USAGE OF MACHINE LEARNING (ML)

OBJECTIVES:

1. PREDICTING MEDIAN HOUSE PRICE USING ML REGRESSION ALGORITHM
2. PREDICTING THE OCCURRENCE OF STROKE USING ML CLASSIFICATION ALGORITHMS

BY EMMA T. (28 AUG 2023)

1. PREDICTING MEDIAN HOUSE PRICE USING ML

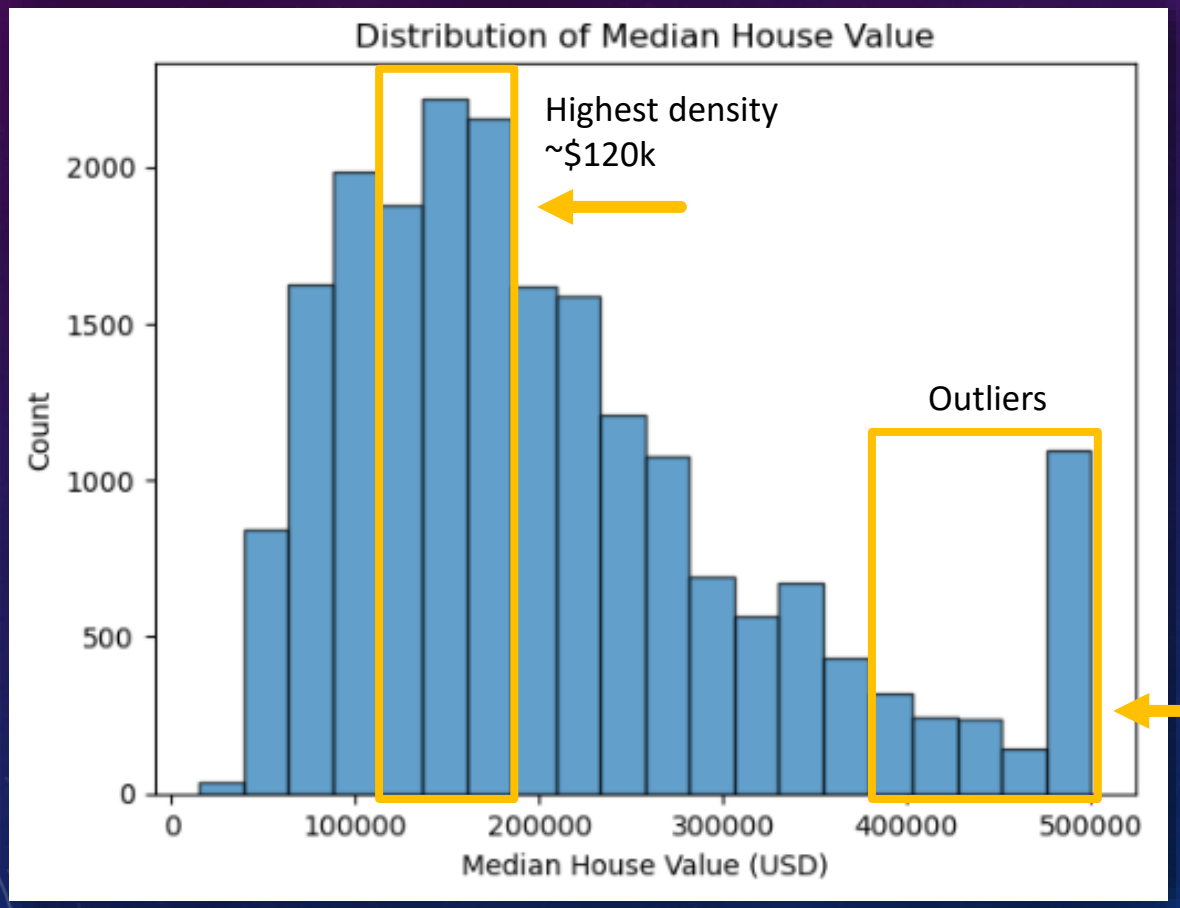
- Problem statements:
  - Are you interested in **purchasing property** or planning to **invest in real estate**? Both requires property value appraisal. In locations like Singapore, where land value has a significant impact on the market, making informed decisions becomes crucial.
  - The **market price of a house** can be **determined using Machine Learning techniques**, such as regression models, that analyze various factors like location, size, amenities, and historical sales data to predict a property's value accurately.
- In this study, California Housing dataset (from 1990 California census data) was obtained from Kaggle.
  - 10 key info was obtained from nearly 21,000 houses
  - i.e. **location, ocean proximity, house age, median income, median house value**.

Notes:

Kaggle is a platform for data science and machine learning enthusiasts, researchers, and professionals to collaborate, compete, and share insights. It provides a community-driven environment where users can find and publish datasets, explore and visualize data, and participate in machine learning competitions.

Prepared by Emma T. (28 Aug 2023)

