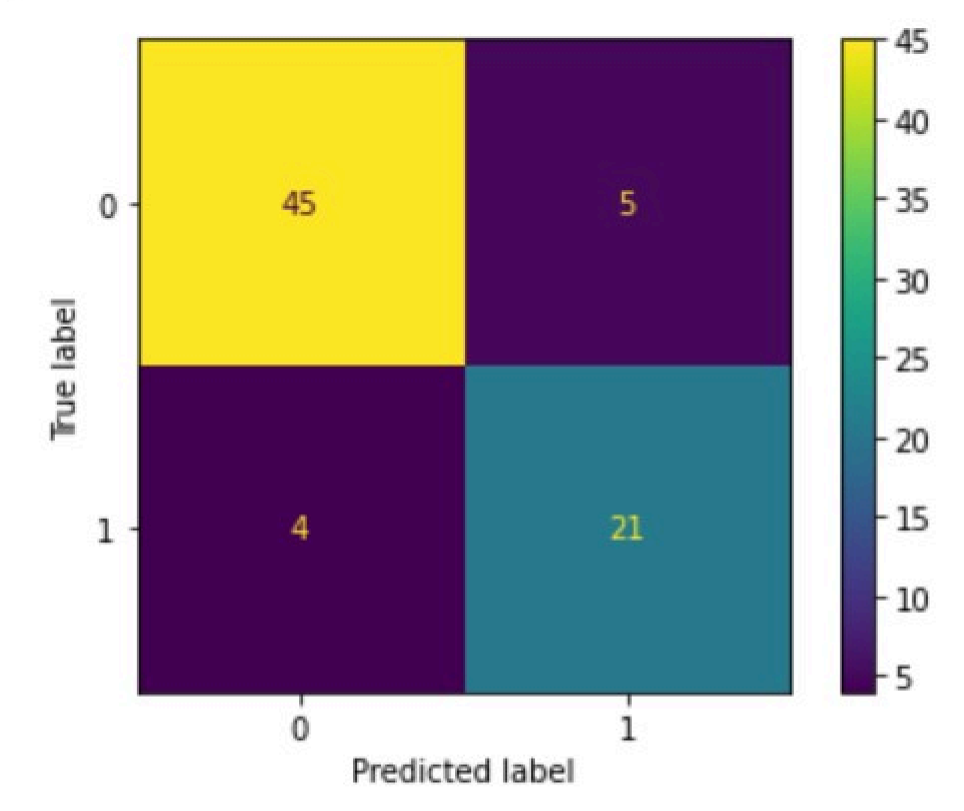
Report on NBA players making it to the hall of Fame

First of all, I picked a data set on the NBA players whether they make it to the hall of Fame. The data consist of 7 features which consists of Name, position, all star appearance, if the player is in hall of fame, height, weight, and the year they are born. I removed the column called "position" since it's a string and not very helpful in predicting training and testing. Then I make dummy variables for the position each players played in and add that to an extra column just for the purpose to make the data look good. Although the data consists of retired and current players, we want to perform the split train and test on the retired player because those players matter since they are the deciding factor of if they made the hall of fame or not. Therefore, we removed the current players from the dataset and perform it on the retired players. We randomized the dataset and calculate the index split. The trainX would be all the features except the "In_Hall_of_fame" column whereas the trainY would just the "In_Hall_of_fame" representing with binary numbers. Then I used the Multinomial Naive Bayes' train data to fit the model and predict it on the testX which produces this graph.

