Stock Price Prediction

1st Bardh Ahmeti  
*Dizajn Softueri*  
*Universiteti ‘Ukshin Hoti’*Prizren , Kosove  
200306019

2nd Mirlinda Abdulla  
*Dizajn Softueri*  
*Universiteti ‘Ukshin Hoti’*Prizren , Kosove  
200306047

*Abstrakti*—Stock Price Prediction e përkthyer në gjuhën shqipe do të thotë ‘parashikim i çmimit të aksioneve’ , ku parashikimi i saktë i tyre është sfidues për arsye se ka faktorë të shumtë : të mëdhenj , të vegjël , politik , ekonomik , ngjarje të papritura , performanca financiare etj , ku ndikimi i tyre është i madh. Analist të ndryshëm ,studiues , shkenctarë po vazhdojnë të eksplorojnë teknikat e ndryshme për të zbuluar tendencat e tregut të aksioneve.

Objektivat tona kryesore për këtë temë janë që përmes metodologjive të ndryshme , e duke u bazuar në gjuhën ‘python’ të analizojmë të dhëna të marra nga dataset , gjithëashtu ti paraqesim në mënyre vizuale apo përmes pamjeve grafike se si do të dukeshin këto të dhëna të ndara sipas kritereve të ndryshme.

Ka metoda të ndryshme të cilat na ndihmojnë në paraqitjen e të dhënave tona , në fillim do të paraqesim disa të dhëna hyrëse përshkak se duhet të kuptohet më qartësishtë tema e zgjedhur e pastaj përmes ketyre metodave ne do ta shtjellojmë temën tonë: Regresioni Logjik , Makina Vektoriale Mbështetëse (SVM) , XGBClassifier të cilat në pjesën hyrese dhe në pjesën metodologjitë do të sqarohen se si funksionojnë.

# Hyrje

Duke e marrë me mend secili prej nesh kemi treg shumë të paqëndrueshem dhe shpesh të paparashikueshëm.Për disa vjet e dekada kanë punuar me të dhënat e serive kohore për të parashikuar vlerat e ardhshme nga të cilat aplikacionet me të ardhura më të mira dhe më sfidues është parashikimi i vlerave të aksioneve për një kompani të caktuar.

Megjithatë siç edhe është pritur ky ndryshim i tregut varet nga shumë parametra nga të cilët vetem një grup mund të matet si : të dhënat historike të aksioneve , vëllimi i tregtisë , çmimet aktuale.

Natyrisht disa nga faktorët themelor të tillë si : vlera e një kompanie , aktivitet , performanca tremujore , investimet e fundit , dhe strategjitë , ndikojnë në besimin e tregëtareve dhe kështu në çmimin e aksioneve të saj.Vetëm disa prej tyre mund të përfshihen në mënyrë efektive në një model matematikor.

Kjo e bën parashikimin e çmimit të aksioneve duke përdorur ‘machine learning’ sfidues dhe jo të besueshem deri në një masë , sepse pothuajse është e pamundur të parashikohet një lajm që do të shkatrrojë ose nxisë tregun e aksioneve në javët e ardhshme – një pandemi , një luftë etj.

Pra , në vend që të fokusohen në vlerat aktuale dhe përputhjen e tyre me saktësi të lartë , analistët fokusohen thjeshtë në bërjen e parashikimeve me afat të shkurtër për të bërë një vlerësim probabilist të asaj se si tregu ‘mund’ të duket së shpejti.

Me të dhëna të mjaftueshme historike dhe veçori të dobishme e të ndryshme modelet e matematikore dhe ‘machine learning’ parashikojne luahtje me afat të shkurtë në treg për një ditë tregu mesatare dhe pa ngjarje.

Stock Price Prediction është procesi i parashikimit të vlerës së ardhshme të një aksioni të tregtuar në bursë për të arritur fitime.

E gjithë ideja e parashikimit të çmimeve të aksioneve është të fitoni fitime të konsiderueshme.

Përfshirja e faktorëve të shumtë në parashikim të çmimeve të aksioneve ndikon edhe në vështirësimin dhe bërjen më sfiduese parashikimin e tyre me saktësi të lartë , dhe kjo është ajo ku mësimi i makinës luan një rol jetësor. [1]

## Rëndësia e parashikimit të çmimit të aksioneve është se :

* Ndihmesë për kompanitë në rritje të kapitalit,
* Ndihmon në gjenerimin e pasurisë personale,
* Shërbejnë si tregues i gjendjes së ekonomisë,
* Është burim i përdorur gjerësisht për njerëzit që të investojnë para në kompani me potencial të lartë.