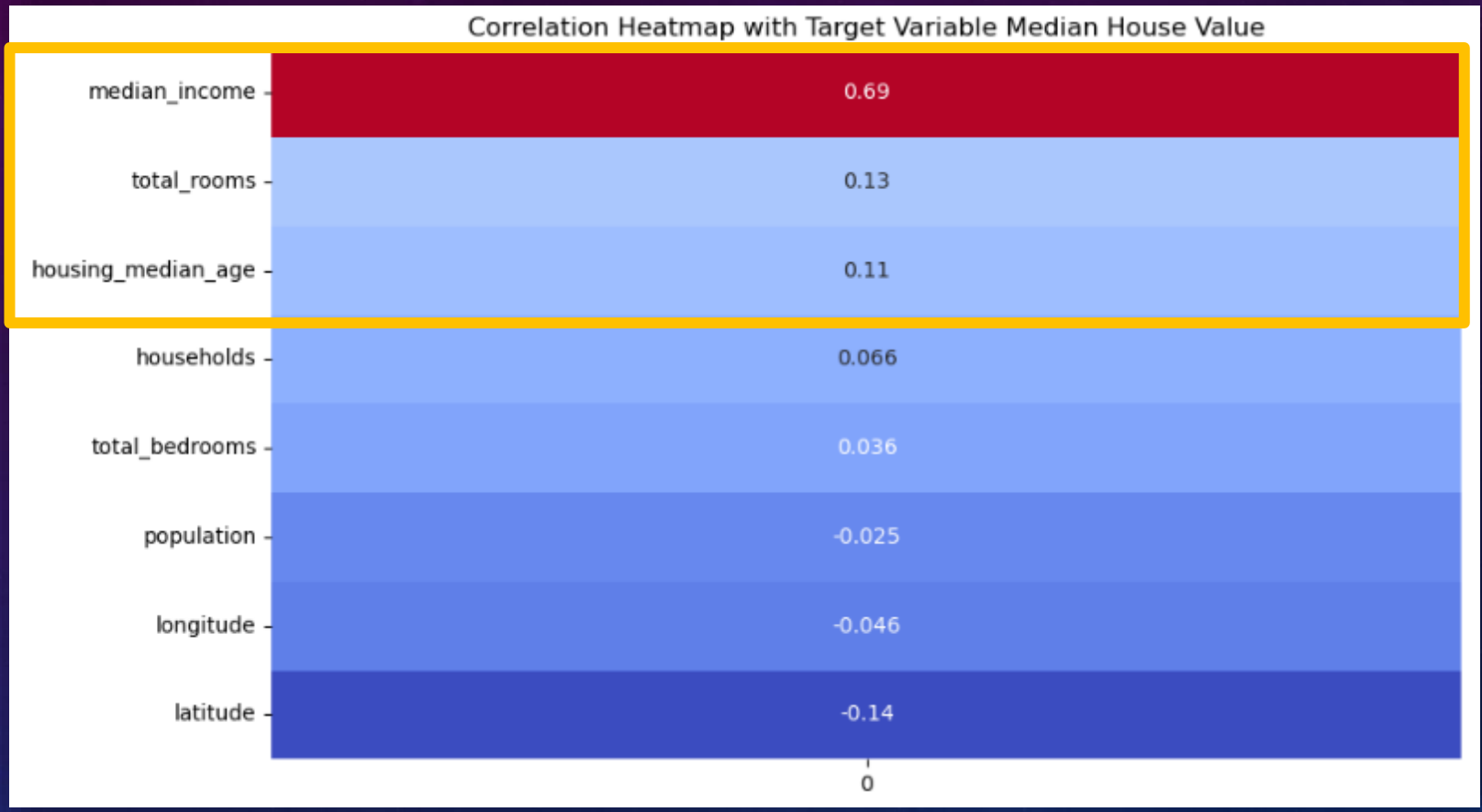
INSIGHTS ON THE CALIFORNIA HOUSING DATA (Y1990)



- Median house price range from $15k to $500k, with the **highest density** lies between **$120k to $180k**.
- Presence of approx. **5% outliers** which shows house value on the pricier side up to $500k.

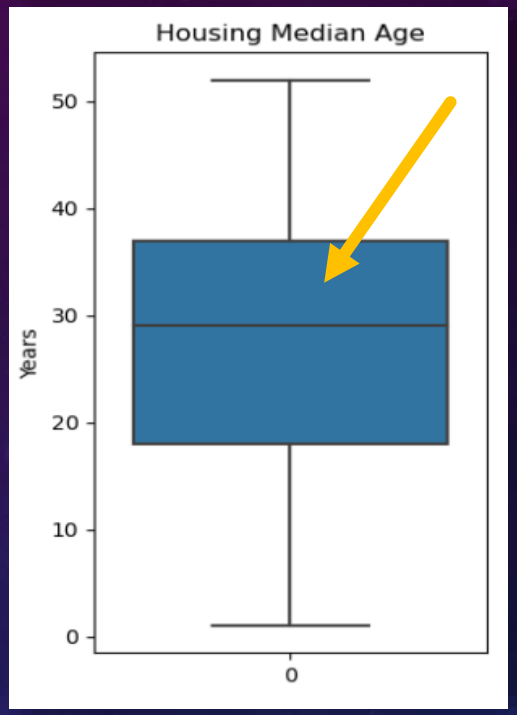
Prepared by Emma T. (28 Aug 2023)

INSIGHTS ON THE CALIFORNIA HOUSING DATA (Y1990) – CONT.

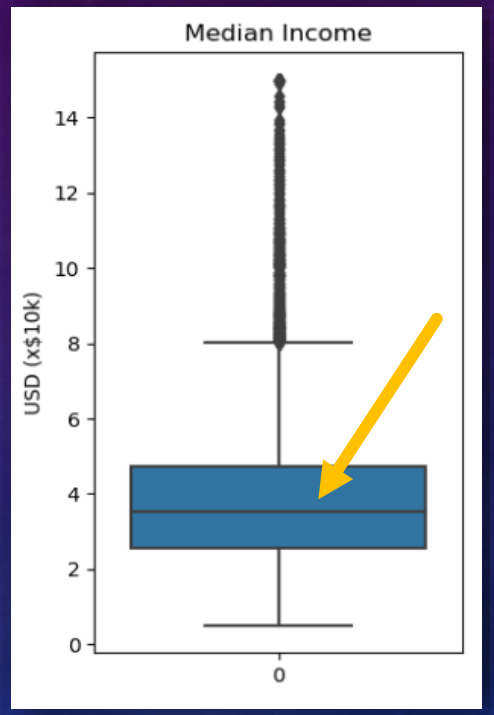


- The **top 3 info** beside ocean proximity that have the **highest correlation** to house value in descending orders are :
  - **Income**
  - **Number of rooms**
  - **House age**

INSIGHTS ON THE CALIFORNIA HOUSING DATA (Y1990) – CONT.



