Using Data Science to improve Stroke Prediction

By:

Josiah Joel – 816030501

Ethan Lee Chong – 816032732

Research Goal

To compare the Kappa algorithm with other traditional data science algorithms using a real-world stroke prediction data set and a synthetic dataset.

The Kappa Algorithm

The Algorithm was designed by Dr. Patrick Hosein to maximise the personalization and robustness of predictions.

Dr. Patrick Hosein

Robustness vs Personalization for Dummies 101

	Sex	Marks
1	F	30
2	M	19
3	F	40
4	O	17

Robust mean:

$$\frac{30 + 19 + 40 + 17}{4} = 26.5$$

Personalized mean for Sex F:

$$\frac{30 + 40}{2} = 35$$

so ez

Note: This is only for Categoric features

- Kappa deals with numeric columns differently

Train set:

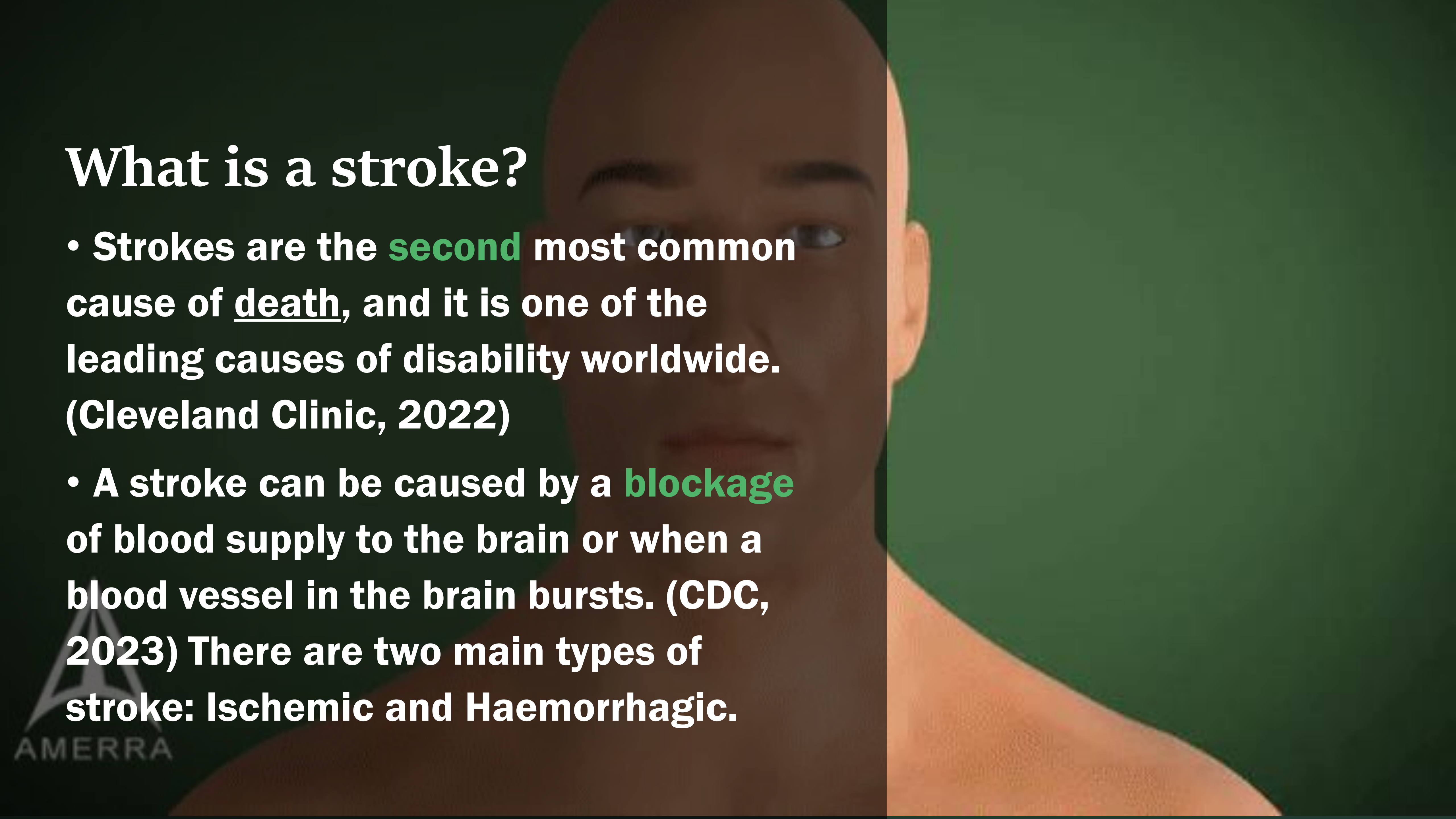
categoric column	NUMERIC COLUMN	categoric column
Sex	Mark	Pass
F	29.3	No
M	31.4	No
O	50.0	Yes
F	51.7	Yes

- if the mark 26.2 appears in the testing the algorithm wasn't trained with that value

so....

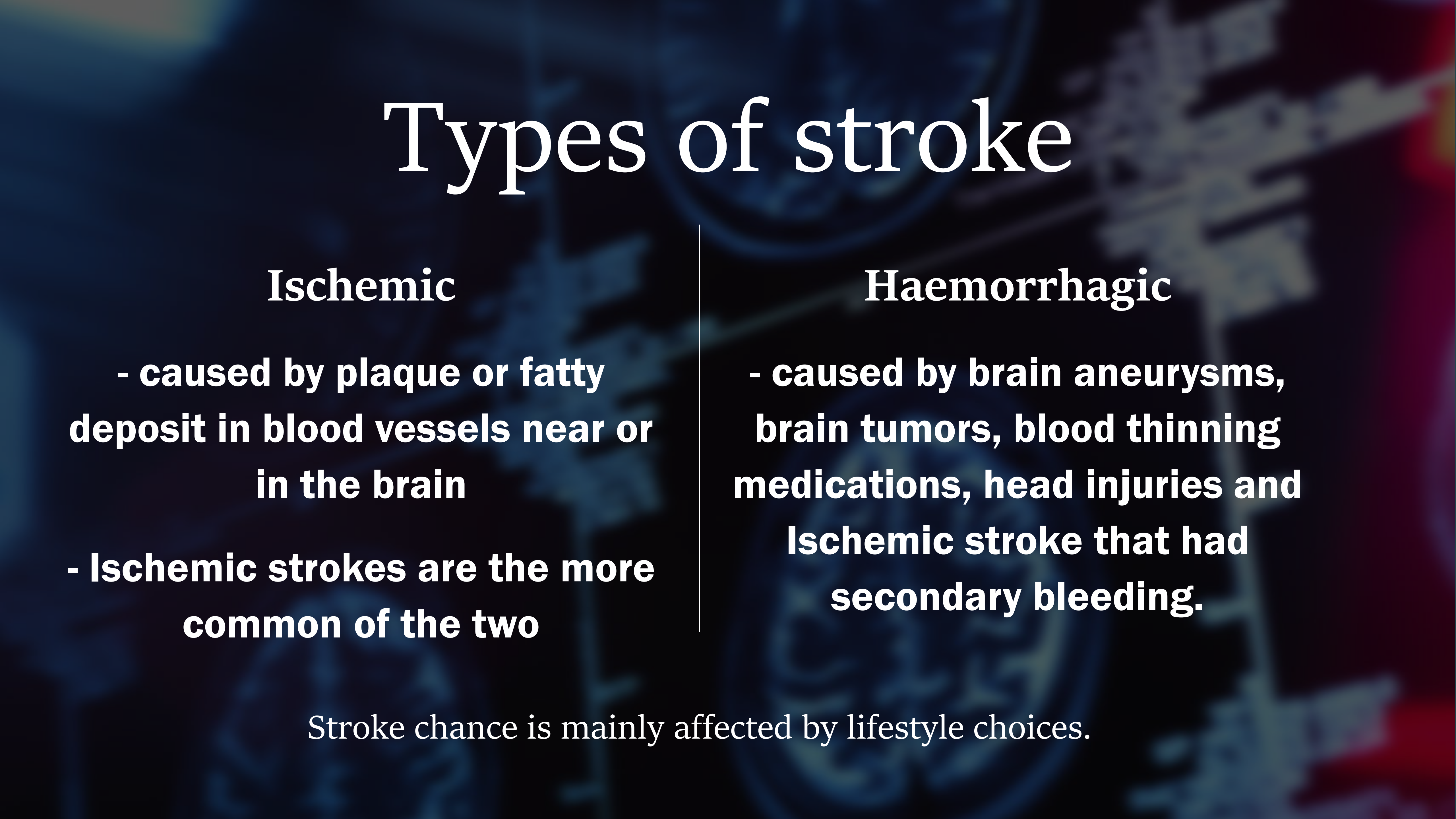
what does the algorithm do?????

