



BREAST CANCER PREDICTION

A series of several thin, white, parallel diagonal lines extending from the bottom right towards the top right of the image, creating a sense of movement and modern design.

BUSINESS PROSPECT

- ▶ Breast Cancer is the 2nd highest type of cancer after lung cancer amongst the women.
 - ▶ A special emphasis has been put on predicting the likelihood or probability of cancer which will help in providing personalized care and remedies to all the patients thereby accordingly focusing the attention towards serious cases and thereby better handling of the ailment.
 - ▶ The plan is to create a machine learning model to predict this probability.
- 
- A series of white lines of varying lengths and orientations are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

DATA UNDERSTANDING

- ▶ We will use the Breast Cancer patients' data of Wisconsin city.
 - ▶ This was collected from the radiology and cancer detection centres of various hospitals spread across the city of Wisconsin.
 - ▶ This will be a csv file containing different attributes like clump thickness, cell size, cell shape, marginal adhesion, bland chromatin, etc which will be used to predict the cancer.
- 
- A series of white lines of varying lengths and orientations are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

DATA CLEANING AND PREPROCESSING

- ▶ We visualize the weightage of each attribute and accordingly drop the ones that least effective to abstain from overfitting the model.
- ▶ We also look for the values in each columns and the ones with invalid values are dropped or if the attribute is of prime importance we replace those invalid values.