## Ne do të mundohemi të sqarojmë disa pjesë më kryesore që janë :

* Analiza e të dhënave
* Inxhinieri të veçorive
* Ndarja dhe normalizimi i të dhënave
* Zhvillimi dhe vlerësimi i modeleve

Shumica do të shtjellohen në pjesën III - ‘Metodologjia’.

# Punë të Ngjashme

Secila analizë që bëhet duhet të jetë e bazuar në ndonjë punë të ngjashme , për arsye se duhet të marrim përvojat më të mira, ti shqyrtojmë ato , të gjejmë mundësinë e përmirësimit të gabimeve të mundshme e në fund të paraqesim rezultatet e mirëfillta të punës tonë.

Për temën të cilën e kemi zgjedhur ka mjaftë material në internet , diskutime në forume të ndryshme , seminare , projekte etj.[2]

Secili projekt i gjetur ka të mirat që mund ti mirren por ka edhe mundësinë që ne të përmirësojmë me mundësitë tona të dijes.[3]

Ka edhe projekte tjera si : ‘Stock Price Prediciton Based on Natural Language Processing’[4] , i cili e ka bërë parashikimin duke u bazuar në modelet BERT dhe NEZHA të Intelegjencës Artificiale pastaj edhe modelin LSTM.

Në secilin projekt serioz që kemi gjetur kemi kuptuar se modeli LSTM është më i sigurti për temën tonë për këtë arsye edhe kemi vendosur që të punojmë me këtë model.[5]

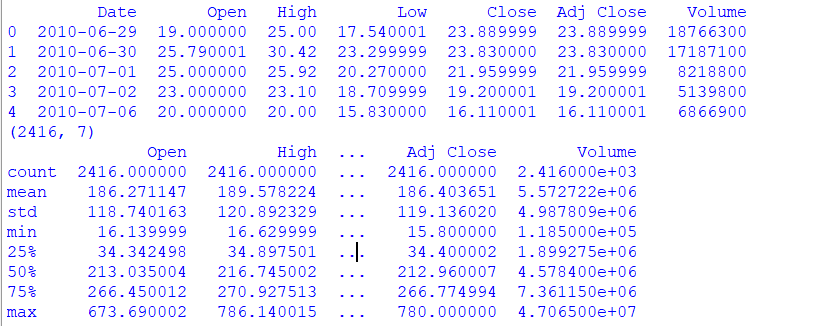
# Metodologjia

Një model është një skedar që është trajnuar për të njohur disa lloje modelesh , ne trajnojmë një model mbi një grup të dhënash , duke i ofruar atij një algoritëm që mund ta përdorë për të aryetuar dhe mësuar nga ato të dhëna.

Mësimi i makinës është jashtëzakonisht i dobishëm në shumë industri në automatizim të detyrave që më parë kanë pasur punë njerëzore , ne do të mësojmë se si të parashikojmë një sinjal që tregon nëse blerja e një aksioni të caktuar do të jetë e dobishme apo jo duke përdorur machine learning.

Do të fillojmë duke importuar librarari të ndryshme të cilat përdoren për qëllime të ndryshme të cilat do të sqarohen më vonë.

Të dhënat që do të përdorim këtu për të kryer analizën dhe për të ndërtuar një model parashikues janë të dhënat e çmimit të aksioneve Tesla , së pari do ta lexojmë datasetin e marrur , dhe duke lexuar pesë rreshtat e parë mund të shohim se mungojnë të dhënat për disa nga datat , për arsye se është gjatë fundjavave dhe festave Bursa mbetet e mbyllur , kështu që nuk ka treg për këto ditë.