puts all numbers
on one scale
so it's easy to guess



What is a stroke?

- Strokes are the second most common cause of death, and it is one of the leading causes of disability worldwide. (Cleveland Clinic, 2022)
- A stroke can be caused by a blockage of blood supply to the brain or when a blood vessel in the brain bursts. (CDC, 2023) There are two main types of stroke: Ischemic and Haemorrhagic.



Types of stroke

Ischemic

- caused by plaque or fatty deposit in blood vessels near or in the brain

- Ischemic strokes are the more common of the two

Haemorrhagic

- caused by brain aneurysms, brain tumors, blood thinning medications, head injuries and Ischemic stroke that had secondary bleeding.

Stroke chance is mainly affected by lifestyle choices.

Beneficiaries

Helps medical industry with accurate predictions.

General public benefits from analytics and predictions.

Predict high-risk individuals for timely intervention.

The Kappa algorithm compares each row of test data with all the training data. This maximizes the robustness.

The weights are also calculated using the numerical and categorical means maximizing the personalization.

$$\hat{y}_j[k] \equiv \frac{\sum_{i\in S} \frac{y_i}{(1+|x_i-x_j|)^{\boxed{\kappa^*}}}}{\sum_{i\in S} \frac{1}{(1+|x_i-x_j|)^{\kappa^*}}}$$

Don’t worry. This is the important part.
Kappa constant.

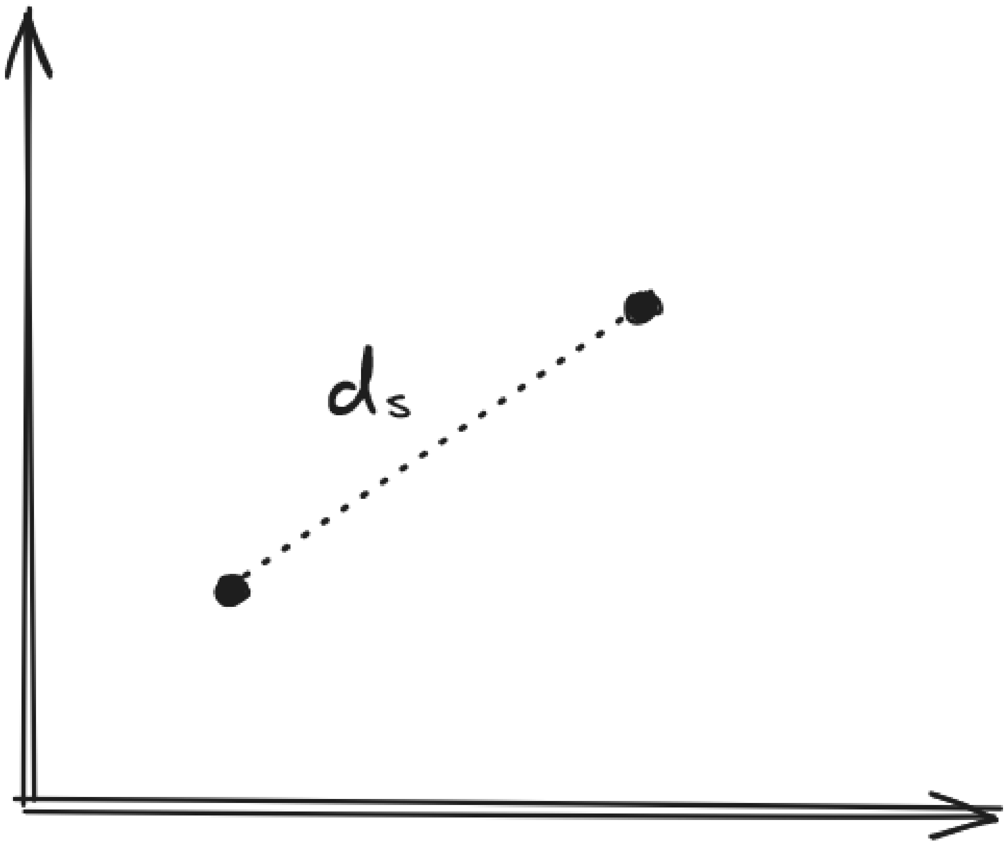
Same formula but less scary:

$$g(\kappa) = \frac{\sum_{s \in S} w(s,\kappa) M(s)}{\sum_{s \in S} \boxed{w(s,\kappa)}}$$

Let's simplify this formula:

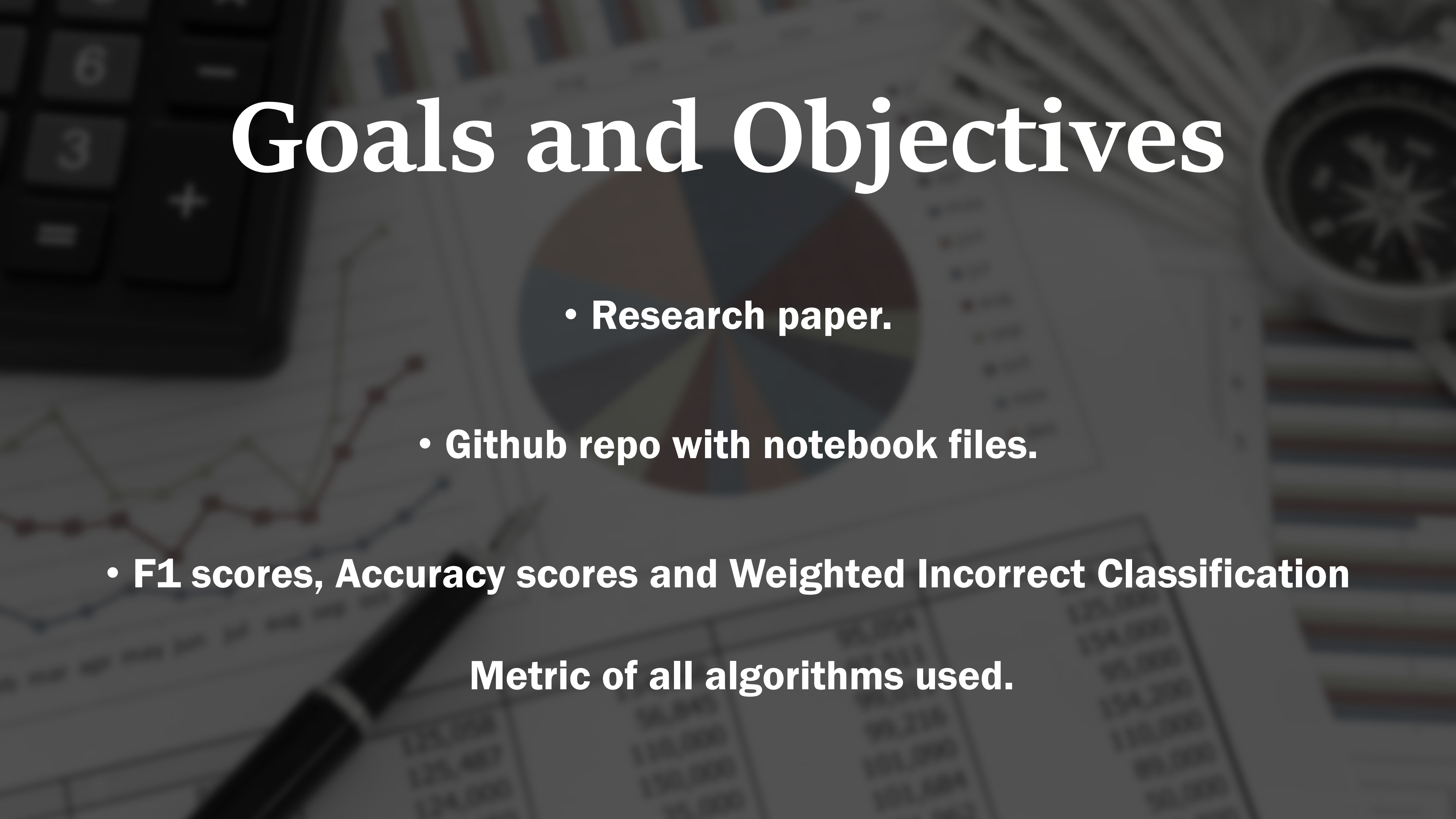
Weighted average w:

$$w(s,\kappa) = \frac{1}{(1 + d_s)^{\kappa}}$$



Significance of the Project Solution

If the algorithm proves to be better at classification in this dataset than traditional approaches it could be applied to other real-world problems and problems with many more features. Many lives can also be saved from improvements.



Goals and Objectives

- Research paper.
- Github repo with notebook files.
- F1 scores, Accuracy scores and Weighted Incorrect Classification

  Metric of all algorithms used.

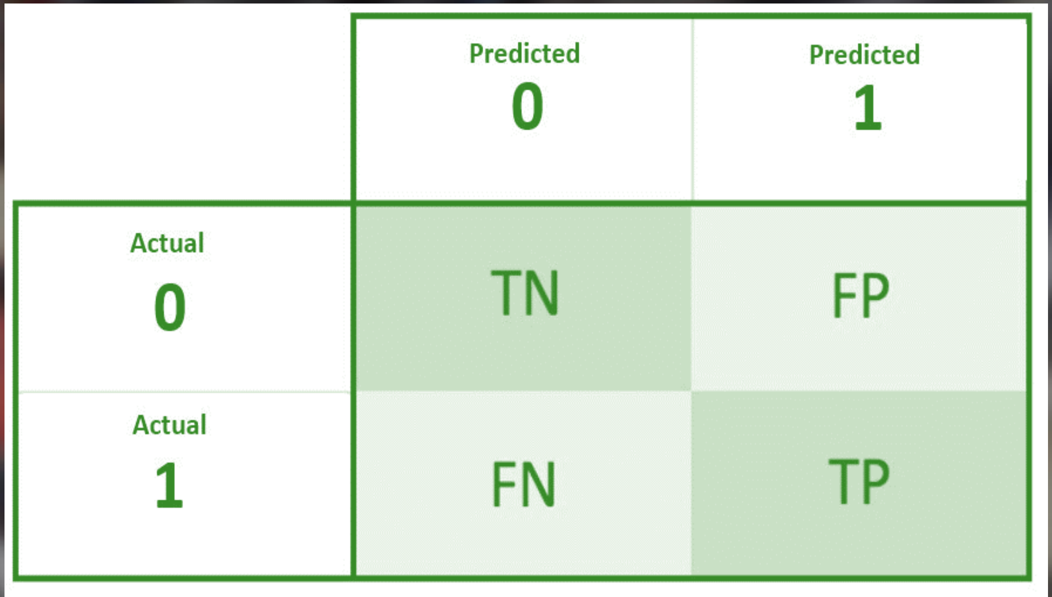
F1 scores

F1 scores

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$$

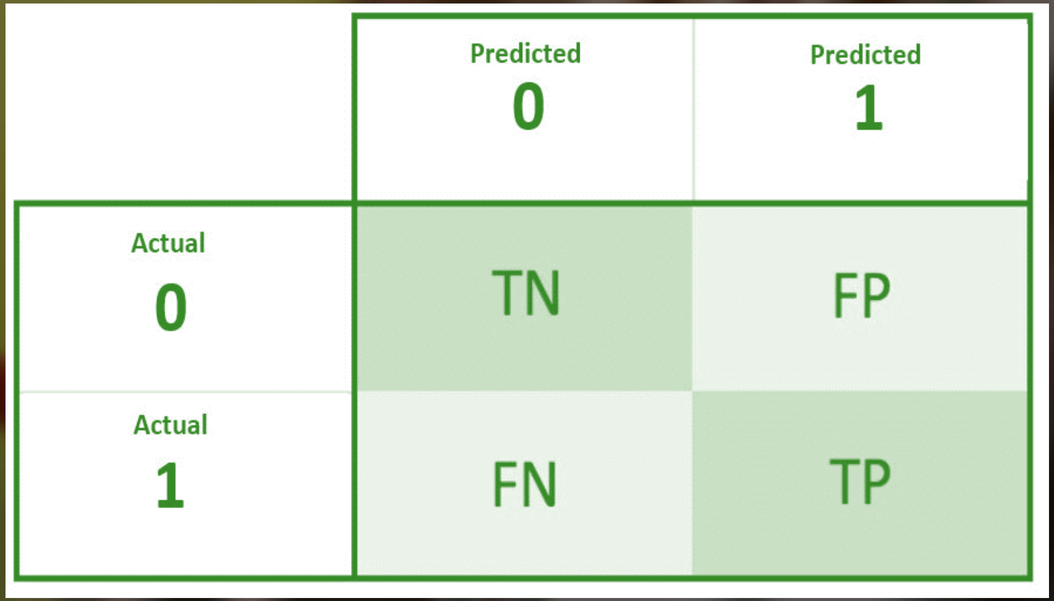
$$F1 = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Accuracy scores