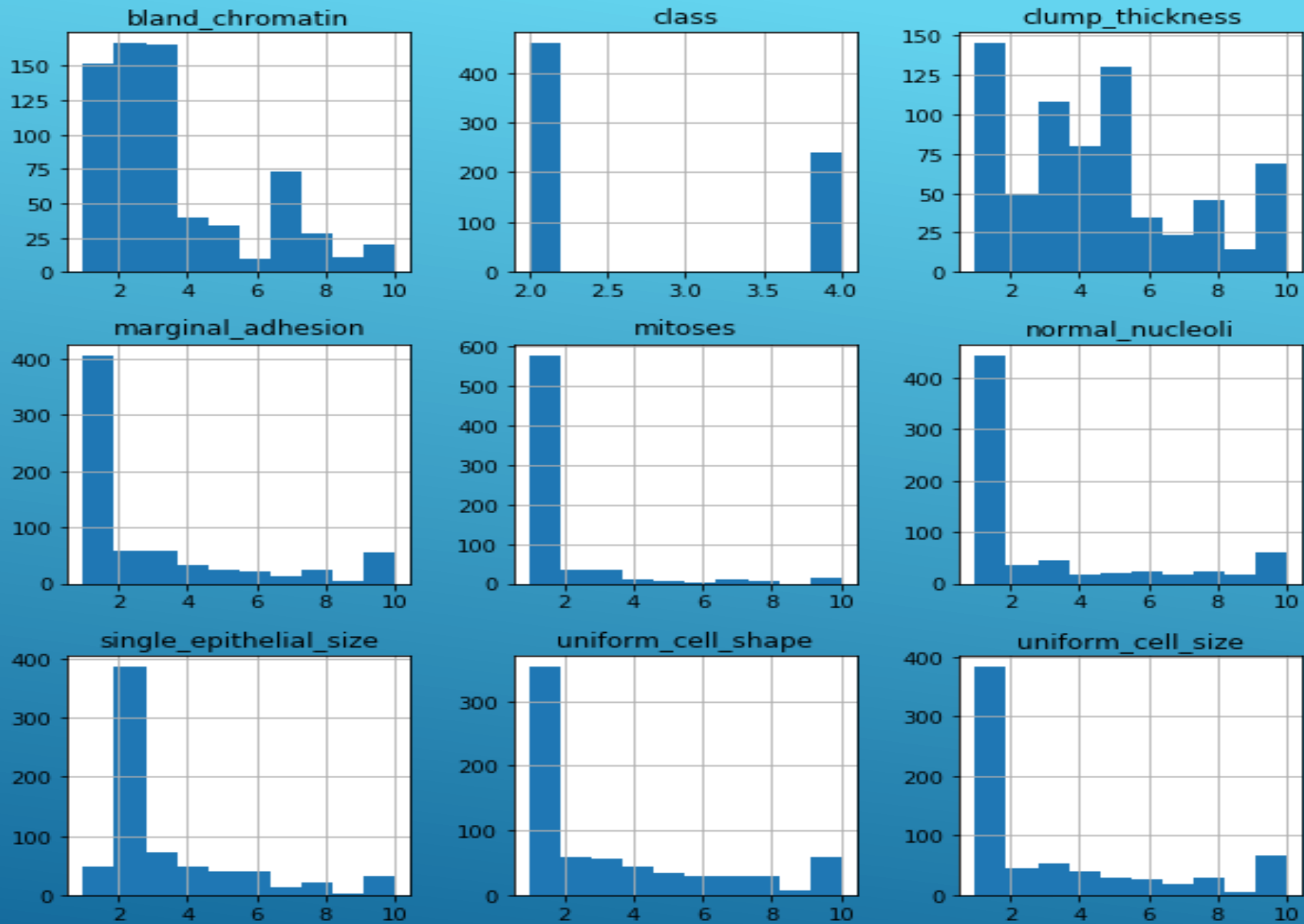


Fig: Attributes Distribution

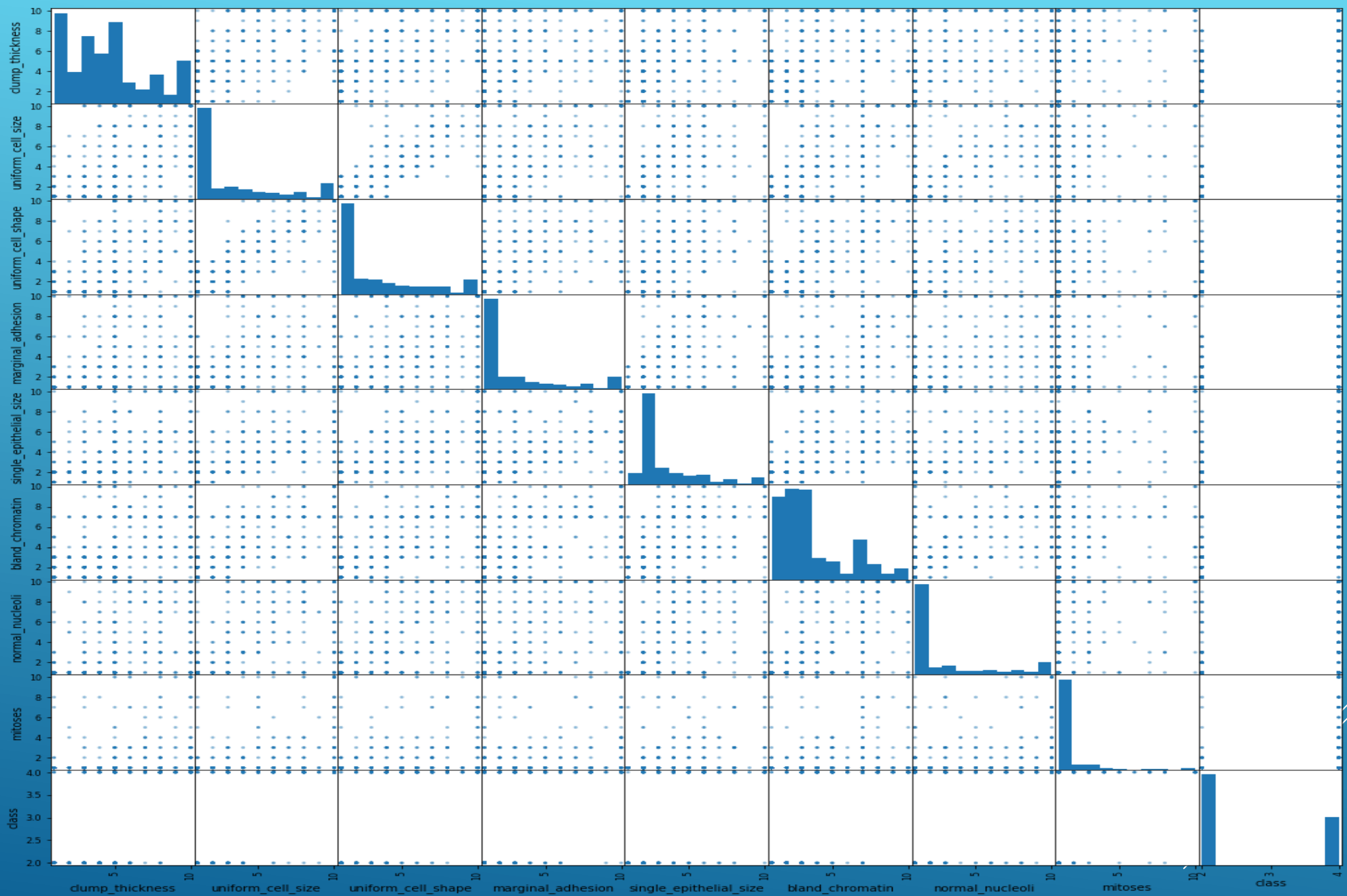


Fig: Attributes Distribution

MODELLING

- ▶ In this project for the prediction of breast cancer we planned to go ahead with 2 models-

- i) K Nearest Neighbours (KNN)

- ii) Support Vector Machine (SVM)

- ▶ We split the dataset into training and testing dataset and used the 'Class' column to make the model.
- ▶ After careful visualization we came with $k=5$ as the optimum value and prepare the KNN and SVM model with the training dataset.


CALCULATING ACCURACY AND PREDICTION

- ▶ We then tested both the KNN and SVM models with the testing dataset to calculate the accuracy. For the sake of better understanding we calculated out the precision, recall f1-score and the support of each models.
- ▶ Based on the values obtained we could clearly see the SVM with a better precision and recall score with the rest parameters remaining same. We thus could finally chose a better and close model and go ahead with the final prediction which will act a precedent for further predictions and patient analysis.

RESULT


- ▶ For the sake of proper prediction we calculated the precision, recall and F1- score of both the KNN and SVM model with the SVM getting a score of 1 against the 0.98 and 0.93 against 0.96 of KNN algorithm for precision and recall respectively. But with the macro average of 0.97 of SVM compared to 0.96 for KNN in recall we went ahead with SVM as a more accurate algorithm for this model.
- ▶ Using the SVM algorithm we finally predicted the probability of getting a breast cancer depending on various attributes for the particular unknown patient and was found to be 0.9571 or 95.71% or very highly likely.

DISCUSSION

- ▶ In order to prevent overfitting we also had to visualize various attributes and select the one that had more weightage.
 - ▶ We came up with 2 approaches- KNN and SVM and we could model using any one of them but to be on a better side we first tried to analyze both the models and finally go ahead with the best one.
 - ▶ On comparing we found SVM to be better one and thus went ahead with this algorithm for our machine learning model.
- 

CONCLUSION

We were therefore successfully able to create a machine learning model using the SVM algorithm to accurately predict the probability of a patient to get breast cancer based on attributes by training our model with attributes like- clump thickness, uniform cell shape, marginal adhesion, bland chromatin, single epithelial size, mitosis etc.

Several white lines of varying lengths and slopes are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.