Majority house age between 20 to 30+ years



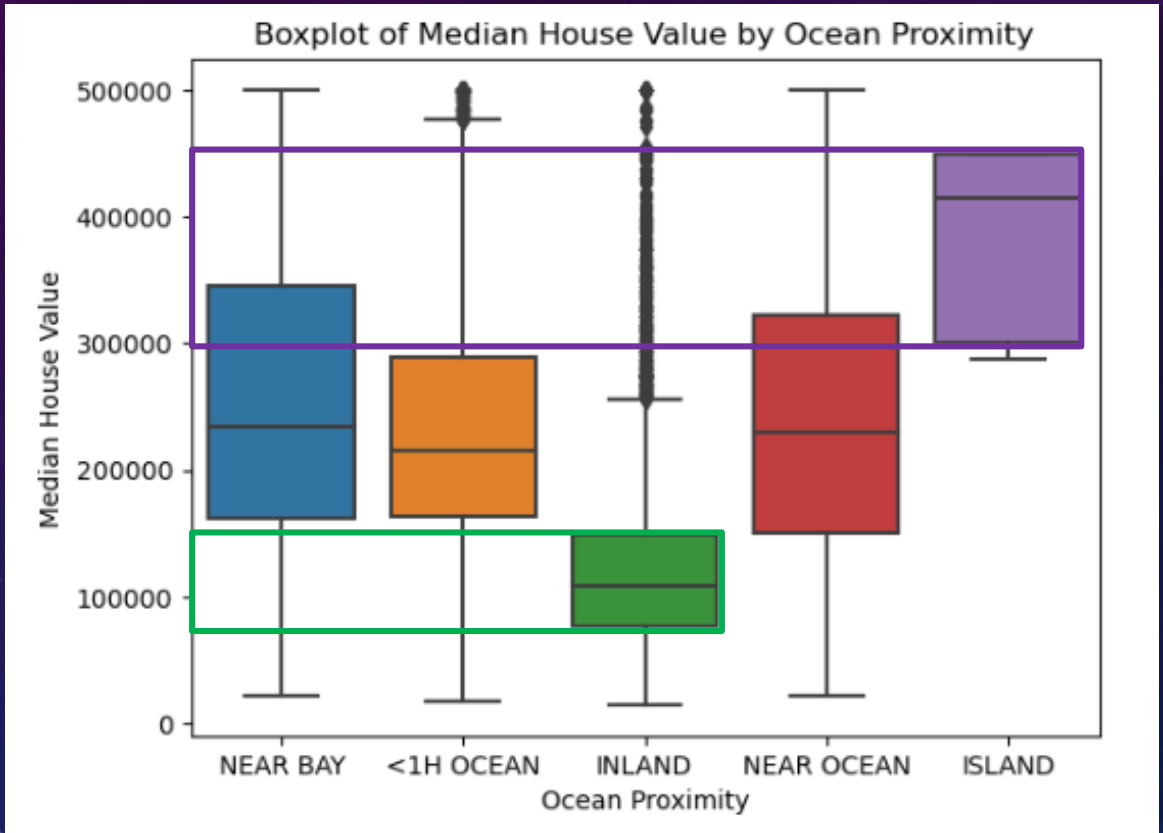
Majority median income of the household is between **$20k -$50k yearly**.

- At a hindsight, a family earning $35k yearly is likely to be able to afford a $120k house of approx. 30 years old.

1. PREDICTING MEDIAN HOUSE PRICE USING ML – CONT.

INSIGHTS ON THE CALIFORNIA HOUSING DATA (Y1990) – CONT.

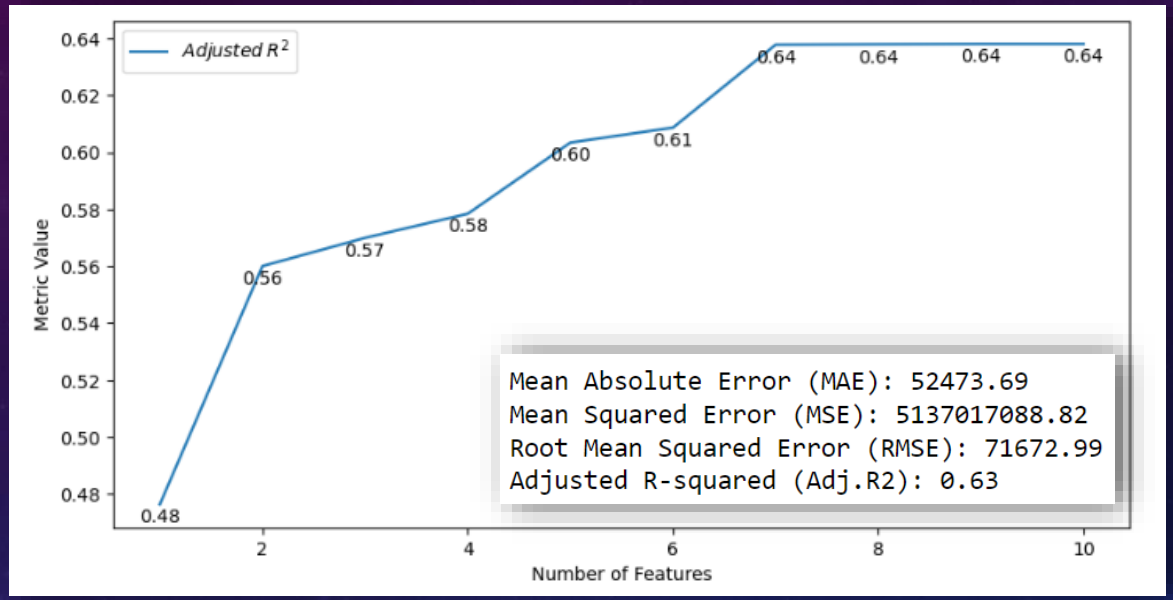
So, does an ocean view means pricier house?



- From data, **house on island** fetch **highest value** (**$300k - $450k**).
- Followed closely by house near bay, near ocean and <1 hour drive from ocean respectively.
- In general, **inland house** is **cheapest** in general, from **$90k**, with some unique cases of reaching $500k.
  - These outliers could be properties with exceptionally large sizes or unique features that justify their higher prices.

Prepared by Emma T. (28 Aug 2023)

MACHINE LEARNING RESULT – LINEAR REGRESSION MODEL



- Model was **trained using Linear Regression (LR) Algorithm** – using **forward feature selection** of all info except house value and cross-validated.
- Model accuracy of a LR can be checked using the evaluation metrics. MSE value show that a large error is to be expected.
- From MAE & RMSE, the model's **predictions are off by $50-$70k** in terms of median house value.

