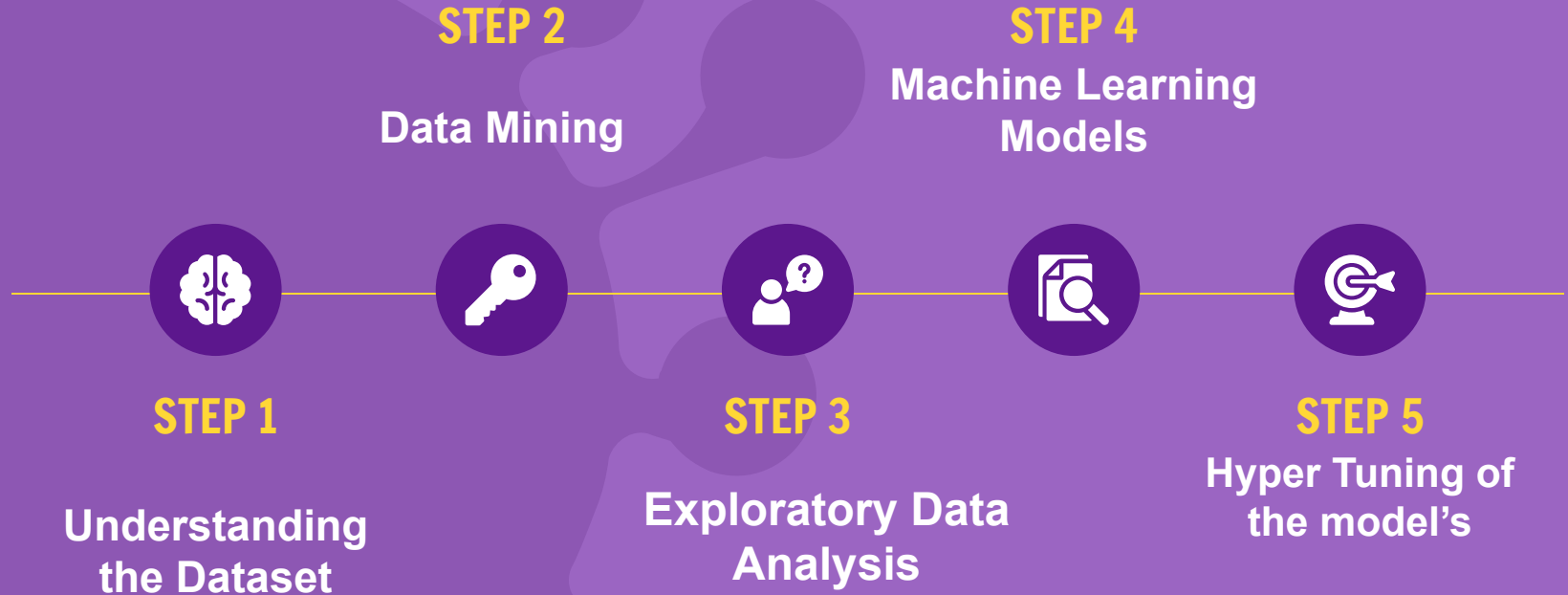


# Covid 19 Prediction Using Supervised Machine Learning

Group Member : = Pritam Channawar  
Chandani Thumar  
Manas Vani

# LET'S START



# Problem Overview

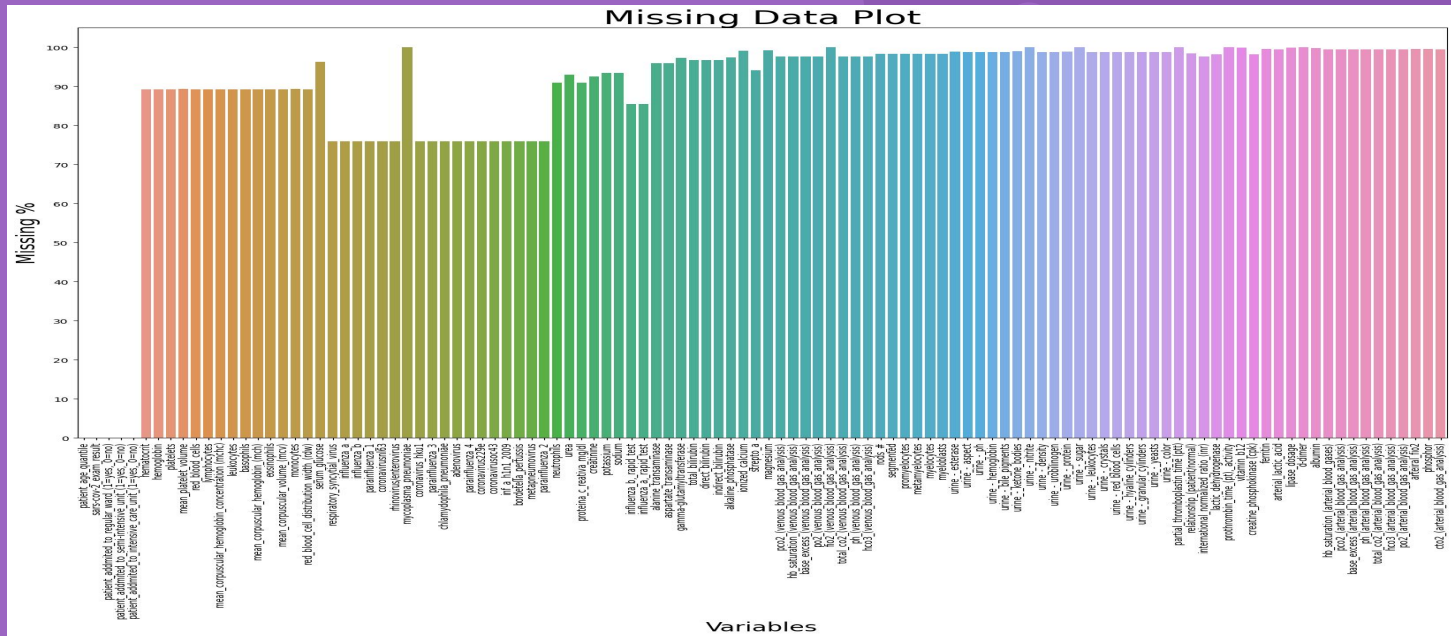
- The problem includes the test result for SARS-Cov-2 (positive/negative) among suspected COVID-19 cases based on the results of laboratory tests commonly collected for a suspected COVID-19 case during a visit to the emergency room.
- The ultimate goal of this project is to develop a model that can help healthcare professionals identify suspected COVID-19 cases with a high degree of accuracy, enabling them to take appropriate measures to prevent the spread of the virus and provide timely treatment to those who test positive.
- This can be achieved by using a machine learning model that can analyze the patterns and trends in the available data to make accurate predictions about the test results.