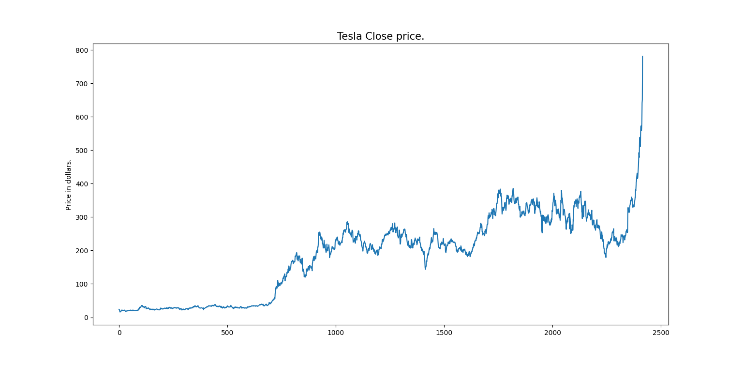
Fig, 1 . Initial information about data on 'Tesla’

## Analiza e të dhënave eksploruese

EDA është qasje analizuese e të dhënave , duke përdorur teknika vizuale , përdoret për të zbuluar tendencat dhe modelet , ose për të kontrolluar supozimet me ndihmën e përmbledhjeve : statistikore dhe grafike.

Ne do të analizojmë se si çmimet e aksioneve kanë lëvizur gjatë periudhës kohore dhe se si fundi i tremujorëve ndikon në çmimet e aksioneve.

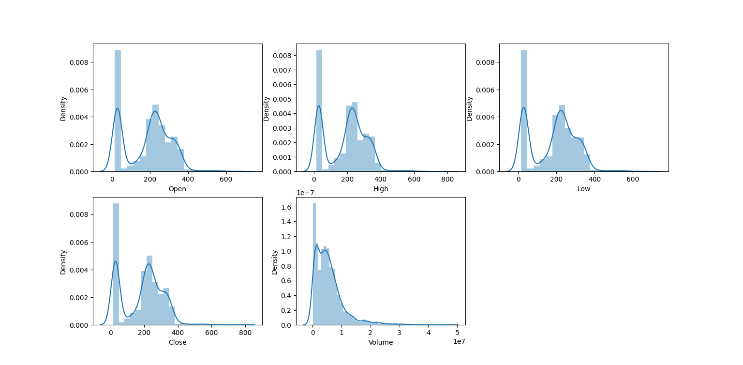
Çmimet e aksioneve të Teslas tregojnë se kanë një prirje rritëse siç përshkruhet nga grafiku i çmimit të mbylljes së aksioneve.



Fig, 2 . Tesla Close Price

Duke vëzhguar ne e shohim se të dhënat në kolonën ‘Close’ dhe ato ‘Adj Close’ janë të njëjta , e kontrollojmë me çdo rresht dhe shohim se janë plotësisht të njëjta , kështu të dhënat e tepërta nuk do të na ndihmojnë , pra ne do të heqim kolonën për analizë të mëtejshme.

Tani ne do ta vizatojmë grafikun e shpërndarjes për veçoritë në grupin e të dhënave , por së pari kontrollojmë për vlerat null nëse ka të pranishme në kornizën e të dhënave.



Fig, 3 . Distribution Plot of the Continuous Variable

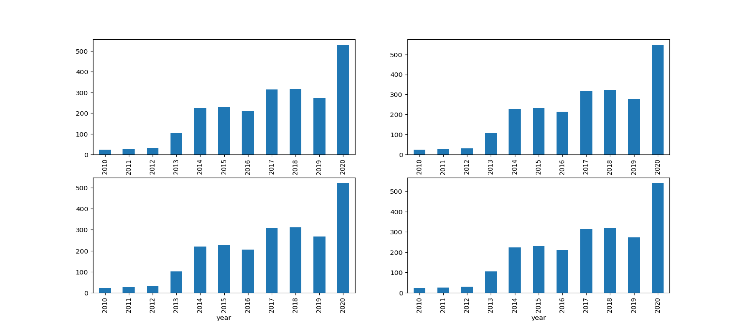
Ne mund të shohim në këtë grafik se janë dy maja që do të thotë se të dhënat kanë ndryshuar ndjeshëm në rajone.

## Inxhinieri të veçorive

Kjo inxhinieri na ndihmon për të nxjerrur disa veçori të vlefshme nga ato ekzistuese , këto shtesë na ndihmojnë në rritjen e ndjeshme të performancës së modelit dhe sigurisht ndihmojnë për të fituar njohuri më të thella mbi të dhënat.

Ne kemi shtuar tri kolona të tjera , përkatësisht ‘dita’ , ‘muaji’ . ‘viti’ të gjitha këto të nxjerra nga kolona ‘Data’ e cila fillimisht ishte dhënë si e vetme.