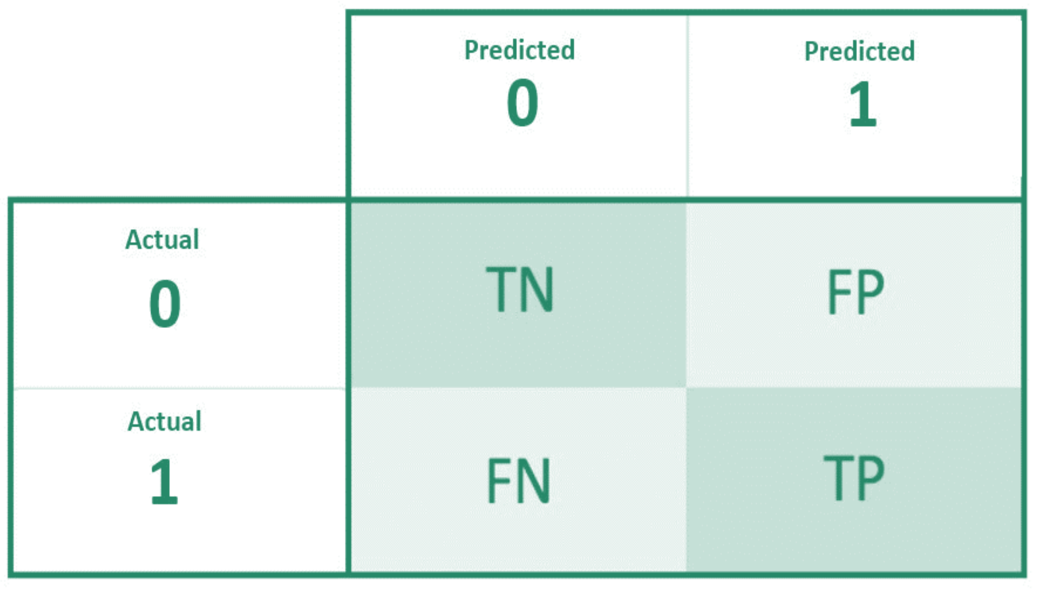
$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Population}}$$



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

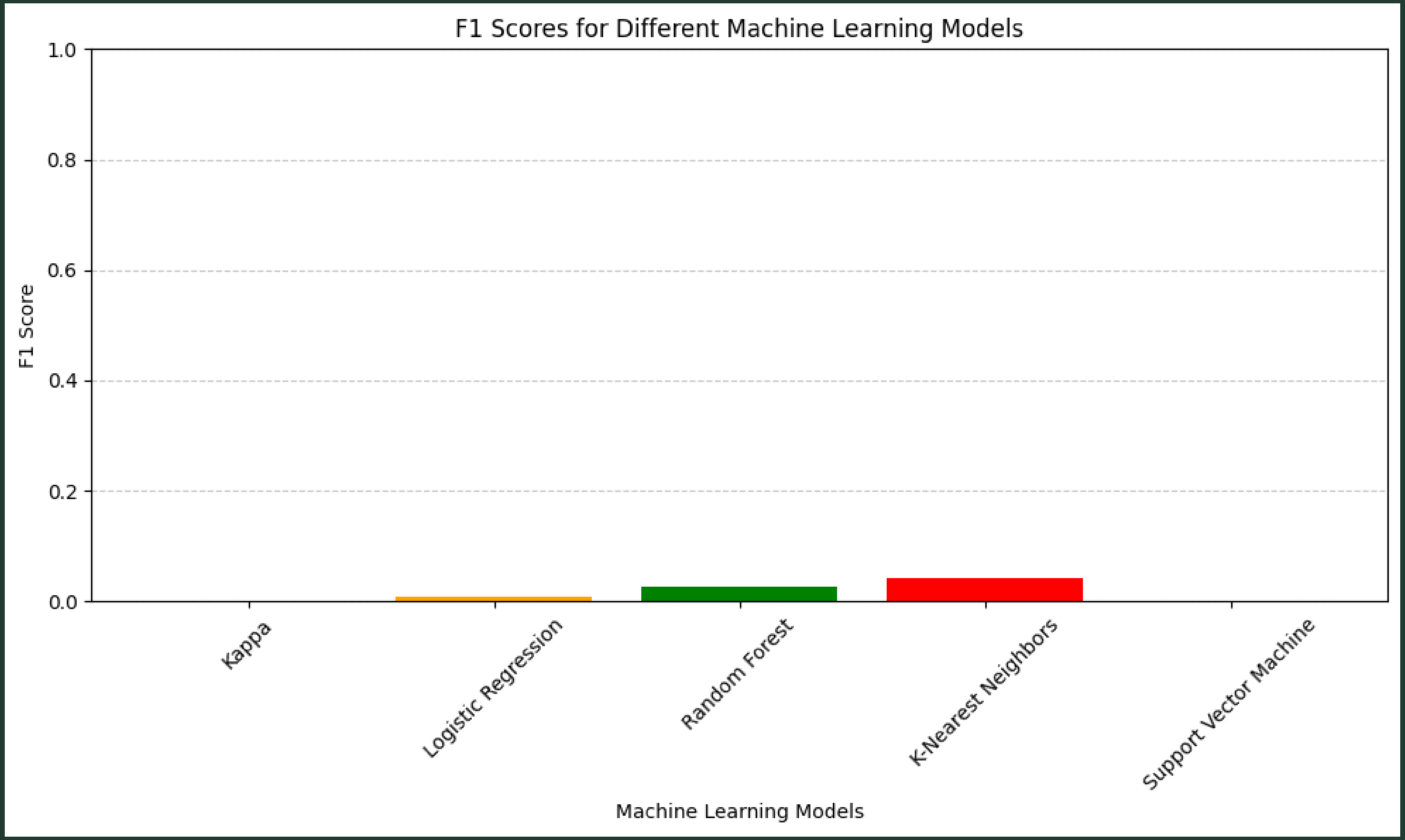
WICM scores

$$\text{Weighted Incorrect Classification Metric} = \frac{\text{Weight} \times \text{False Negatives} + \text{False Positives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

We set our weight arbitrarily to 10



Our Dataset is very skewed

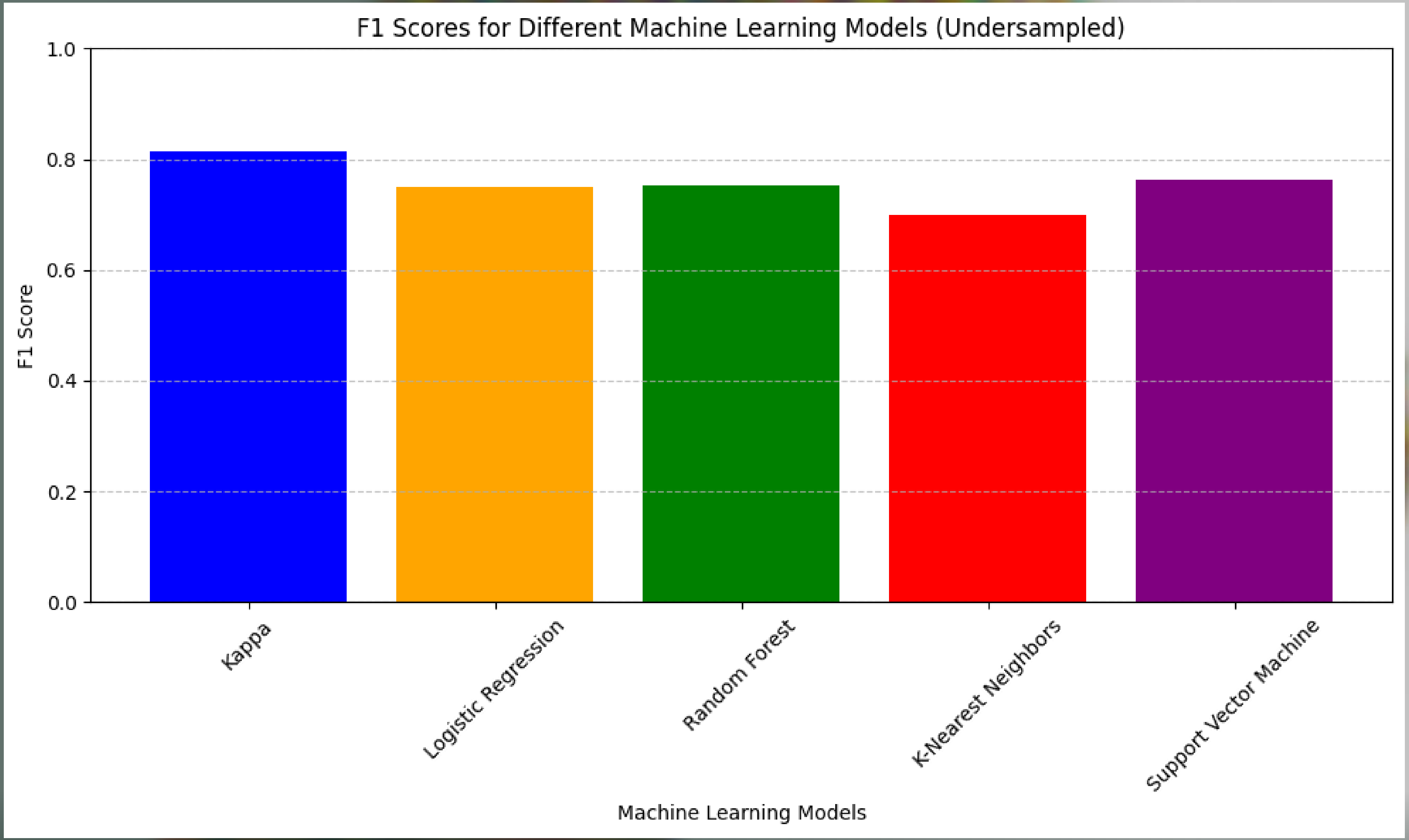
Around 95% of the rows in the dataset are negative for stroke. We are predicting stroke, and this is not good. To solve this issue, we use:

SMOTE OVERSAMPLING

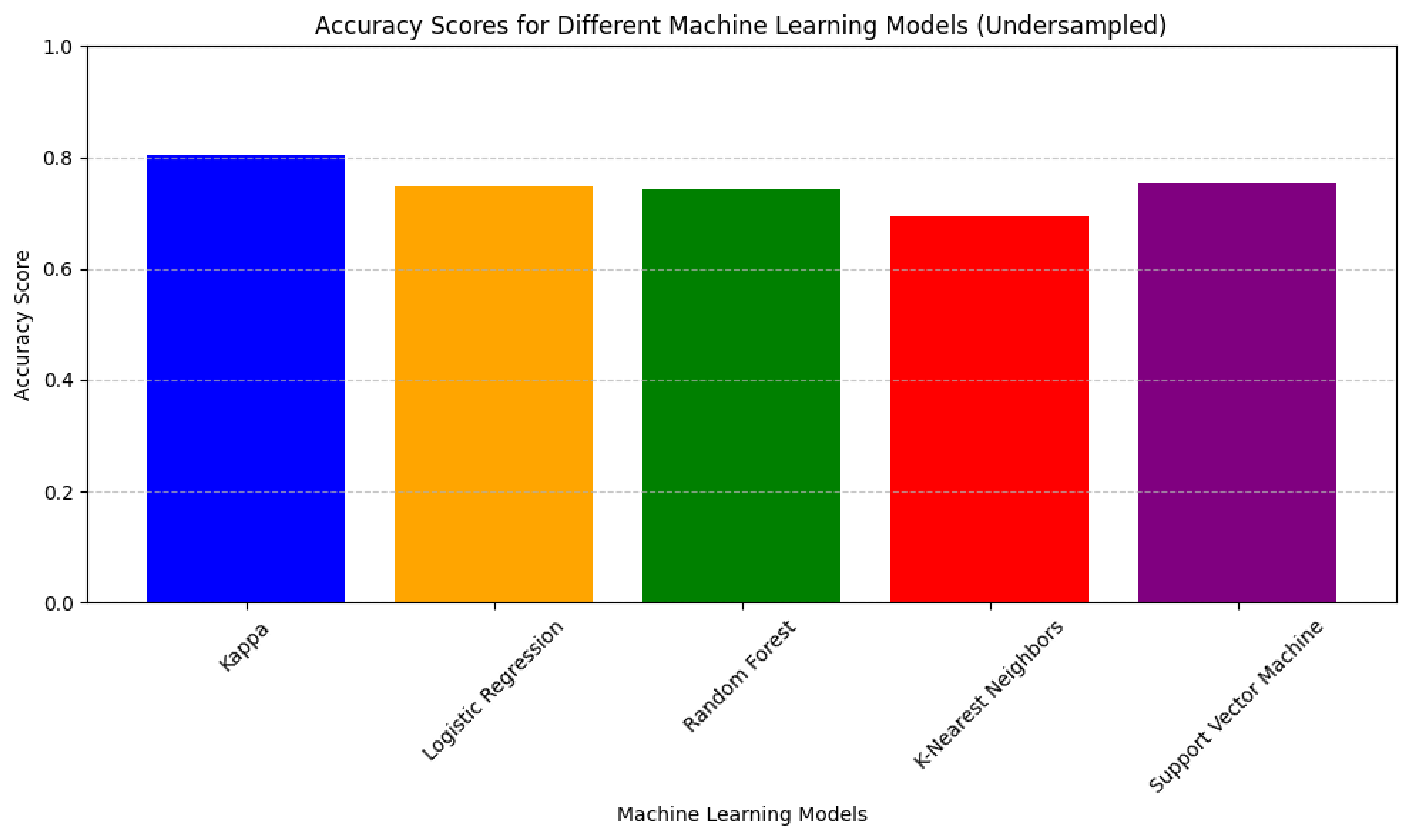
and

UNDERSAMPLING

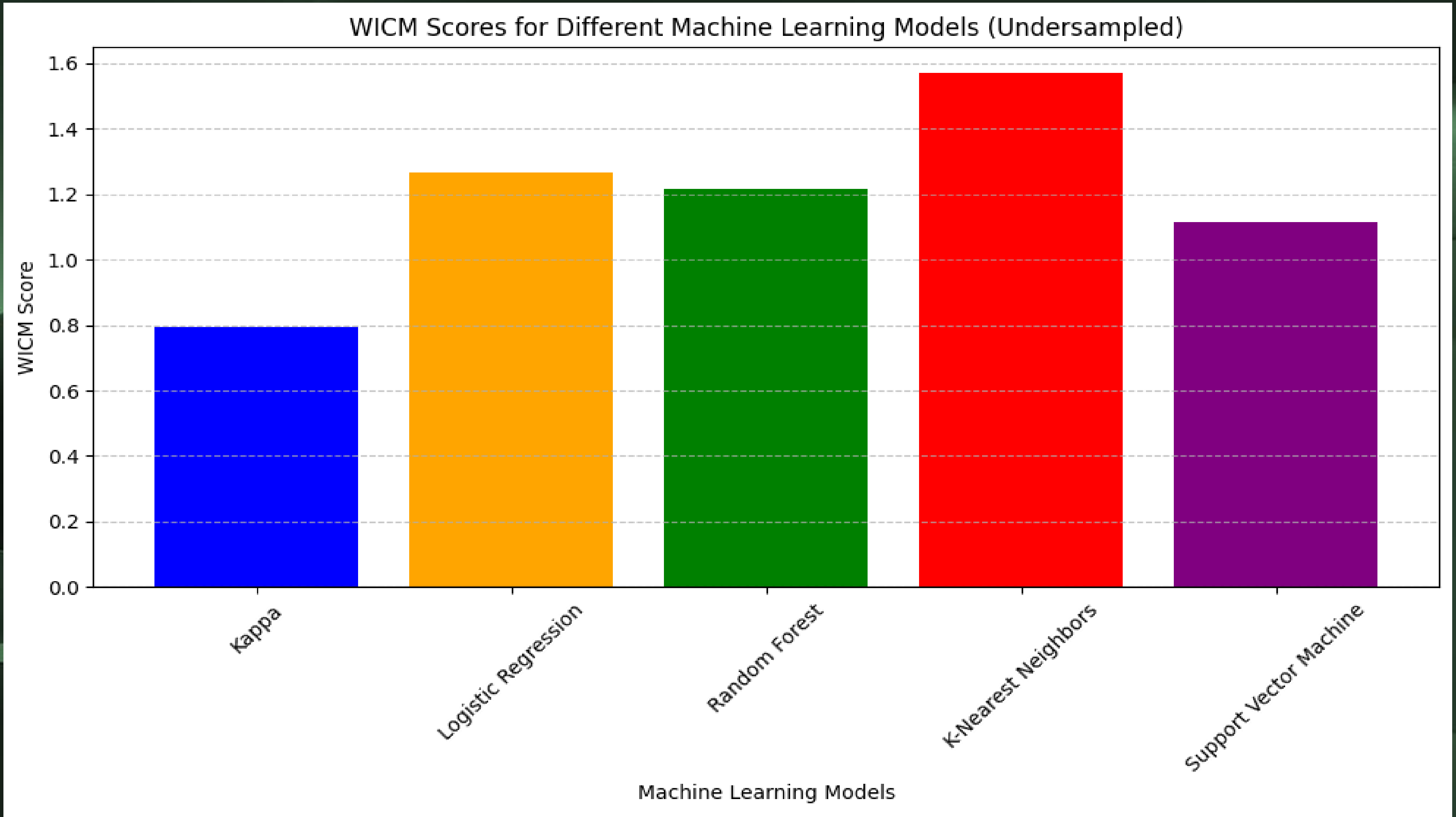
Undersampled F1 scores



Undersampled Accuracy scores

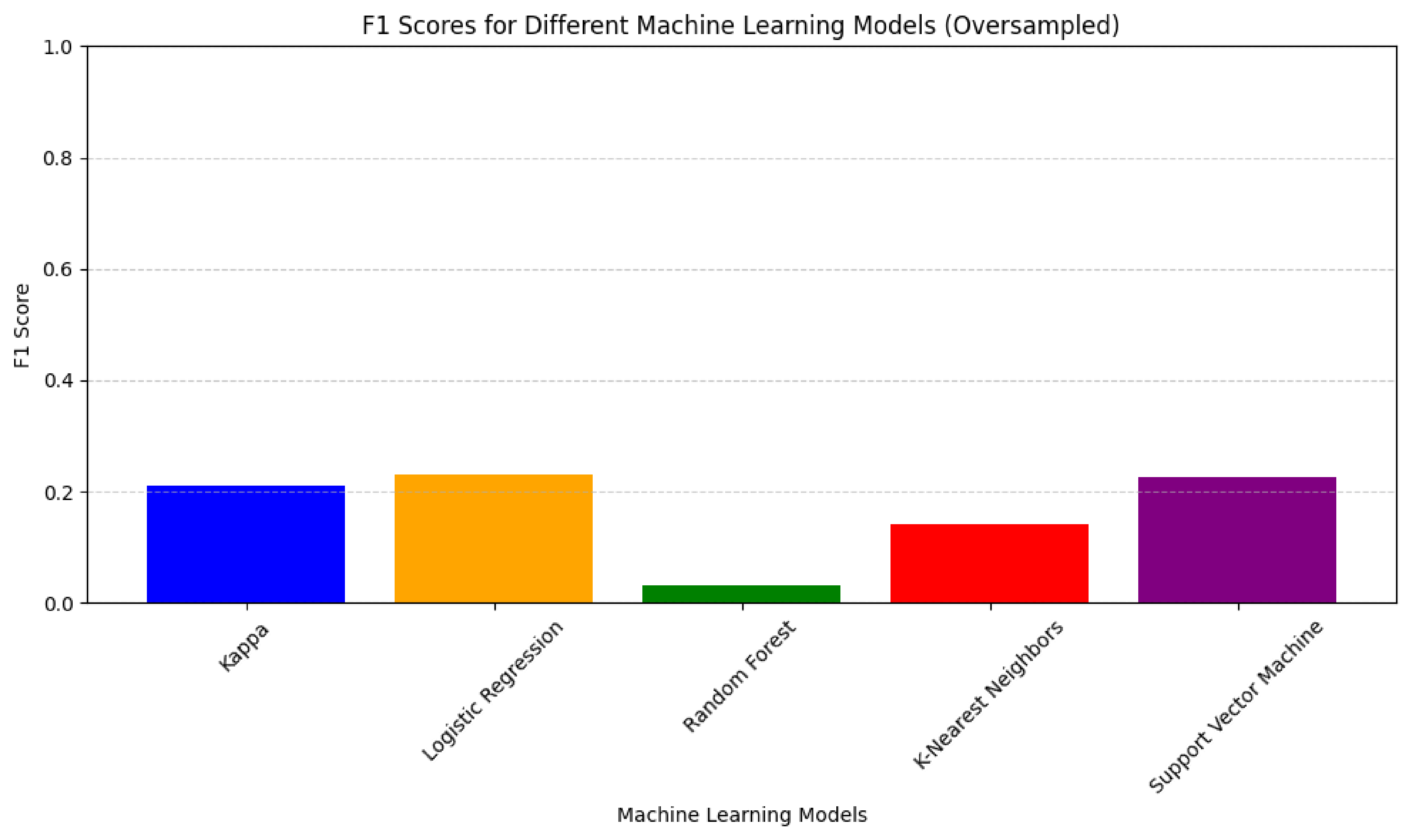


Undersampled WICM scores

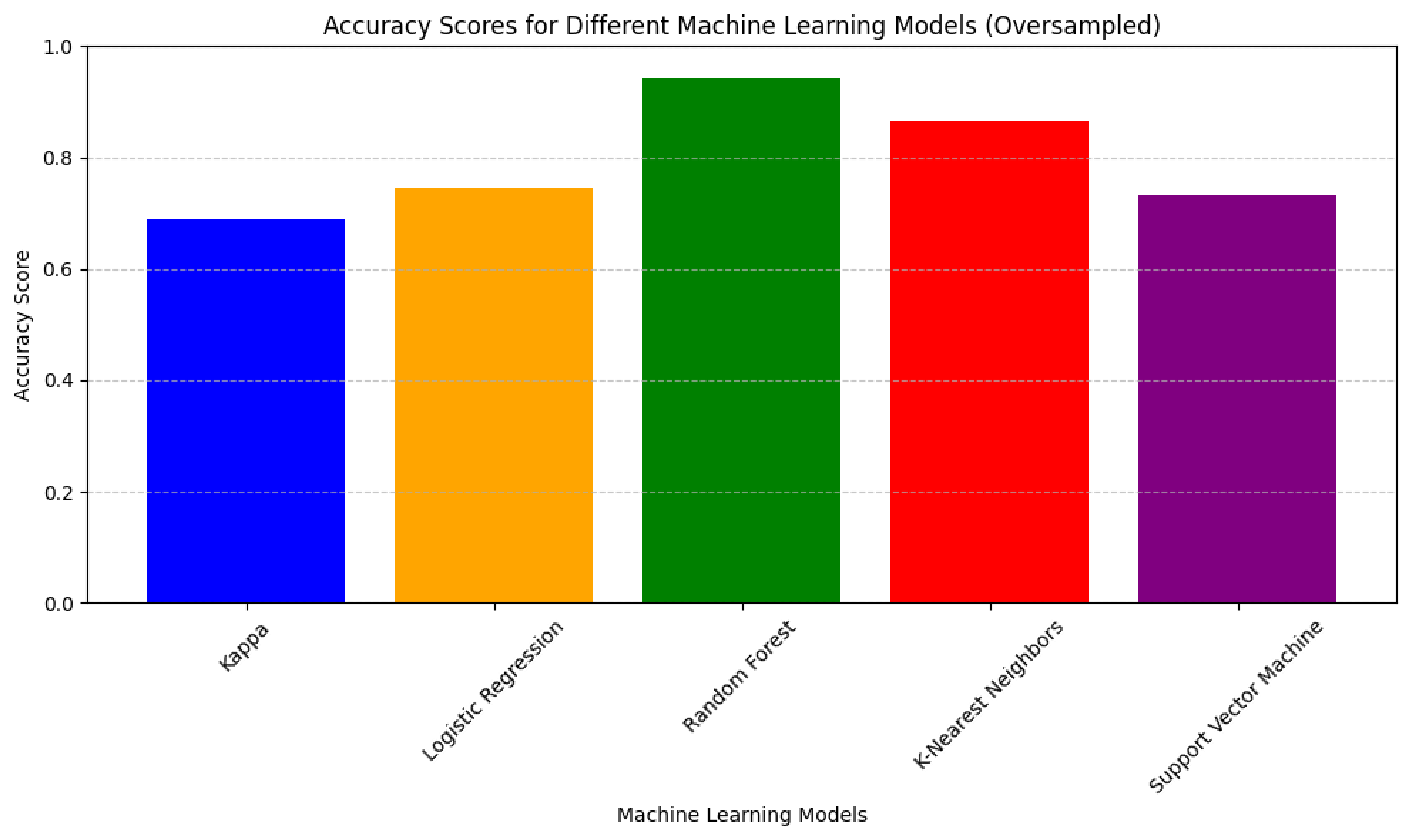


Lower is better

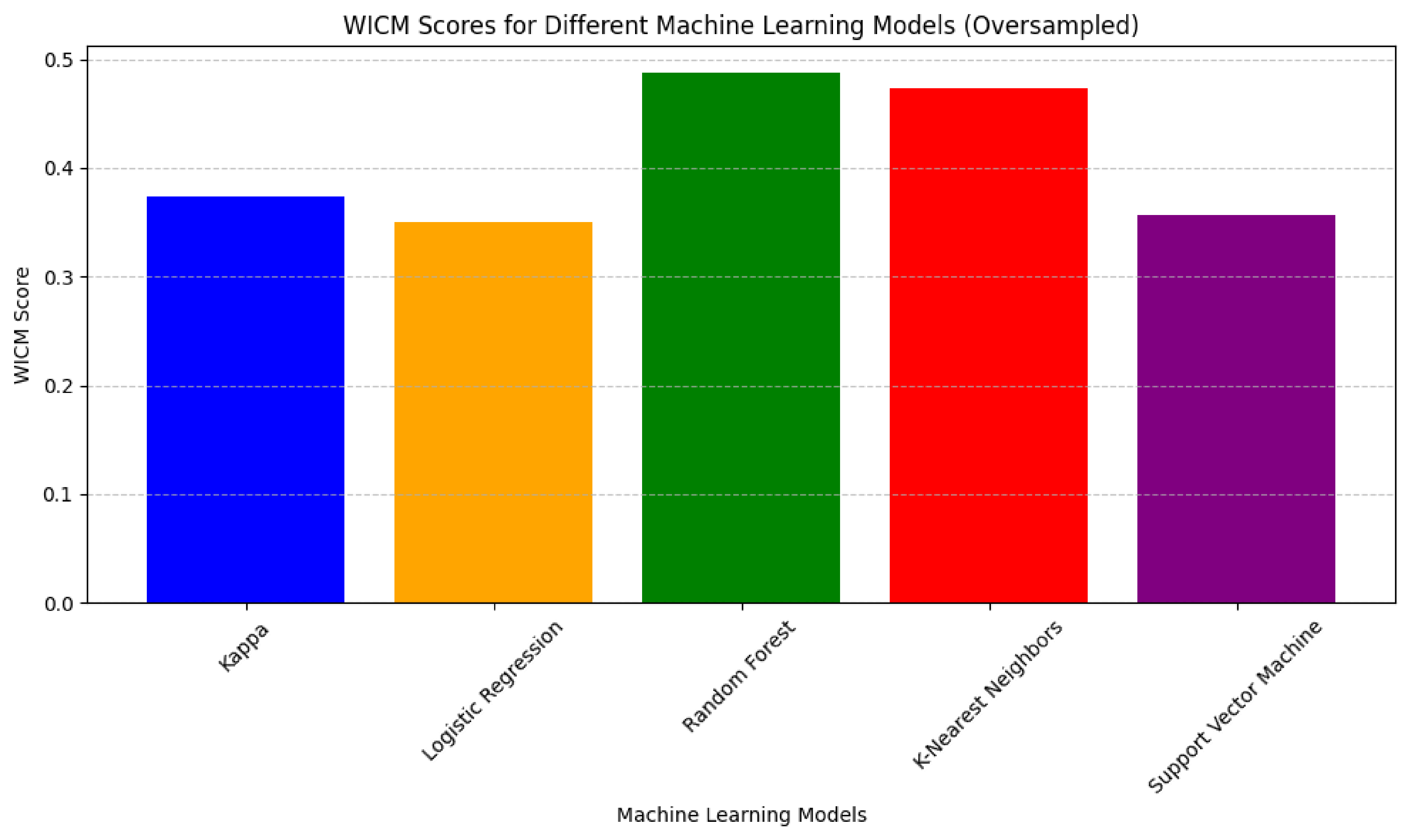
Oversampled F1 scores



Oversampled F1 scores



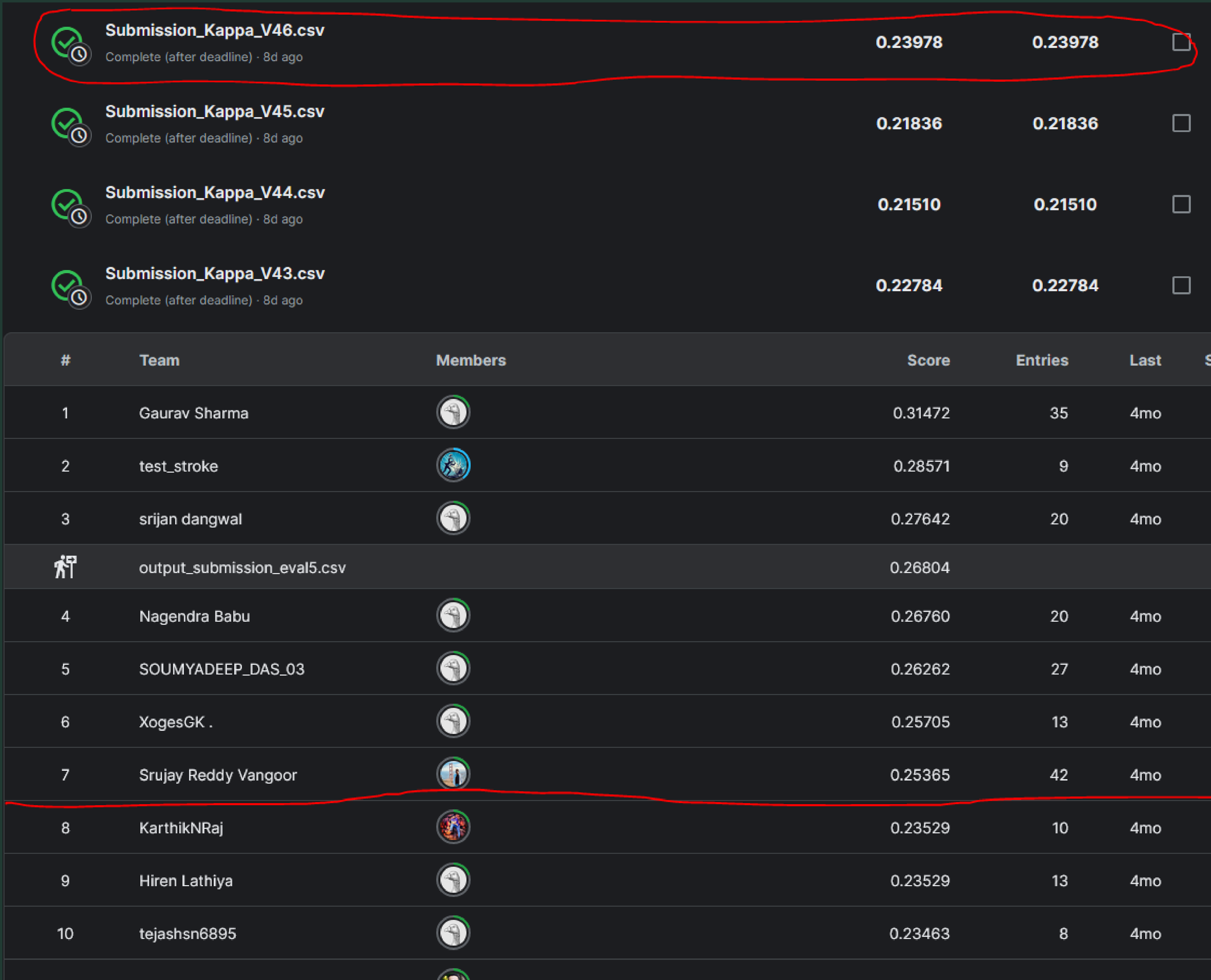