# Understanding the Dataset :

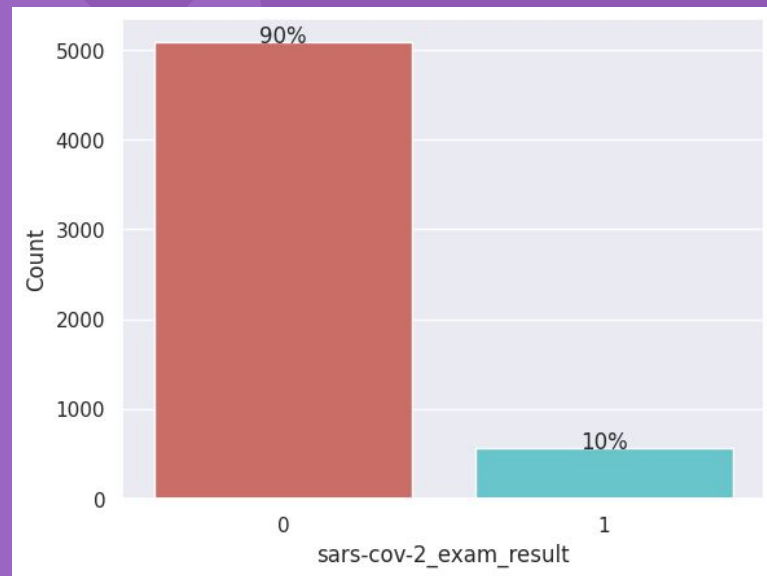
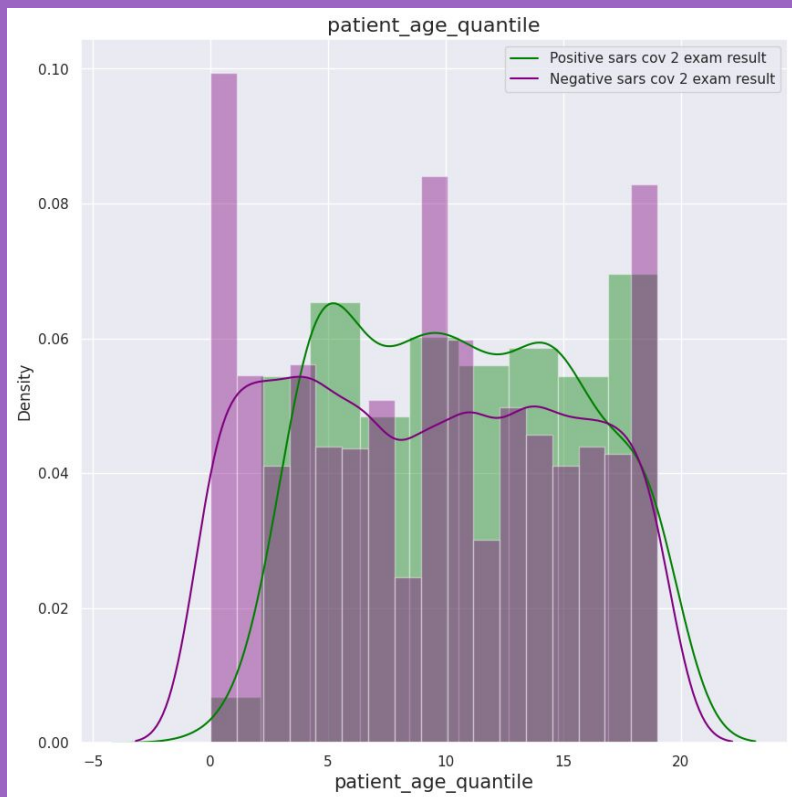
- The dataset contains anonymized clinical data of patients who visited the Hospital Israelita Albert Einstein in São Paulo, Brazil, and had samples collected to perform the SARS-CoV-2 RT-PCR and other laboratory tests.
- The dataset contains 5644 rows and 111 columns and it have 70 columns as datatype float, 4 columns as int and 37 column as object.
- The purpose of this dataset is to develop a predictive model that can assist in the diagnosis of COVID-19 cases among suspected cases . The data has been standardized to have a mean of zero and a unit standard deviation, and all patient information has been anonymized following international best practices and recommendations.
- This dataset is valuable for developing a predictive model that can help in the diagnosis of COVID-19 cases, especially in situations where testing is limited or delayed.