Confusion Matrix Naive Bayes'

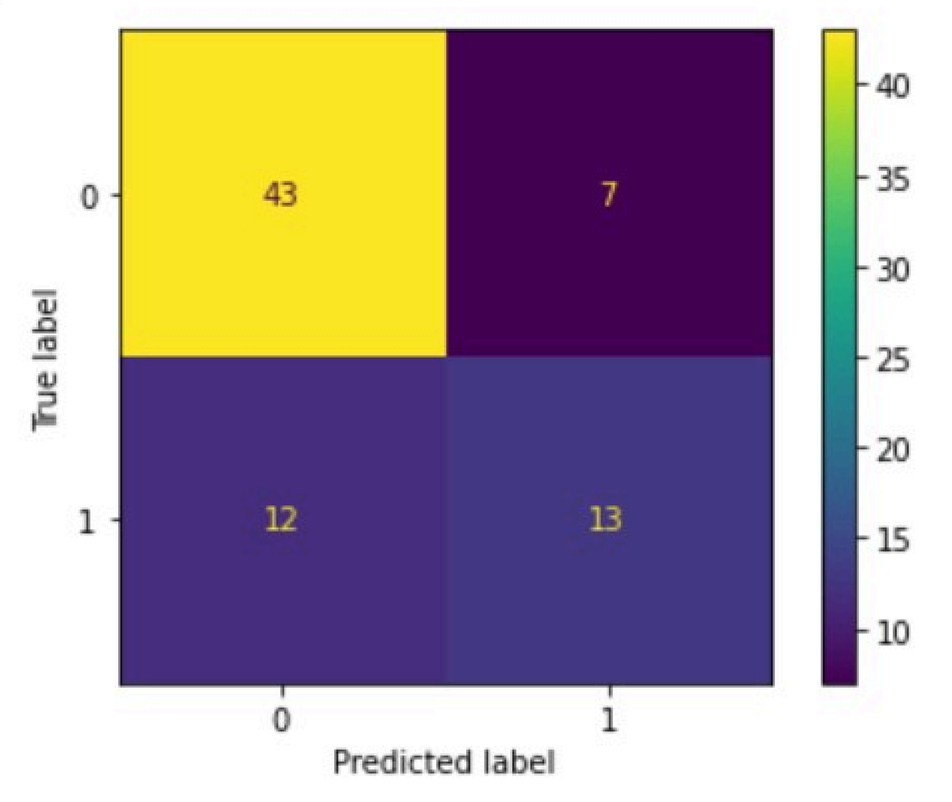
In [620]:
```
plot_confusion_matrix(clf, testX, testY)
plt.show()
```



On this confusion matrix plot on Naive Bayes' prediction, we can see less error and the accuracy score is 0.88