Oversampled F1 scores



kaggle

Competition

8th Place



Submission_Kappa_V46.csv
Complete (after deadline) · 8d ago
0.23978 0.23978 □

Submission_Kappa_V45.csv
Complete (after deadline) · 8d ago
0.21836 0.21836 □

Submission_Kappa_V44.csv
Complete (after deadline) · 8d ago
0.21510 0.21510 □

Submission_Kappa_V43.csv
Complete (after deadline) · 8d ago
0.22784 0.22784 □

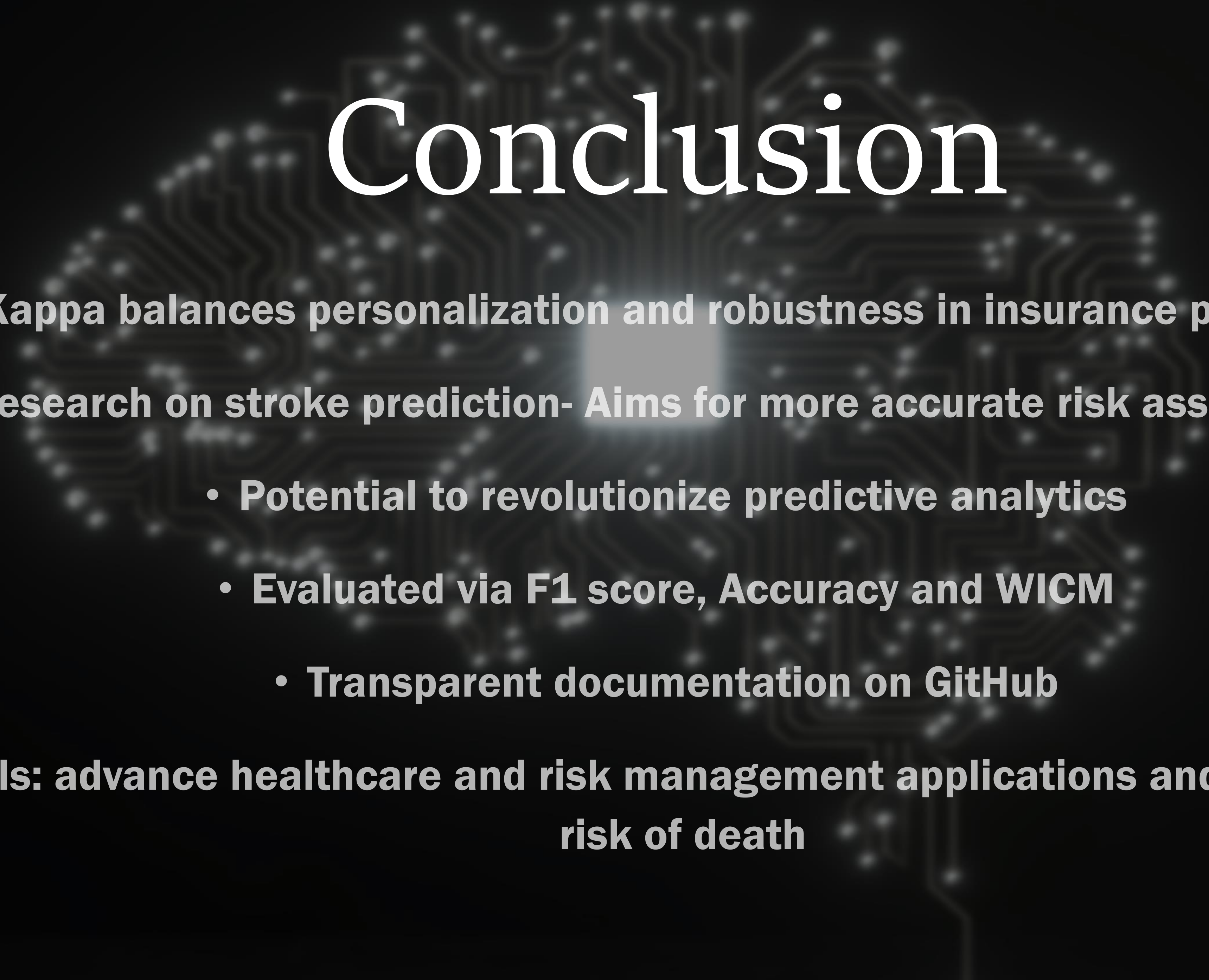
#	Team	Members	Score	Entries	Last
1	Gaurav Sharma		0.31472	35	4mo
2	test_stroke		0.28571	9	4mo
3	srijan dangwal		0.27642	20	4mo
	output_submission_eval5.csv		0.26804		
4	Nagendra Babu		0.26760	20	4mo
5	SOUMYADEEP_DAS_03		0.26262	27	4mo
6	XogesGK .		0.25705	13	4mo
7	Srujay Reddy Vangoor		0.25365	42	4mo
8	KarthikNRaj		0.23529	10	4mo
9	Hiren Lathiya		0.23529	13	4mo
10	tejashsn6895		0.23463	8	4mo

Future Works

TEST WITH MORE VARIED DATASETS

OPTIMIZE TO FIND IDEAL HYPER-PARAMETERS

CONDUCT FURTHER EXPERIMENTS



Conclusion

- Kappa balances personalization and robustness in insurance premiums
- Research on stroke prediction- Aims for more accurate risk assessments
- Potential to revolutionize predictive analytics
- Evaluated via F1 score, Accuracy and WICM
- Transparent documentation on GitHub
- Goals: advance healthcare and risk management applications and reduce the risk of death