# Missing Data

This dataset has 88.9% missing values



# Patient age quantile by sars cov2 exam result





# Data Cleaning

- The names of attributes were trimmed to reduce unnecessary whitespace.
- Binary zeros and ones were used to represent categorical variables stated in various ways.
- Replaced null values with -999
- Negative and positive target classes were substituted with 0 and 1 for covid negative and positive respectively.
- The flu variables that were modified from 'detected' and 'not detected' to 'present' and 'absent' were also subjected to the same procedures
- Attributes with missing values of greater than 98 percent were discarded.



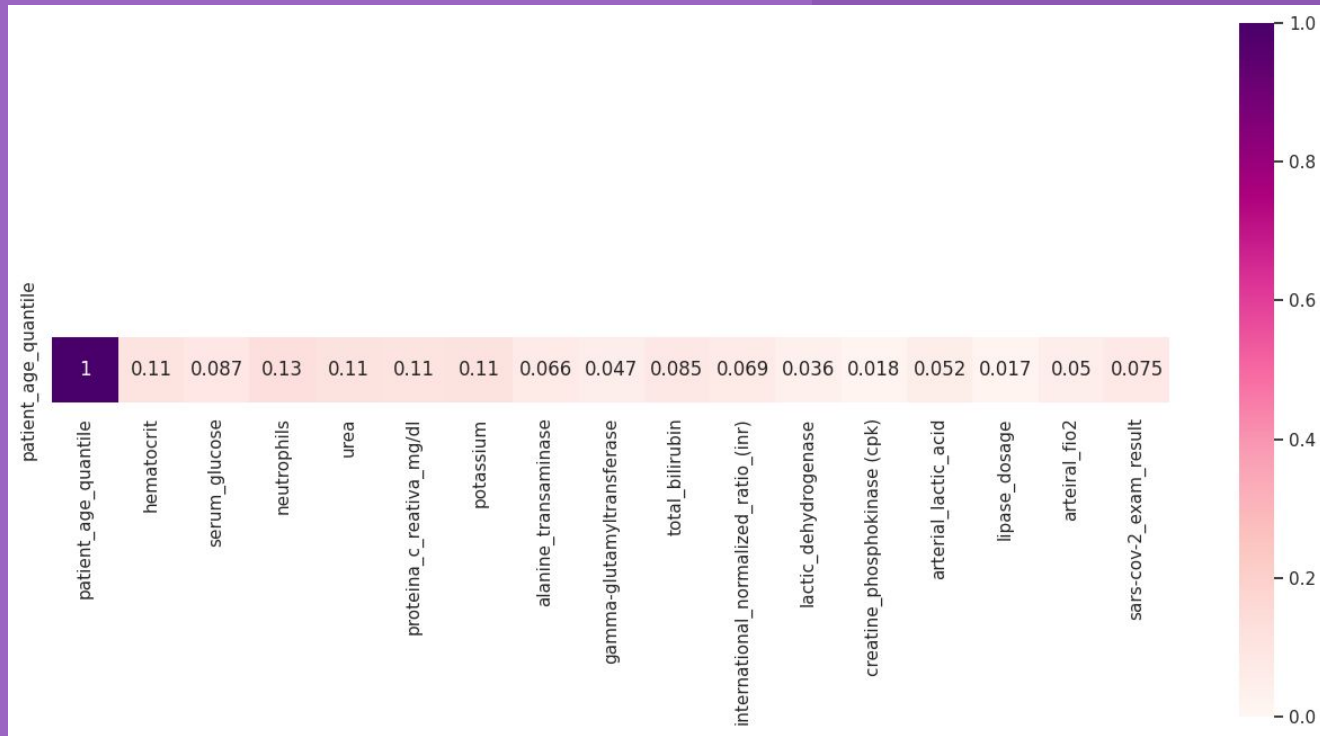