Prepared by Emma T. (28 Aug 2023)

MACHINE LEARNING RESULT – LINEAR REGRESSION MODEL

	Training R2	Test R2	Training RMSE	Test RMSE
1	0.635251	0.638116	69530.642205	70043.796111
2	0.640800	0.613917	69178.226996	71616.559595
3	0.641102	0.613560	69376.904635	70674.581745
4	0.631952	0.651167	70009.771739	68134.530496
5	0.630981	0.654823	69981.333296	68242.216508

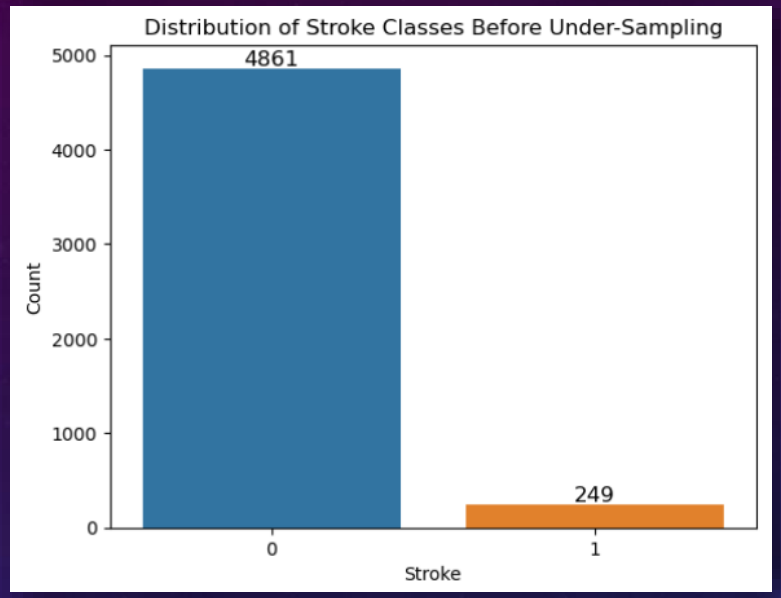
- **Similar R2 training and testing results** indicate the model's **consistent** and **effective** generalization to new data without overfitting or underfitting.
- R2 value suggests that the model explains about **63% of the variability** in the target variable which is good.
- Model prediction may be further improved with more data on house size & amenities etc

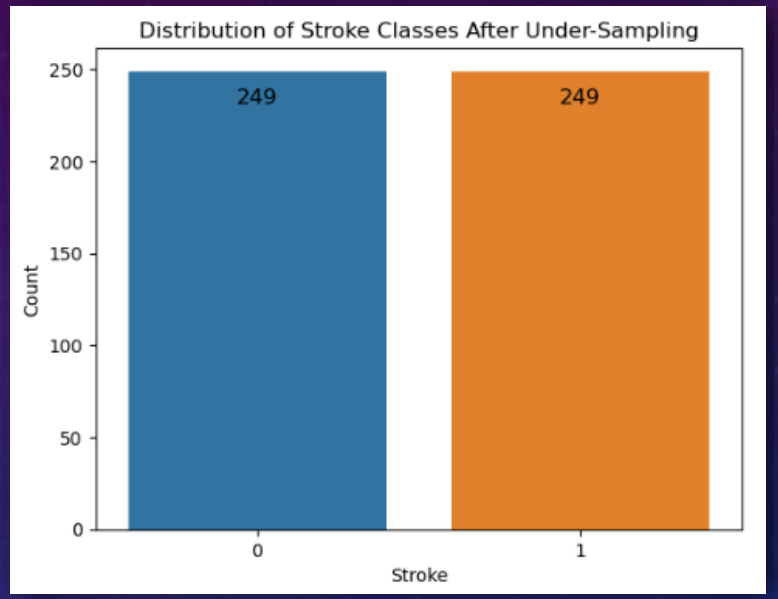
2. PREDICTING THE OCCURRENCE OF STROKE

- Problem Statements:
  - In our community, a vital concern is **identifying individuals at risk of strokes** before it's too late, as strokes can lead to severe outcomes.
  - For doctors, this can **help to set priority** on which patients to focus on **based on likelihood of stroke occurrence**.
  - This is where machine learning steps in – a proactive approach that analyzes comprehensive medical and lifestyle data. By uncovering hidden patterns and connections, machine learning enhances the accuracy of identifying risks, offering a promising way to save lives within our community.

- Health care dataset was obtained from Kaggle, provided by Fedesoriano in 2020.
  - 11 key info was obtained from over 5,000 patients
  - i.e. **age, history of hypertension, heart disease** and **smoking status** are among the key info.

INSIGHTS ON THE HEALTH CARE DATA (Y2020)





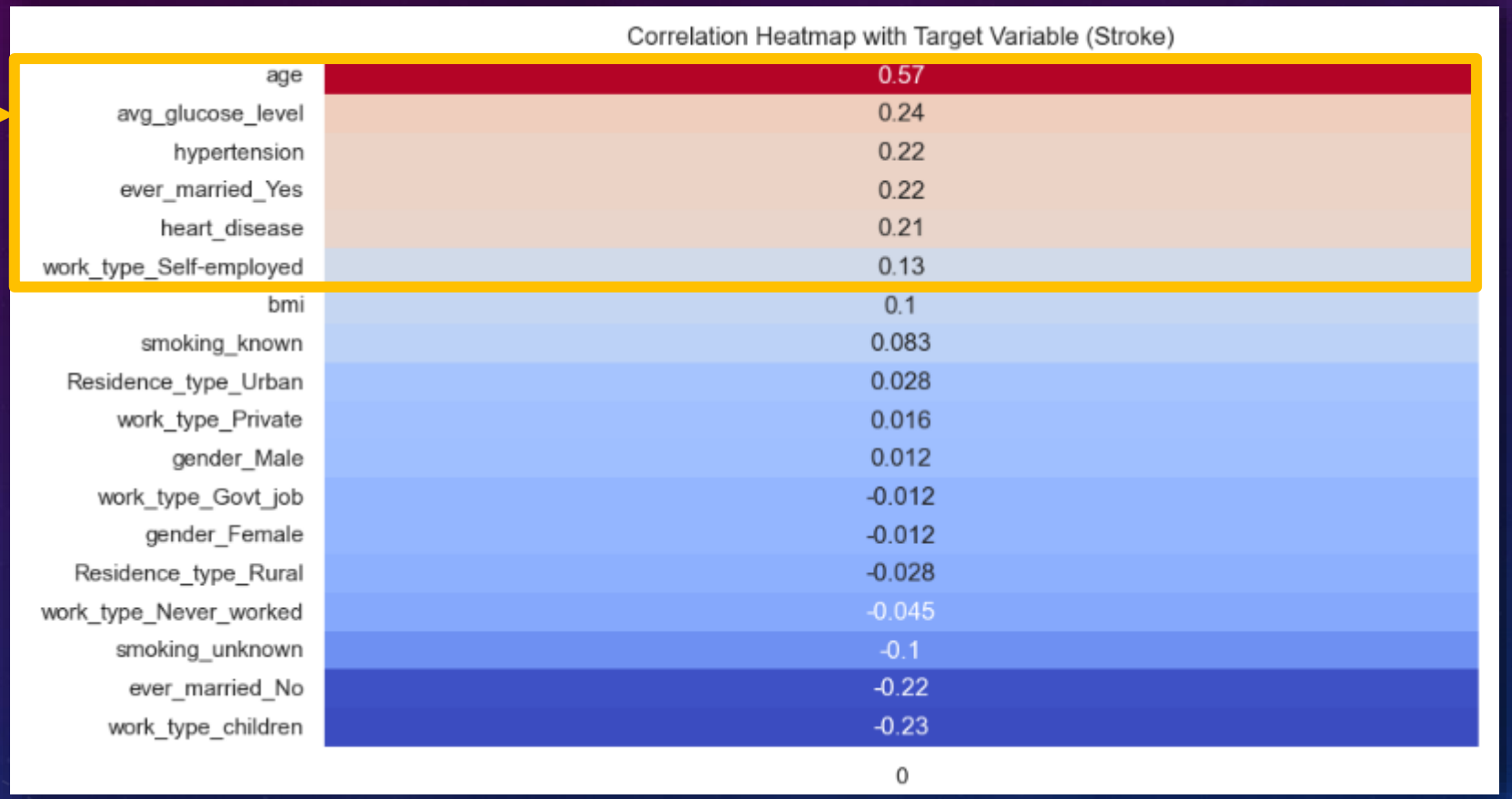
Notes:

- While experimenting with class weight balancing, the model frequently exhibited a bias towards predicting no strokes, rendering it ineffective.
- The utilization of oversampling techniques like SMOTE was avoided to prevent the introduction of synthetic patterns that could potentially distort the natural distribution of data.

- The **original dataset** was **severely skewed towards no stroke patient (95%)**. In stroke prediction through machine learning, having a balanced dataset is indeed essential to prevent bias and ensure accurate model performance and real-world applicability.
- **Undersampling method** is applied to balance the dataset. While 90% of data was sacrificed, the trade-off is justified to effectively addresses the under-represented class.

Prepared by Emma T. (28 Aug 2023)

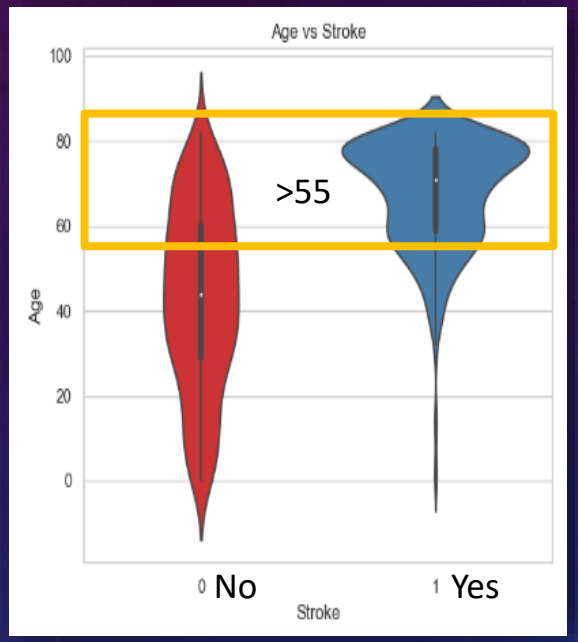
INSIGHTS ON THE HEALTH CARE DATA (Y2020) – CONT.

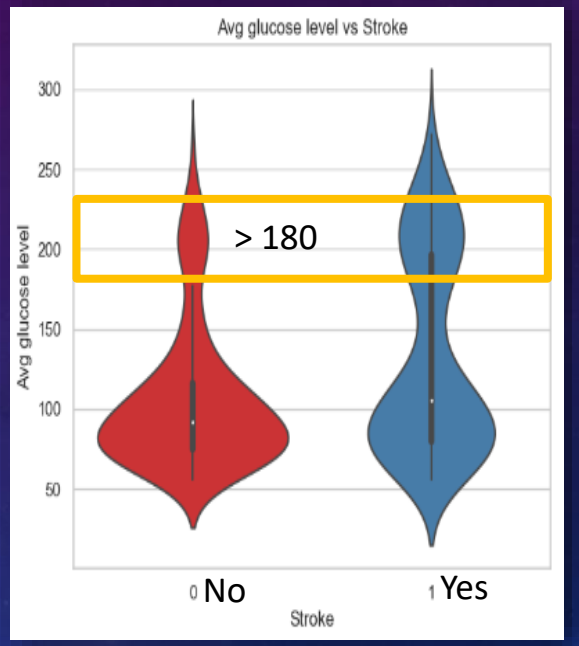


- The **top 6** info that have the highest correlation to stroke in descending orders are :
  - **Age**
  - **Average glucose level**
  - **Hypertension & Being married**
  - **Heart disease**
  - **Being self employed**

INSIGHTS ON THE HEALTH CARE DATA (Y2020) – CONT.