Kemi shtuar edhe kolonën ¼ (is\_quarter\_end) sepse çdo kompani i përgadit rezultatet e saj tre-mujore dhe i publikon ato , publikimi i tyre ndikojnë shumë në çmimin e aksioneve, prandaj kemi shtuar këtë veçori , sepse mund të jetë e dobishme për modelin e të mësuarit.



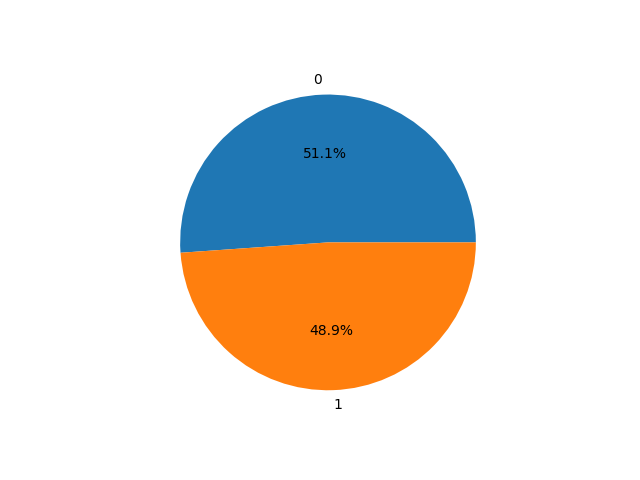
Fig, 4 . Doubling prices 2013-2014

Ne nga ky grafik i mësipërm me shtylla mund të konkludojmë se çmimet e aksioneve janë dyfishuar nga viti 2013 deri në vitin 2014 gjithëashtu edhe nga viti 2019 në vitin 2020.

Duke e përdorur një funksion groupby() , ne mund të nxjerrim disa nga vëzhgimet e rëndësishme të të dhënave të grupuara :

* Çmimet janë më të larta në muajt që janë në fund të tremujorit në krahasim me muajt e fundit jo tre-mujorit.
* Vëllimi i tregut është më i ulët në muajt që janë në fund të tre-mujorit.

Për t’i ndihmuar trajnimit të modelit tonë , ne kemi shtuar edhe target feature apo ‘veçorinë e synuar’ që është një sinjal nëse do të blejmë apo jo , ne do të trajnojmë modelin tonë për të parashikuar vetëm këtë , por përpara saj ne do të kontrollojmë nëse objektivi është i balancuar ose jo përmes pie chart :



Fig, 5 . Balancing Target

Por duhet të kemi kujdes që kur të shtojmë veçori në grupin e të dhënave tona , duhet të sigurohemi që nuk ka veçori shumë të ndërlidhura pasi ato nuk ndihmojnë mësimin e algoritmit.

Ne kemi konstatuar se veçoritë e shtuara nuk janë shumë të lidhura me njëra-tjetrën ose me veçoritë e mëparshme , që do të thotë se ne mund të ndërtojmë modelin tonë.

Nga parqitja e hartës së nxehtësisë apo heatmap , ne mund të themi se ekziston korrelacion i lartë midis kolonave Open, High , Low dhe Close , i cili është mjaftë i dukshëm .

Gjithëashtu edhe veçoritë e shtuara nuk janë shumë të lidhura njëra me tjetrën ose me veçoritë e ofruara më parë që do të thotë se ne jemi mirë të shkojmë e ta ndërtojmë modelin tonë.

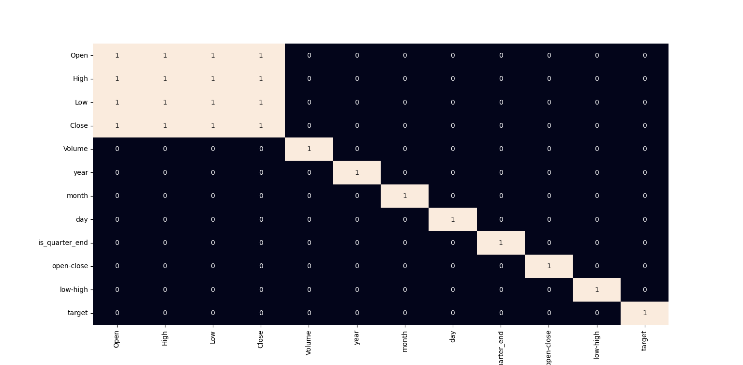


Fig 6. Heatmap of OHLC and other features

## Ndarja dhe normalizimi i të dhënave

Pasi që kemi zgjedhur veçoritë për të trajnuar modelin , ne duhet të normalizojmë të dhënat sepse të dhënat e normalizuara na qojnë në trajnim të qëndrueshem dhe të shpejtë të modelit.