# Data Cleaning

- Imputations were avoided since they might affect the results, and any records with more than 10% missing column values were removed from the datasets.
- we were able to construct two separate databases from which to run the modelling process.
- The first subset was constructed using features linked to the full blood count, age quantile, and three categorical variables reflecting the hospitalization and the target variable ('sars cov2')... RBC, WBC, and platelets make up three of the clinical spectrum's independent properties.

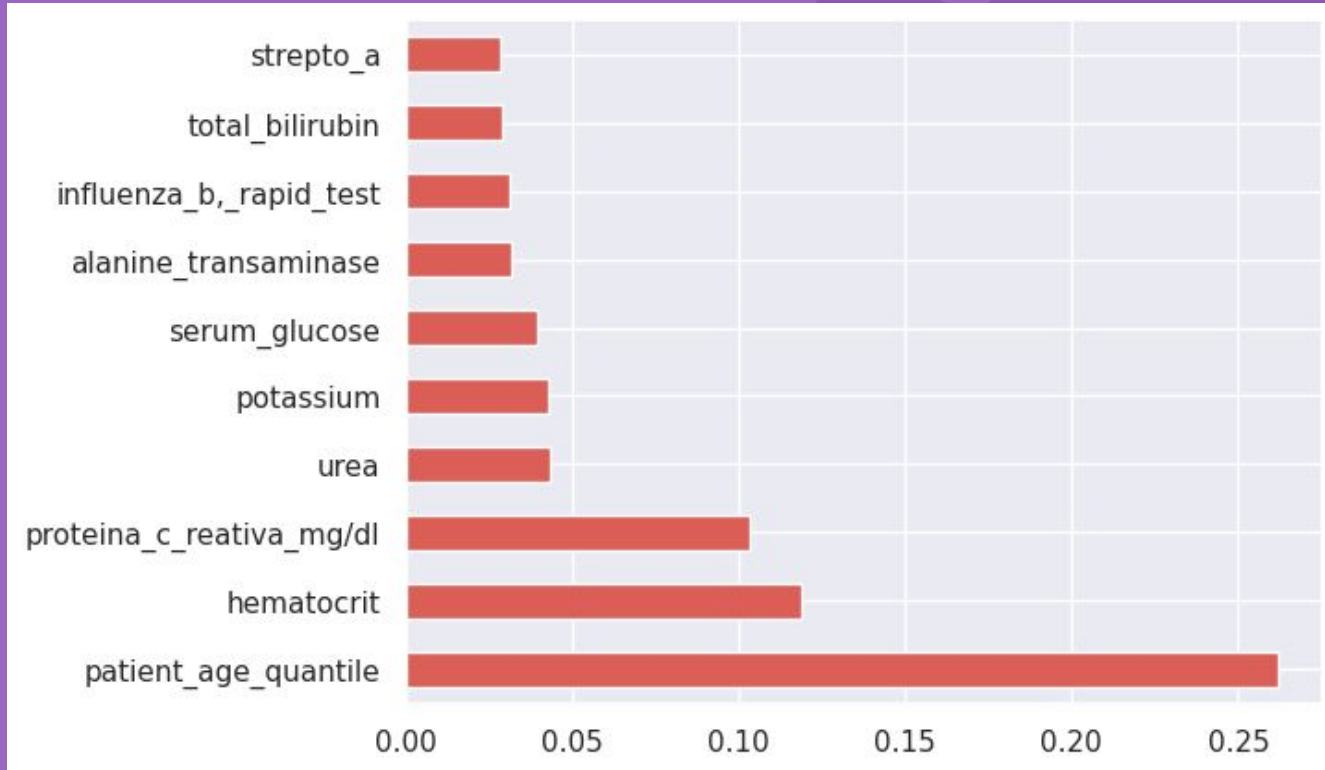




# Feature Correlation



# Feature Importance using Extra Trees Classifier





## Baseline models

	Model	Accuracy_score
0	Logistic Regression	90.552872
1	Random Forest	90.552872
2	XGBOOST	90.391841
3	NN	89.694042