A closer look at Top 5 info with highest correlation

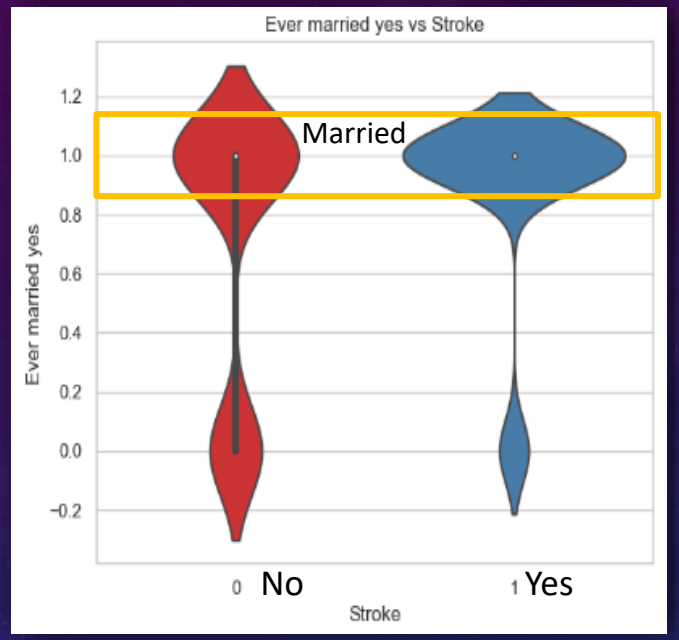


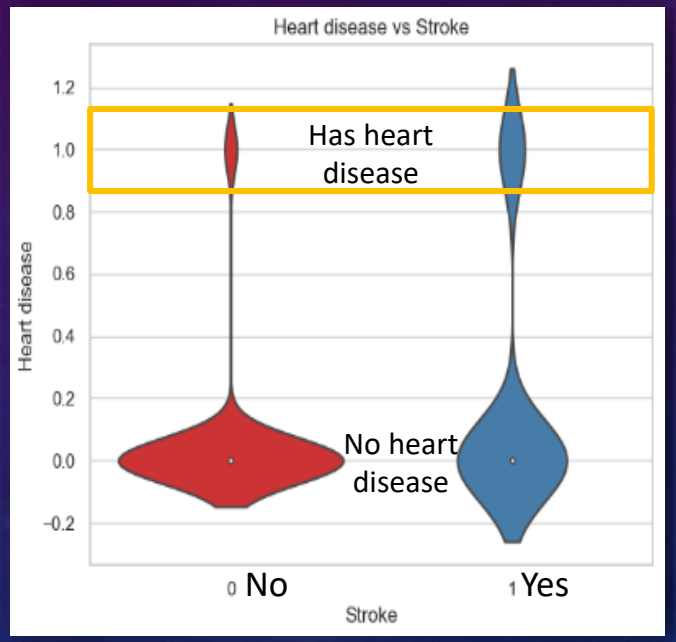


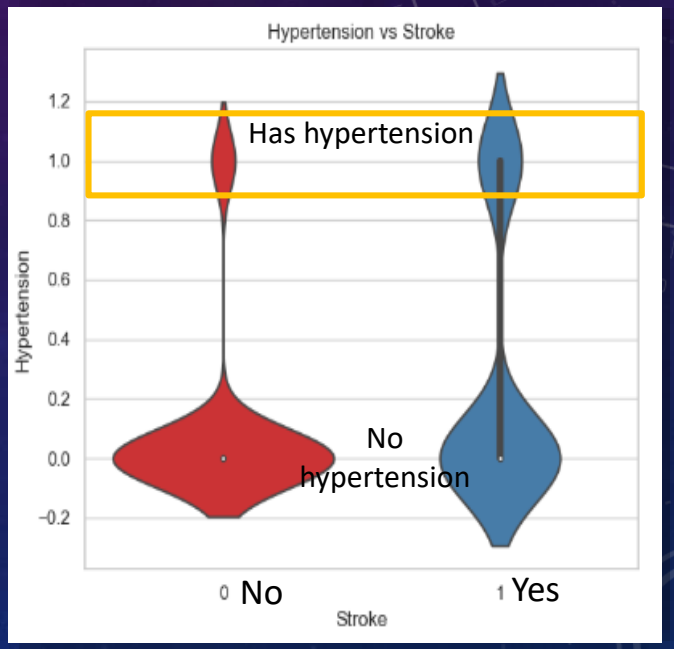
- Upon closer examination, individuals **over the age of 55** and **average glucose level >180** are notably more **prone to experiencing a stroke** (wider spread observed from violin plot).

INSIGHTS ON THE HEALTH CARE DATA (Y2020) – CONT.

A closer look at Top 5 info with highest correlation



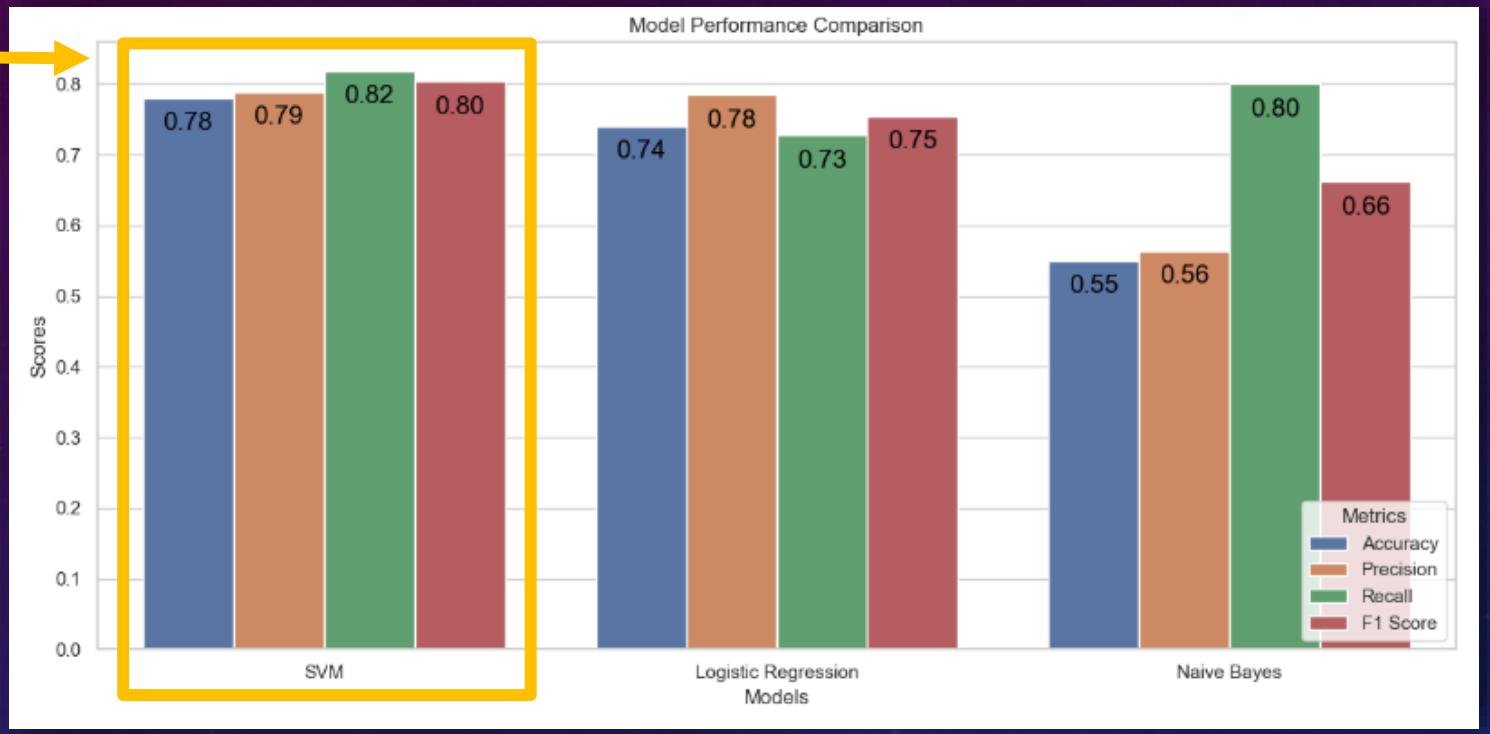




- While **being married** and having a **history of hypertension** and **heart disease** does not guarantee a stroke, data shows that these does **elevate the chances**.

