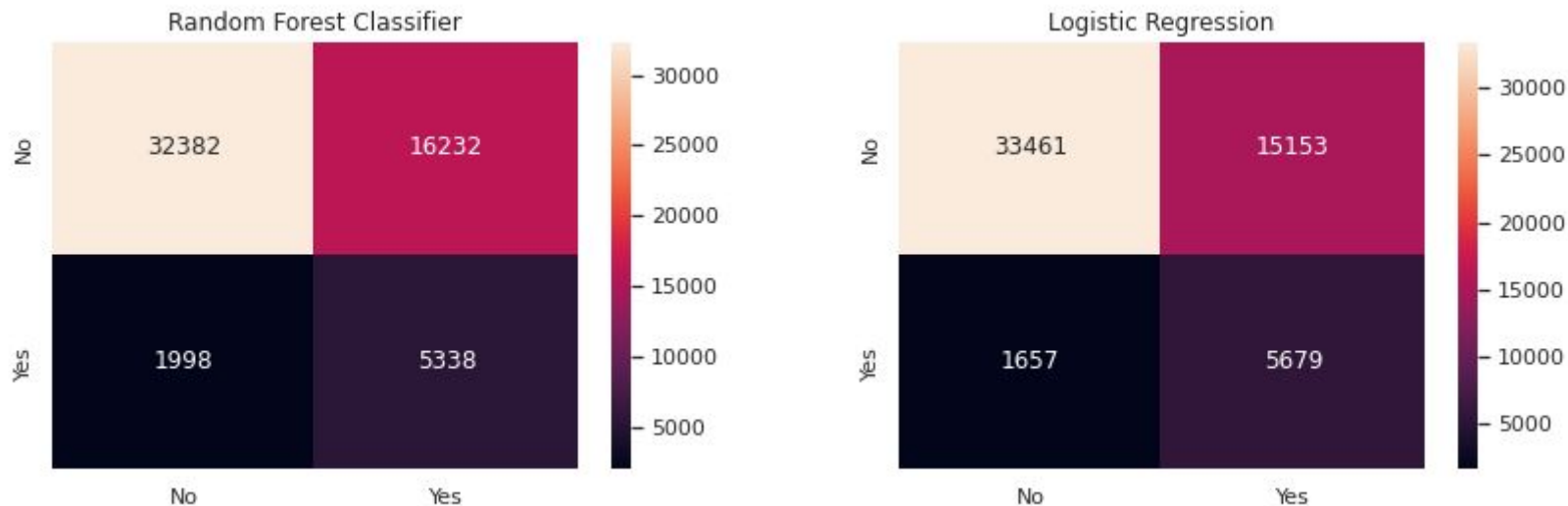

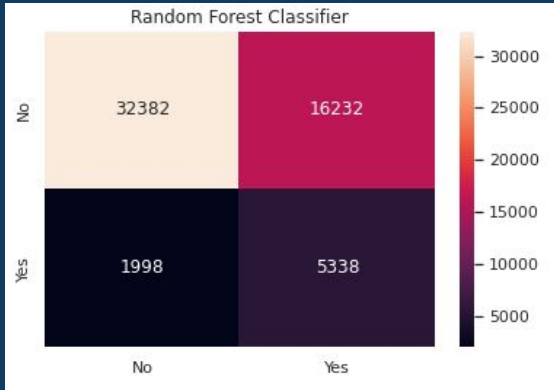


Fig. 12: *Confusion metrics for random forest classifier and logistic regression models*

PERFORMANCE OF EACH MODEL

<div>Model</div> <div>Metrics</div>	<div>Logistic Regression</div> 		<div>Random Forest Classifier</div> 	
	Accuracy	0.699553		0.674173
	FPR	0.3117		0.333896
	FNR	0.225872		0.272356
	AUC	0.802888		0.761871

