**BLUEPRINT**

**Problem Statement:**

-To devise a stock trading strategy which utilizes concepts of Reinforcement Learning Techniques such as discretized Q – Learning, Hill Climbing, Deep Q Learning etc. Our aim is to devise a strategy that can make trades on a single exchange traded fund in the S&P 500 Stock Index, that outperforms simple common trading strategies such as the MACD indicator technique or the buy and hold method.

**Scraping data**

* Scraping the data using selenium and beautifullsoup

**Data Cleaning**

* Removing the null values
* or imputing the null values

**Detecting outliers**

* An outlier is a data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the dataset.

**Should outliers be removed?**

Depending on the context, outliers either deserve special attention or should be completely ignored. Take the example of revenue forecasting: if unusual spikes of revenue are observed, it's probably a good idea to pay extra attention to them and figure out what caused the spike. In the same way, an unusual transaction on a credit card is usually a sign of fraudulent activity, which is what the credit card issuer wants to prevent. So in instances like these, it is useful to look for and investigate further outlier values.

If outliers are however, due to mechanical error, measurement error or anything else that can't be generalised, it’s a good idea to filter out these outliers before feeding the data to the modeling algorithm.

**What defines an outlier?**

Data point that falls outside of 1.5 times of an interquartile range above the 3rd quartile and below the 1st quartile Data point that falls outside of 3 standard deviations. we can use a z score and if the z score falls outside of 2 standard deviation

**Why do outliers exist in a dataset?**

* Variability in the data
* An experimental measurement error

**Impact of an Outlier**

* It causes various problems during our statistical analysis
* It may cause a significant impact on the mean and the standard deviation

**How to find an outlier?**

* Outlier analysis and anomaly detection are a huge field of research devoted to optimise methods and create new algorithms to reliably identify outliers. There are a huge number of ways optimised to detect outliers in different situations. These are mostly targeted to identify outliers when those are the observations that we indeed want to focus on, for example for fraudulent credit card activity.

**Using the following methods we can detect an outlier**

* IQR interquantile range
* z score
* Scatter plots
* Box plot

**EDA**

* knowing about the data
* knowing the didtribution
* if required doing transformation
* Finding any correlation between feature variables and label
* finding any patteren between feature variables and label

**Selecting features and label**

* Selecting the variables which has an impact on output variable

**Checking for data imbalance**

* if we found any dataimbalance then we will proceed with sampling technique like upsampling ,SMOTE and downsampling

**Train test split**

* splitting the dataset into train and test

**Scaling the data**

-types

* Standardscalar
* Min Max scalar
* robust Scalar

**Model Building**

* training the model on train dataset and testing on test dataset

**Model validation**

* validating the different models
* selecting the best model