Prepared by Emma T. (28 Aug 2023)

MACHINE LEARNING RESULT



Quick study on metrics used:

- Accuracy provides an overall gauge of stroke prediction model's performance
- Precision helps avoid false positive predictions
- Recall helps avoid missing actual stroke cases
- F1 score offers a way to strike a balance between precision and recall based on study specific priorities.

- Model was trained using 3 commonly used algorithms; Support Vector Machine (SVM), Logistic Regression and Naïve Bayes.
  - The best performing model is **SVM (scaled)** with **highest score** across all 4 metrics tested with **78% accuracy on test data**.
- Model prediction may be further improved with more data on positive stroke cases, using different data processing method and streamline feature (info) selections.

Prepared by Emma T. (28 Aug 2023)

End of Presentation. Thank you.

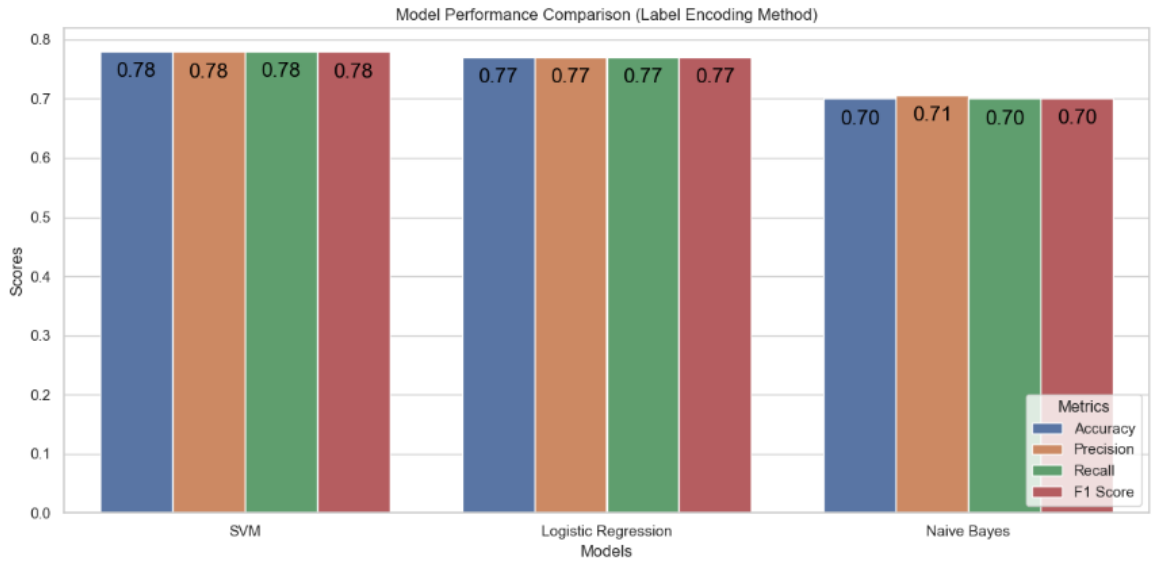
Q&As Session

Dataset, codes & models are available in GitHub Link:
https://github.com/EmmaT0611/mp2_supervisedmachinelearning

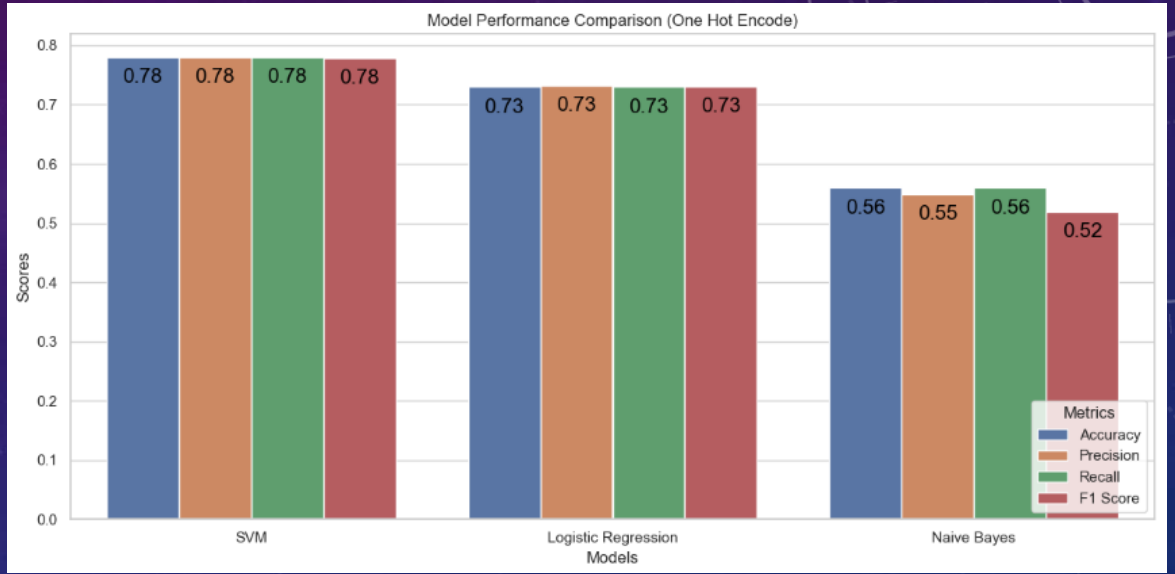
Prepared by Emma T. (28 Aug 2023)