Pas kësaj , të gjitha të dhënat janë ndarë në dy pjesë me një raport 90/10 , në mënyrë që të mund të vlerësojmë performancën e modelit tonë në të dhëna të padukshme.

features = df[['open-close', 'low-high', 'is\_quarter\_end']]

target = df['target']

scaler = StandardScaler()

features = scaler.fit\_transform(features)

X\_train, X\_valid, Y\_train, Y\_valid = train\_test\_split(

features, target , test\_size=0.1, random\_state=2022)

print(X\_train.shape, X\_valid.shape)

## Zhvillimi dhe vlerësimi i modeleve

Tani është koha për të trajnuar disa modele më të avansuara të machine learning (Regresioni Logjik , Makina Vektoriale Mbështetëse , XGBClassifier) dhe më pas bazuar në performancën e tyre në të dhënat e trajnimit dhe vërtetimit ne do të zgjedhim se cili model ML i shërben qëllimit më mirë.

Për matjen (metrikën) e vlerësimit ne do të përdorim një lakore ROC-AUC e cila përdoret përgjithësisht për të matur saktësinë e parashikimeve.

Regresioni Logjik është një nga algoritmat më të njohura të mësimit të makinerisë , i cili hyn në teknikën e mësimit të mbikqyrur , përdoret për të parashikuar variablën kategorike duke përdorur një grup të caktuar variablash të pavarura. [6]

Rezultati tek Regresioni Logjik është vlerë kategorike ose diskrete , mund të jetë ose po ose jo , ose 0 ose 1 , e vërtetë pse e gabuar etj. por në vend që të japë vlerën e saktë si 0 dhe 1, jep vlerat probabilistike që qëndrojnë midis 0 dhe 1. Përdoret për zgjidhjen e problemeve të klasifikimit. [6]

Machine Vector Support (SVM) është një algoritëm i mbikëqyrur i mësimit të makinerisë që përdoret si për klasifikimin ashtu edhe për regresionin. Edhe pse themi problemet e regresionit, janë më të përshtatshmet për klasifikim. Objektivi i algoritmit SVM është të gjejë një hiperplan në një hapësirë N-dimensionale që klasifikon qartë pikat e të dhënave. [6]

Dimensioni i hiperplanit ,varet nga numri i veçorive. Nëse numri i veçorive hyrëse është dy, atëherë hiperplani është vetëm një vijë. Nëse numri i veçorive hyrëse është tre, atëherë hiperplani bëhet një plan 2-D. Bëhet e vështirë të imagjinohet kur numri i veçorive i kalon tre. [6]

Klasifikuesi XGBoost është një algoritëm i mësimit të makinës që aplikohet për të dhëna të strukturuara dhe tabelare. XGBoost është një zbatim i pemëve të vendimeve të rritura me gradient të krijuar për shpejtësi dhe performancë. XGBoost është një algoritëm i rritjes së gradientit ekstrem. Dhe kjo do të thotë se është një algoritëm i madh i mësimit të makinerive me shumë pjesë. XGBoost punon me grupe të dhënash të mëdha dhe të ndërlikuara. XGBoost është një teknikë e modelimit të ansamblit. [6]

Në vazhdim do të paraqesim kodin për të tre modelet të cilat i kemi zhvilluar :

models = [LogisticRegression(), SVC(

kernel='poly', probability=True), XGBClassifier()]

for i in range(3):

models[i].fit(X\_train, Y\_train)

print(f'{models[i]} : ')

print('Training Accuracy : ', metrics.roc\_auc\_score(

Y\_train, models[i].predict\_proba(X\_train)[:,1]))