RELATIVE FEATURE IMPORTANCES FOR *LOGISTIC REGRESSION*

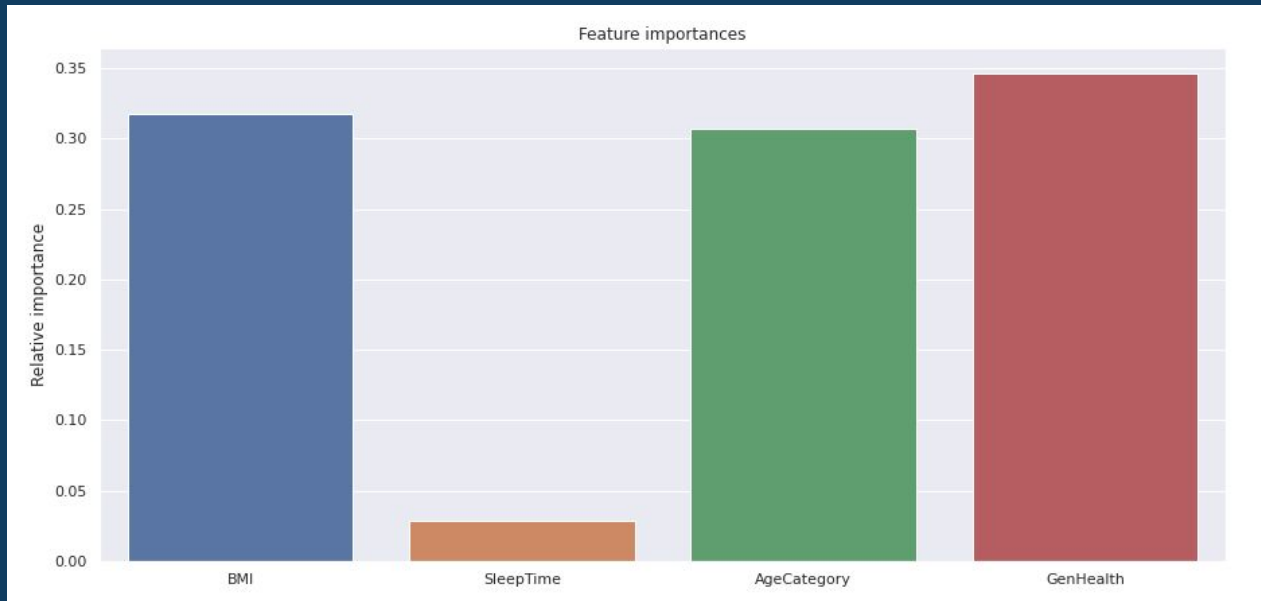


Fig. 13: *Feature importances proportion in logistic regression model*

RELATIVE FEATURE IMPORTANCES FOR *RANDOM FOREST CLASSIFIER*

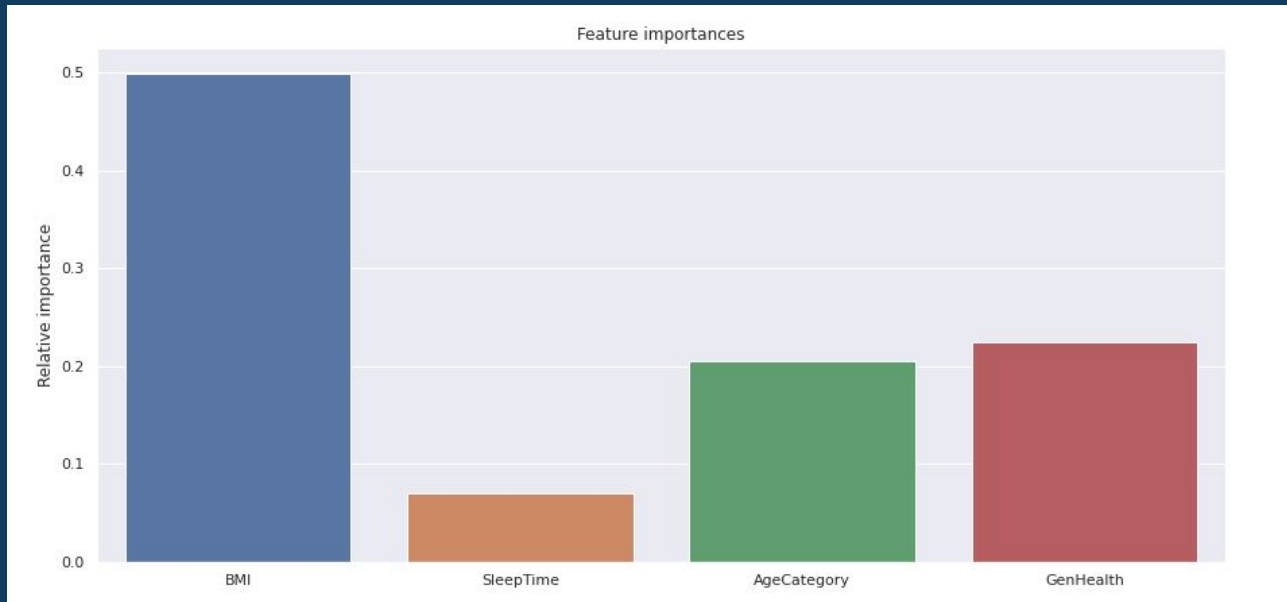


Fig. 14: *Feature importances proportion in random forest classifier model*



CONCLUSIONS



LOGISTIC REGRESSION > RANDOM FOREST CLASSIFIER

Logistic Regression was more accurate (70%) as compared to Random Forest Classifier (67%).