Confusion Matrix for KNN

In [625]:
```
plot_confusion_matrix(neigh, testX, testY)
plt.show()
```

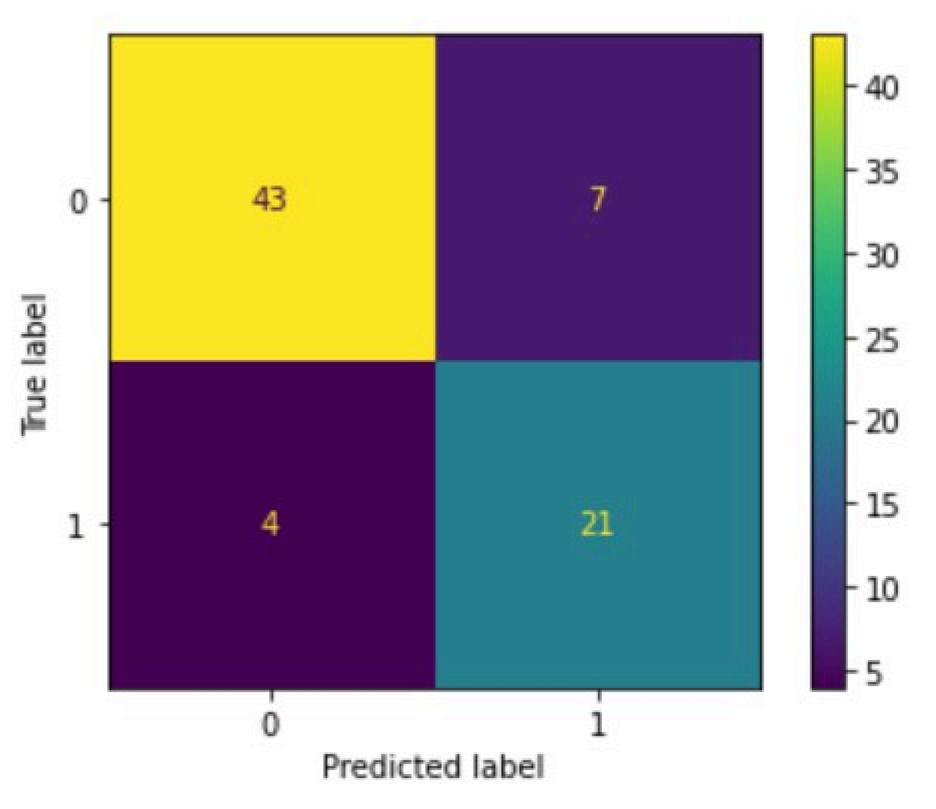


Confusion Matrix for KNN is also not very bad. However, it's showing unnecassary maount of true negative. The accuracy score, approximately, for it is 0.75

In addition, I added Random Forest classifier to see if the accuracy is more or less than the Naive Bayes' or KNN. It turns out that it's less than but close to Naive Baye's accuracy which is 0.85. Therefore, we still conclude that Naive Baye's is best for this dataset. Below is the confusion Matrix for Random Forest Classifier and it shows a reasonable result.

Confusion Matrix for Random Forest Classifier

In [482]:
```
plot_confusion_matrix(rf_clf, testX, testY)
plt.show()
```



In addition, I did Random Forest 5 fold Cross Validation. First, I set the random.seed(123) then for the number of tree for the parameter I have [100, 500, 1000, 2000]. Then I try to fit the train X and Y for the test. Then I predict the testX and got the following.

Random Forest 5 Fold Cross Validation ¶

In [483]:
```
Y = retired_player['In_Hall_of_fame']
train_data = retired_player.drop(['In_Hall_of_fame'], axis=1)
random.seed(123)
num_trees = [100, 500, 1000, 2000]
skf = StratifiedKFold(n_splits = 5, shuffle = True)
error = []
for param in num_trees:
    rf_mod = RandomForestClassifier(n_estimators=param)
    for train_index, test_index in skf.split(train_data, Y):
        trainX, testX = train_data.iloc[train_index],train_data.iloc[test_index]

        trainY, testY = Y.iloc[train_index],Y.iloc[test_index]
        rf_mod = rf_mod.fit(trainX, trainY)
        y_pred = rf_mod.predict(testX)
        error.append(accuracy_score(testY, y_pred))
    print("Accuracy:", sum(error)/len(error), "with", param, "tree")
```

```
Accuracy: 0.8552072072072072 with 100 tree
Accuracy: 0.8579279279279278 with 500 tree
Accuracy: 0.8543903903903904 with 1000 tree
Accuracy: 0.8552792792792794 with 2000 tree
```

```
Accuracy: 0.8133333333333334 with 100 tree
Accuracy: 0.8533333333333333 with 100 tree
Accuracy: 0.8577777777777778 with 100 tree
Accuracy: 0.8494144144144145 with 100 tree
Accuracy: 0.8633153153153155 with 100 tree
Accuracy: 0.8616516516516518 with 500 tree
Accuracy: 0.8661776061776063 with 500 tree
Accuracy: 0.8695720720720722 with 500 tree
Accuracy: 0.8570370370370372 with 500 tree
Accuracy: 0.861873873873874 with 500 tree
Accuracy: 0.8562489762489764 with 1000 tree
Accuracy: 0.854894894894895 with 1000 tree
Accuracy: 0.858877338877339 with 1000 tree
Accuracy: 0.8622007722007723 with 1000 tree
Accuracy: 0.8578738738738739 with 1000 tree
Accuracy: 0.8575900900900901 with 2000 tree
Accuracy: 0.8589083200847908 with 2000 tree
Accuracy: 0.8600800800800802 with 2000 tree
Accuracy: 0.8603319108582269 with 2000 tree
Accuracy: 0.8598828828828831 with 2000 tree
```

Our best model happened to be around 86% accuracy. This is the same accuracy as the random forest I generated on the very first one. We know that this is quite a good performance for such a model. Random Forest models are often used for testing the performance of complex models, so we shouldn't be surprise since our model is not too complex.