UPDATES FROM PREVIOUS ANALYSIS - TECHNICAL VISUALISATION

Label Encoding Method



One Hot Encoding Method

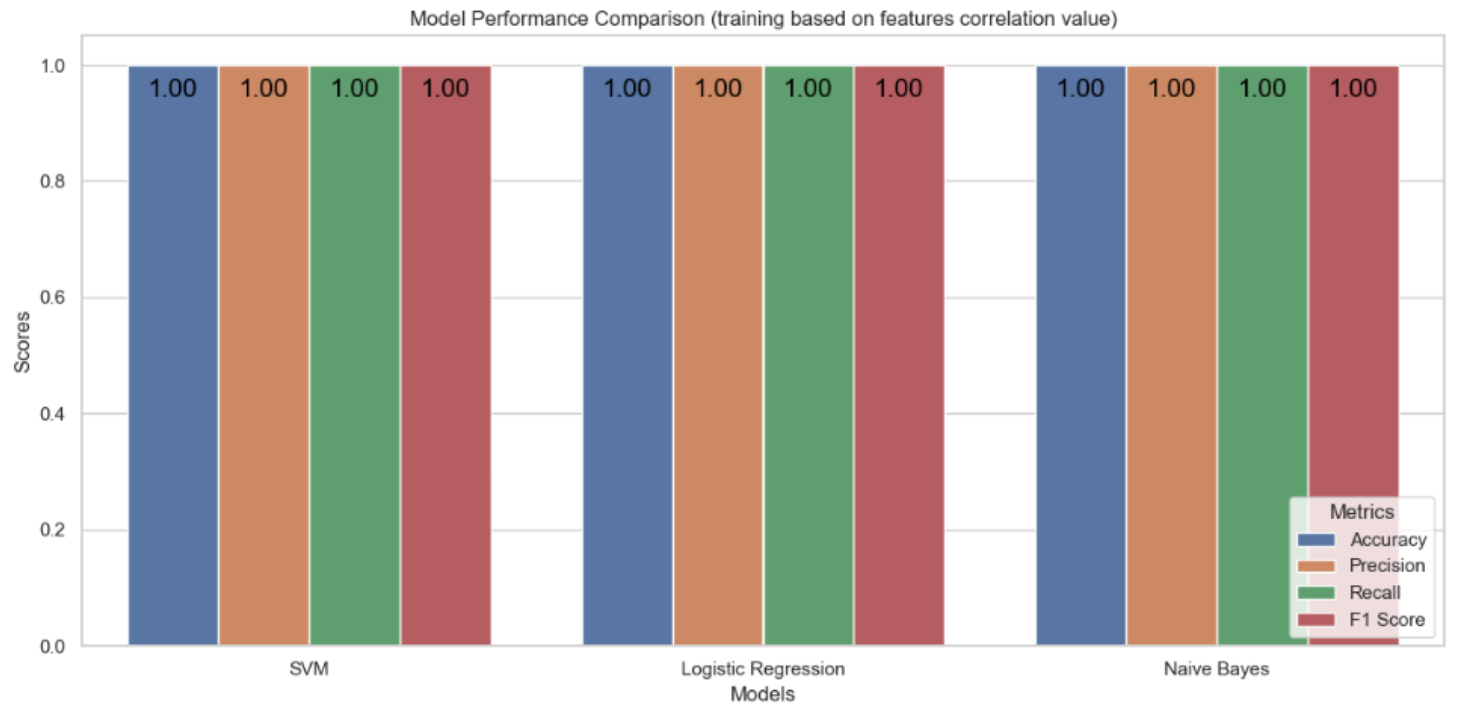


- With the exception of SVM, model trained using Label Encoding score higher accuracy, precision, recall and F1 for Logistic Regression and Naïve Bayes.

Prepared by Emma T. (28 Aug 2023)

UPDATES FROM PREVIOUS ANALYSIS - TECHNICAL VISUALISATION

Model trained based on features correlation value



Perfect accuracy was obtained from **training the model** starting from the **most relevant features** that have the highest correlation with target variable, regardless of the encoding method used.

- Cross-validation, hyperparameter tuning(GridSearchCV), changing random states and random undersampling has been performed to ensure that the model's performance is consistent and not simply a result of chance.
- Given the dataset is relatively small, achieving perfect accuracy could be possible due to limited data. It's essential to evaluate the model on a larger dataset or perform further analysis to ensure robustness.

Prepared by Emma T. (28 Aug 2023)

APPENDIX 1.

Key Info from California Housing Dataset Explained:

1. Longitude: A measure of how far west a house is; a higher value is farther west
2. Latitude: A measure of how far north a house is; a higher value is farther north
3. Housing Median Age: Median age of a house within a district; a lower number is a newer building
4. Total Rooms: Total number of rooms within a district
5. Total Bedrooms: Total number of bedrooms within a district
6. Population: Total number of people residing within a district
7. Households: Total number of households, a group of people residing within a home unit, for a district
8. Median Income: Median income for households within a district of houses (measured in tens of thousands of US Dollars)
9. Median House Value: Median house value for households within a district (measured in US Dollars)
10. Ocean Proximity: Location of the house with respect to ocean/sea

APPENDIX 2.

Key Info from Health Care Dataset Explained:

- Gender: "Male", "Female" or "Other"
- Age: age of the patient
- Hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- Heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- Ever_married: "No" or "Yes"
- Work_type: "children", "Govt_job", "Never_worked", "Private" or "Self-employed"
- Residence_type: "Rural" or "Urban"
- Avg_glucose_level: average glucose level in blood
- Bmi: body mass index
- Smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
- Stroke: 1 if the patient had a stroke or 0 if not
- *Note: "Unknown" in smoking_status means that the information is unavailable for this patient

REFERENCES

- House Price Prediction Using Linear Regression (2021) by Simran Kaur Link: https://linuxhint.com/house-price-prediction-linear-regression/
- Predicting House Prices with Linear Regression | Machine Learning from Scratch (2019) by Venelin Valkov Link: https://towardsdatascience.com/predicting-house-prices-with-linear-regression-machine-learning-from-scratch-part-ii-47a0238aeac1

- Analyzing the Performance of Stroke Prediction using ML Classification Algorithms (2021) by Gangavarapu Sailasya & Gorli L Aruna Kumari.
  Link:https://thesai.org/Downloads/Volume12No6/Paper_62Analyzing_the_Performance_of_Stroke_Prediction.pdf
- Stroke Disease Detection and Prediction Using Robust Learning Approaches (2021) by Tahia Tazin and team. Link: https://www.hindawi.com/journals/jhe/2021/7633381/
- Stroke Risk Prediction with Machine Learning Techniques (2022) by Elias Dritsas* and Maria Trigka.
  Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9268898/