IMPORTANCE OF VARIABLES: SIMILARITIES



Both models considered sleep time an unimportant factor.





CONCLUSIONS



IMPORTANCE OF VARIABLES: DIFFERENCES

The two models also placed a different level of importance on each of the variables.

NUMERIC VARIABLES > CATEGORICAL VARIABLES



+ PREVENTION BETTER THAN CURE +



PHYSICAL ACTIVITY

Regular aerobic and
resistance exercises.



HEALTHY DIET

Less sweet fruits and
starchy foods. More
grains and legumes.



PORTION SIZES

Smaller portions reduce
calorie intake and regulate
insulin fluctuations.





THANKS!

CREDITS: This presentation template was
created by [Slidesgo](#), including icons by [Flaticon](#)
and infographics & images by [Freepik](#)



REFERENCES



Slide 4:

Diabetes. (2022). World Health Organization.

https://www.who.int/health-topics/diabetes#tab=tab_1

Slide 5:

James, T. (n.d.). *Diabetes in Singapore*. <https://acetutors.com.sg/diabetes-in-singapore>

Ministry of Health, Singapore (2011, Oct). *National Health Survey 2010*.

<https://www.moh.gov.sg/docs/librariesprovider5/resources-statistics/reports/nhs2010---low-res.pdf>

Singapore's War on Diabetes. (2021, May 26). HealthHub.

<https://www.healthhub.sg/live-healthy/1273/d-day-for-diabetes>

Slide 7:

Pytlak, K. (2022). *Personal Key Indicators of Heart Disease*.

<https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>





REFERENCES



Slide 19:

Seb. (2021, April 8). *Chi-Square Distribution Table*. Programmatically.
<https://programmatically.com/chi-square-distribution-table/>

Slide 22

Freepik. (n.d.). *Pulse on heart* [Icon]. Flaticon. https://www.flaticon.com/free-icon/pulse_1240841

Freepik. (n.d.). *Hourglass* [Icon]. Flaticon. https://www.flaticon.com/free-icon/hourglass_1255391

Photo3idea_studio. (n.d.). *Bed time* [Icon]. Flaticon.
https://www.flaticon.com/free-icon/bed_1257318



Smalllikeart. (n.d.). *Weighing scale* [Icon]. Flaticon.
https://www.flaticon.com/free-icon/weight-scale_1256551



Smashicons. (n.d.). *Height scale* [Icon]. Flaticon. https://www.flaticon.com/free-icon/height_950746





REFERENCES



Slide 23:

Imbalanced Data. (2021, Nov 11). Google Developers.

<https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data>

Slide 24:

Polamuri, S. (2017, Mar 2). *How The Logistic Regression Works.* Dataaspirant.

<https://dataaspirant.com/how-logistic-regression-model-works/>

Sharma, A. (2020, May 12). *Decision Tree vs. Random Forest – Which Algorithm Should you Use?* Analytics Vidhya.

<https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/>





REFERENCES



Slide 33:

Becris. (n.d.). *Food pyramid* [Icon]. Flaticon. https://www.flaticon.com/free-icon/diet_1240823

Mayo Clinic Staff. (2021, June 25). *Diabetes prevention: 5 tips for taking control*. Mayo Clinic. <https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/in-depth/diabetes-prevention/art-20047639>

Nawicon. (n.d.). *Plate with fork and spoon* [Icon]. Flaticon. https://www.flaticon.com/free-icon/portion_7126726?term=portion&page=1&position=38&page=1&position=38&related_id=7126726&origin=search

Streit, L. (2022, Jan 28). *11 Ways to Prevent Type 2 Diabetes*. <https://www.healthline.com/nutrition/prevent-diabetes>

Turkkub. (n.d.). *A person stretching* [Icon]. Flaticon. https://www.flaticon.com/free-icon/running_815082?related_id=815119&origin=search





REFERENCES



Music:

Tissot, B. (n.d.). *Cute* [Song]. <https://www.bensound.com/royalty-free-music/track/cute>.