print('Validation Accuracy : ', metrics.roc\_auc\_score(

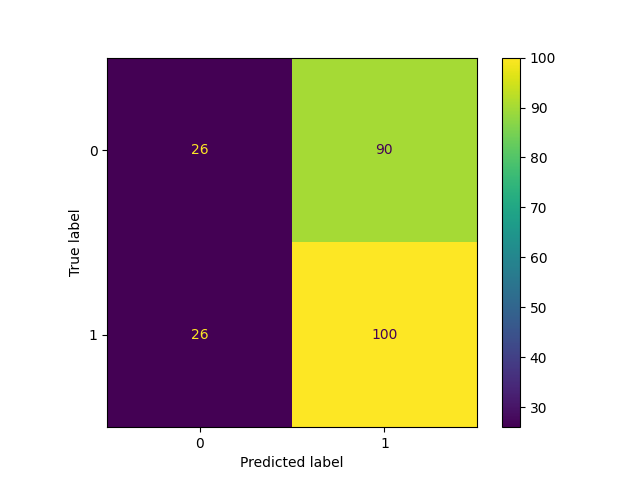
Y\_valid, models[i].predict\_proba(X\_valid)[:,1]))

print()

Modelet me të dhënat të cilat tashmë i kemi normalizuar dhe ndarë në vlera test , validation , pastaj me metodën fit janë mbushur secili , pastaj përmes predict\_proba ne kemi parashikuar vlerat të cilat do të paraqiten në pjesën ‘Rezultatet e Fituara’.

Për secilin model ne kemi vizatuar matricat e konfuzionit të cilat gjithë ashtu do ti paraqesim tek rezultatet e fituara.

Tani e vizatojmë një matricë të konfuzionit për të dhënat e vërtetimit :



Fig, 7 . Confusion matrix for validation data

Kjo është matrica për tre modelet – ne kemi paraqitur në skriptën tonë edhe matricat e secilit model.

Secilin model e kemi zhvilluar veç e veç por për mos ta ngarkuar punimin tonë me kod të shkruar ne kemi vendosur që atë ta bëjmë përmes një for-loop ,që të mund të paraqesim të tre modelet e zhvilluara bashkë në rezultate.

# Rezultatet e Fituara

Secili punim që bëhet , ka synim arritjen e rezultateve të caktuara , ne kemi punuar në këtë projekt për të zhvilluar këto tre modele : Regresioni Logjik , SVM , XGBClassifier .

Secili model na ka sjellur rezultate të ndryshme për shkak të ndryshimit që kanë ndërmjet vete secili nga ta.

Ne do të paraqesim rezultatet si në vijim :

-Saktësinë e testimit

-Saktësinë e parashikimit

-Matrica e Konfuzionit për modelin.

Mund të vërejmë se saktësia e arritur nga modeli më i avancuar ML nuk është më i mirë sesa thjesht të hamendësosh me një probabilitet prej 50%. Arsyet e mundshme për këtë mund të jenë mungesa e të dhënave ose përdorimi i një modeli shumë të thjeshtë për të kryer një detyrë kaq komplekse si parashikimi i bursës.

Regresioni Logjik – ka këto rezultate :

*Training Accuarcy : 0.5228802330060918*

*Validation Accuarcy : 0.4923371647509579*

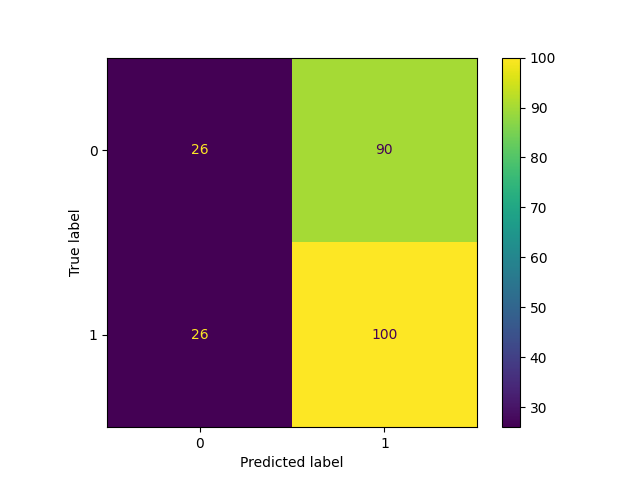


Fig 8. Confusion matrix for LR

SVC – ka këto rezultate :