For another additional method I use tree boosting. I run a single boosting model with 5-fold CV using sci-kit learn's implementation.

Gradient Boosting Classifier

16]:
```
random.seed(123)
boost_mod = GradientBoostingClassifier(n_estimators = 100, learning_rate = 1,max_depth = 2)
skf = StratifiedKFold(n_splits = 5, shuffle = True)
error = []
for train_index, test_index in skf.split(train_data, Y):
    trainX, testX = train_data.iloc[train_index],train_data.iloc[test_index]
    trainY, testY = Y.iloc[train_index], Y.iloc[test_index]
    boost_mod = boost_mod.fit(trainX, trainY)
    y_pred = boost_mod.predict(testX)
    error.append(accuracy_score(testY,y_pred))
    print(sum(error)/len(error))
```

```
0.8
0.8666666666666667
0.8355555555555556
0.8361261261261261
0.804036036036036
```

Additional information on if the current NBA players will make it to the Hall of Fame

So we can see that for tree boosting we got the ones on the bottom which is around 0.83 accuracy. Which is still okay compare to the 5 fold random forest.

I made addition research on if the current NBA players will make the Hall of fame or not. I used the previous yhat from the Naive Bayes' to predict the current NBA probability of making Hall of fame since the Naive Bayes' accuracy score is 0.88 which is higher than KNN. Below is the finding of it.

Current players who will or will not make it to the Hall of Fame

In [453]:
```
print(curr_pls.to_string())
```

	index	Name	position	All_star_selections	In_Hall_of_fame	height	weight	born	pred_HOF
0	2	LeBron James	F	17	2	203	113	1984	1
1	30	Chris Paul	G	11	2	183	79	1985	1
2	31	Kevin Durant	F	11	2	206	108	1988	1
3	40	Carmelo Anthony	F	10	2	203	108	1984	1
4	46	Russell Westbrook	G	9	2	190	90	1988	1
5	47	James Harden	G	9	2	196	99	1989	1
6	59	Anthony Davis	F	8	2	206	97	1968	1
7	60	Dwight Howard	C	8	2	211	120	1985	1
8	80	Stephen Curry	G	7	2	190	86	1988	1
9	81	Kyrie Irving	G	7	2	190	87	1992	1
10	82	Paul George	F	7	2	206	99	1990	1
11	83	LaMarcus Aldridge	F	7	2	211	117	1985	1
12	104	Kyle Lowry	G	6	2	183	92	1986	1
13	105	Damian Lillard	G	6	2	190	88	1990	1
14	106	Blake Griffin	F	6	2	208	113	1989	1
15	133	John Wall	G	5	2	193	88	1990	0
16	134	Jimmy Butler	F	5	2	201	99	1989	1
17	135	Kawhi Leonard	F	5	2	201	104	1991	1
18	136	Klay Thompson	G	5	2	201	97	1990	1
19	137	Al Horford	C	5	2	208	111	1986	1
20	138	Kevin Love	F	5	2	208	113	1988	1
21	139	Giannis Antetokounmpo	F	5	2	211	100	1994	1
22	177	Rajon Rondo	G	4	2	185	84	1986	0