# Hyperparameter-tuning & Final models

```
Logistic Regression Accuracy: 0.9207283829462436
Logistic Regression Best Params: {'LR__C': 0.01, 'LR__max_iter': 100, 'LR__penalty': 'l1', 'LR__solver': 'liblinear'}

Random Forest Accuracy: 0.9232362883000763
Random Forest Best Params: {'RF__criterion': 'gini', 'RF__max_depth': 8, 'RF__max_features': 'auto', 'RF__n_estimators': 200}

XGBoost Accuracy: 0.9273797841020608
XGBoost Best Params: {'XGB__colsample_bytree': 0.6, 'XGB__gamma': 1, 'XGB__max_depth': 5, 'XGB__n_estimators': 500, 'XGB__subsample': 1.0}
```

```
accuracy_df.sort_values(by='Accuracy Score', ascending=False)
```

	Model	Accuracy Score
0	Logistic Regression	90.552872
1	Random Forest	90.552872
4	Best Logistic Regression	90.552872
5	Best Random Forest	90.552872
2	XGBOOST	90.391841
6	Best XGBoost	90.338164
7	Best NN	90.338164
3	NN	89.694042

# Voting Classifier

```
[189] from sklearn.ensemble import VotingClassifier
      eclf = VotingClassifier(estimators=[('Logistic Regression', lg), ('Random Forest', rforest), ('XGBoost', xgb), ('Tensor Flow', nn)], voting='hard')
      #test our model on the test data
      eclf.fit(X_train, y_train)
      eclf.fit(X_test, y_test)
      eclf.score(X_test, y_test)
```

```
0.9146537842190016
```

```
▶ eclf.predict_proba(X.iloc[284].values.reshape(1,-1))
```

```
↳ array([[0.94768973, 0.05231027]])
```



## Findings from Data Mining :

- Patient age plays an important role in covid-19 result, as we the plot of Patient age against covid-19 test results it shows that :
  - Patients with age between 25-55 years and greater than 90 years, has highest risk of being covid positive
  - Childrens with age between 0-20 years have 0 chances of being covid positive
- Features which are responsible for drawing covid positive/negative results :
  - 'patient\_age\_quantile', 'hematocrit', 'serum\_glucose',
  - 'respiratory\_syncytial\_virus', 'mycoplasma\_pneumoniae', 'neutrophils',
  - 'urea', 'proteina\_c\_reativa\_mg/dl', 'potassium',
  - 'influenza\_b,\_rapid\_test'





## Conclusion :

- In conclusion, the COVID-19 dataset provided by Hospital Israelita Albert Einstein in São Paulo, Brazil, has provided a valuable resource for developing machine learning models to aid in the diagnosis of COVID-19 cases.
- The dataset includes anonymized data from patients who had samples collected to perform the SARS-CoV-2 RT-PCR test and additional laboratory tests during a visit to the hospital.
- Through exploratory data analysis, we have observed that the dataset contains a variety of features related to clinical measurements and laboratory results. We have also observed that the dataset is imbalanced, with a relatively small number of positive COVID-19 cases compared to negative cases.
- We have performed machine learning experiments on the dataset, including cross-validation and ensemble models. Our experiments have demonstrated that it is possible to predict COVID-19 test results based on laboratory and clinical data, with accuracy ranging from 90-95%.






## Future Work:

### **Healthcare providers:**

Healthcare providers can use these machine learning models to predict the likelihood of a patient having COVID-19 based on their symptoms and lab results. Additionally, providers can use the model to monitor the diseases spread and identify high-risk areas to target interventions and prevent outbreaks.

### **Travel industry:**

Travel companies can use this dataset and machine learning models to evaluate the risk of COVID-19 transmission in different locations and adjust their travel policies accordingly.





The background is a solid purple color. It features several stylized virus particles. There are four large, dark purple virus particles with prominent spikes: one on the left side, one on the right side, and two smaller ones at the bottom corners. Additionally, there are several smaller, light purple virus particles scattered across the top and bottom edges. The word "THANKS!" is centered in the middle of the image in a bold, yellow, sans-serif font.

**THANKS!**