*Training Accuarcy : 0.5294589297129334*

*Validation Accuarcy : 0.46257525998905313*

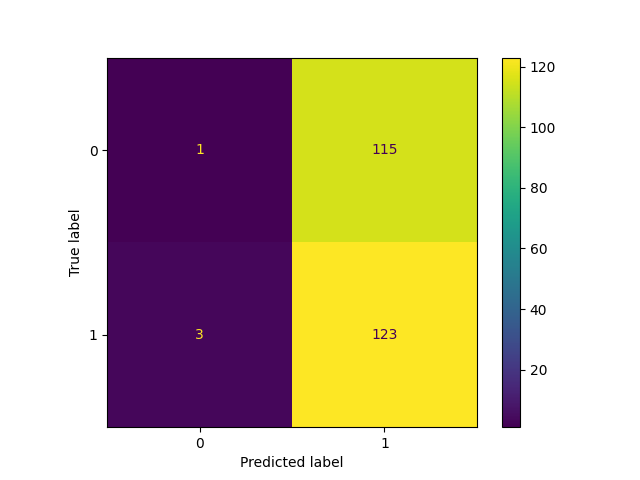


Fig 9. Confustion Matrix for SVC

XGB Classifier – ka këto rezultate :

*Training Accuarcy : 0.968013309795654*

*Validation Accuarcy : 0.41772714833059665*

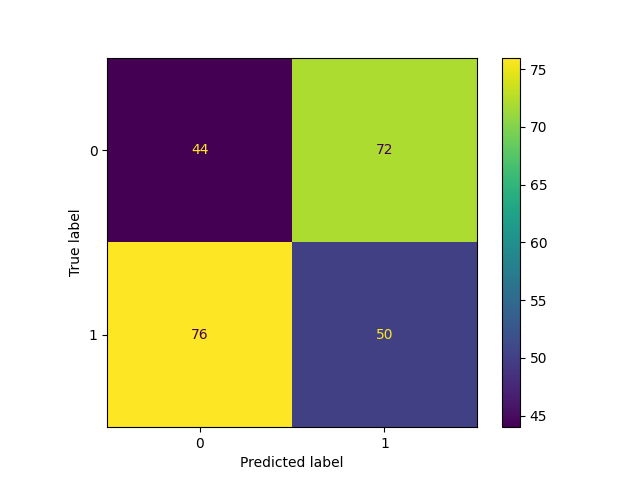


Fig 10. Confusion Matrix for XGBClassifier

# Konkluzione

Jemi munduar të jemi të qartë në analizimin dhe shpjegimin e temës.

Stock Price Predicition ka qenë temë e qasshme dhe e lehtë e organizueshme deri në momentin kur erdhi pandemia COVID-19 , apo edhe një luftë, e këtu arsyeja kryesore është se gjithçka ndryshoi , çdo parashikim ishte i pamundshëm.

Parashikimi i çmimeve mund të bëhet edhe me modelet më të avansuara të machine learning të cilat quhen edhe deep -learning.

TensorFlow dhe LSTM [7] kanë treguar rezultate më të mira por gjithmonë duke u bazuar në mos ndryshim të tregut nga faktorë të mëdhenj siç është pandemia apo një luftë.

Tre modelet tona të cilat i kemi implementuar nuk kanë treguar rezultate shumë të kënaqshme dhe mund të jenë jo shumë të bësueshme.

Është vështirë e hamendsueshme se me 50% propabilitet apo saktësi mund të bëhet parashikim i saktë i çmimeve.

Mund të ketë edhe arsye të tjera si : mungesa e të dhënave të mjaftueshme ose edhe përdorimi i modeleve të thjeshta për parashikim.

Ne konkludojmë se në kushte normale parashikimi i çmimeve të aksioneve do të ishte më i saktë por faktorë të shumtë ndikojnë në punën tonë.

##### Referencat

1. <https://www.projectpro.io/article/stock-price-prediction-using-machine-learning-project/571>
2. <https://www.simplilearn.com/tutorials/machine-learning-tutorial/stock-price-prediction-using-machine-learningJ>
3. <https://neptune.ai/blog/predicting-stock-prices-using-machine-learning>
4. <https://www.hindawi.com/journals/complexity/2022/9031900/>
5. <https://www.sciencedirect.com/science/article/pii/S1877050920307924>
6. [Stock Closing Price Prediction using Machine Learning Techniques - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S1877050920307924)
7. <https://www.datacamp.com/tutorial/lstm-python-stock-market>
8. <https://www.youtube.com/watch?v=KUFmCwCVXWs>
9. <https://www.youtube.com/watch?v=1O_BenficgE>
10. <https://www.youtube.com/watch?v=H6du_pfuznE>

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published**