22	177	Rajon Rondo	G	4	2	185	84	1986	0
23	178	DeMar DeRozan	G	4	2	201	100	1989	0
24	179	Paul Millsap	F	4	2	203	111	1985	0
25	180	Kemba Walker	G	4	2	203	95	1964	0
26	181	DeMarcus Cousins	C	4	2	211	122	1990	1
27	182	Joel Embiid	C	4	2	213	113	1994	1
28	214	Derrick Rose	G	3	2	190	86	1988	0
29	215	Bradley Beal	G	3	2	196	93	1993	0
30	216	Draymond Green	F	3	2	201	104	1990	0
31	217	Nikola Jokic	C	3	2	208	113	1995	0
32	218	Ben Simmons	G	3	2	211	109	1996	0
33	219	Marc Gasol	C	3	2	216	115	1985	0
34	288	Donovan Mitchell	G	2	2	185	98	1996	0
35	289	Isaiah Thomas	G	2	2	185	81	1961	0
36	290	Victor Oladipo	G	2	2	193	95	1992	0
37	291	Devin Booker	G	2	2	198	93	1996	0
38	292	Luka Doncic	G	2	2	201	104	1999	0
39	293	Khris Middleton	F	2	2	203	106	1991	0
40	294	Jayson Tatum	F	2	2	203	95	1998	0
41	295	Andre Drummond	C	2	2	211	126	1993	0
42	296	Domantas Sabonis	F	2	2	211	108	1996	0
43	297	Karl-Anthony Towns	C	2	2	213	110	1995	0
44	298	Nikola Vucevic	C	2	2	213	117	1990	0
45	299	Rudy Gobert	C	2	2	216	111	1992	0
46	419	Mike Conley	G	1	2	185	79	1987	0
47	420	Trae Young	G	1	2	185	74	1998	0
48	421	D'Angelo Russell	G	1	2	188	86	1994	0
49	422	Jeff Teague	G	1	2	188	84	1988	0

37	291	Devin Booker	G	2	2	198	93	1996	0
38	292	Luka Doncic	G	2	2	201	104	1999	0
39	293	Khris Middleton	F	2	2	203	106	1991	0
40	294	Jayson Tatum	F	2	2	203	95	1998	0
41	295	Andre Drummond	C	2	2	211	126	1993	0
42	296	Domantas Sabonis	F	2	2	211	108	1996	0
43	297	Karl-Anthony Towns	C	2	2	213	110	1995	0
44	298	Nikola Vucevic	C	2	2	213	117	1990	0
45	299	Rudy Gobert	C	2	2	216	111	1992	0
46	419	Mike Conley	G	1	2	185	79	1987	0
47	420	Trae Young	G	1	2	185	74	1998	0
48	421	D'Angelo Russell	G	1	2	188	86	1994	0
49	422	Jeff Teague	G	1	2	188	84	1988	0
50	423	Goran Dragic	G	1	2	190	86	1986	0
51	424	Jaylen Brown	F	1	2	193	97	1992	0
52	425	Jrue Holiday	G	1	2	193	92	1990	0
53	426	Zach LaVine	G	1	2	196	85	1995	0
54	427	Andre Iguodala	G	1	2	198	97	1984	0
55	428	Zion Williamson	F	1	2	198	129	2000	0
56	429	Gordon Hayward	F	1	2	203	102	1990	0
57	430	Bam Adebayo	C	1	2	206	116	1997	0
58	431	Brandon Ingram	F	1	2	206	86	1997	0
59	432	Julius Randle	F	1	2	206	113	1994	0
60	433	Pascal Siakam	F	1	2	206	104	1994	0
61	434	DeAndre Jordan	C	1	2	211	120	1988	0
62	435	Brook Lopez	C	1	2	213	124	1988	0
63	436	Kristaps Porzingis	F	1	2	221	108	1995	0

:
```
NBcurr_pls['pred_HOF'].value_counts(dropna = False)
```

```
0    41
1    23
Name: pred_HOF, dtype: int64
```

All together there are 63 rows. Among them, there is only 23 hall of famer out of 63 according to the Naive Bayes theorem that is tested on the current players who are in the league.

In [503]:
```
KNNcurr_pls['pred_HOF'].value_counts(dropna = False)
```

Out[503]:
```
0    60
1     4
Name: pred_HOF, dtype: int64
```

For KNN on the current players list. We got only 4 hall of famers out of 64. Since KNN accuracy is around 75%, it's not one of the reliable method for this dataset.

```
RFcurr_pls['pred_HOF'].value_counts(dropna = False)
```

```
0    59
1     5
Name: pred_HOF, dtype: int64
```

For random forest on the current players list. We can see that there are 5 players among the 64 who will got to Hall of fame in the future.

Additional Graph