Էլեն Դավթյան

Pima Indians Diabetes Dataset

Documentation

Բովանդակություն

Ընդհանուր բնութագիր

Նախագիծը նպատակ ունի ցույց տալու, թե ինչպես են կիրառվում առաջին կիսամյակում մեր անցած գիտելիքները մեքենայական ուսուցման խնդիրները լուծելիս: Ինքս ընտրելով Pima Indians Diabetes տվյալների համակարգը սկսեցի աշխատել դրա վրա ըստ հատուկ քայլերի հաջորդականության: Dataset – ի մանրամասն ներկայացմանը կանդրադառնանք ավելի ուշ: Նպատակը՝ առկա տվյալների հիման վրա սովորելով՝ մոդելը պետք է կարողանա ճիշտ կանխագուշակել, թե արդյո՞ք մարդը(այս դեպքում հանդես են գալու միայն 21 և բարձր տարիքի կանայք) հիվանդ է շաքարային դիաբետով: Կուսումնասիրենք և կնկարագրենք dataset-ի առանձնահատկությունները և պիտակները: Կծանոթանանք այն մոդելներին, որոնց հետ աշխատել եմ, ինչպես նաև կներկայացվի աշխատանքի ընթացքը: Մոդելների բաժնում նաև կներկայացվի մոդելի գնահատումը և հիպերպարամետրի ընտրությունը:

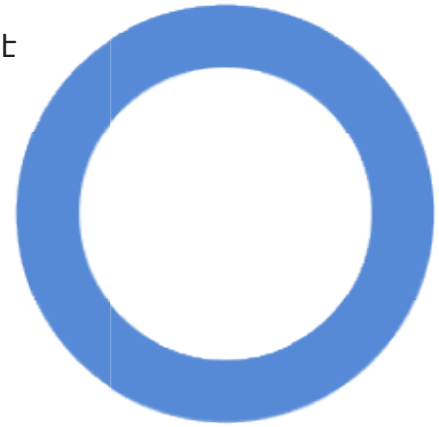
Սակայն մինչև հիմնական աշխատանքի ներկայացումը, նախնական գիտելիքներով զինվենք շաքարային դիաբետի մասին և հասկանանք, թե այն ինչ է իրենից ներկայացնում:

Հ.Գ. Հիմնականում տերմինները հանդես են գալու անգլերենով:)

Ի՞նչ է շաքարային դիբետը

Շաքարային դիաբետ, ներզատիչ հիվանդությունների խումբ է, որոնց առաջացման պատճառը կապված է **գլյուկոզի յուրացման խանգարման հետ**, որն էլ իր հերթին հարաբերականորեն կամ բացարձակ ձևով կախված է **ինսուլին հորմոնների** անբավարարությամբ (խանգարումը բջիջների), որի արդյունքում էլ առաջանում է **հիպերգլիկեմիան**՝ արյան մեջ գլյուկոզայի քանակության կայուն բարձրացումը:

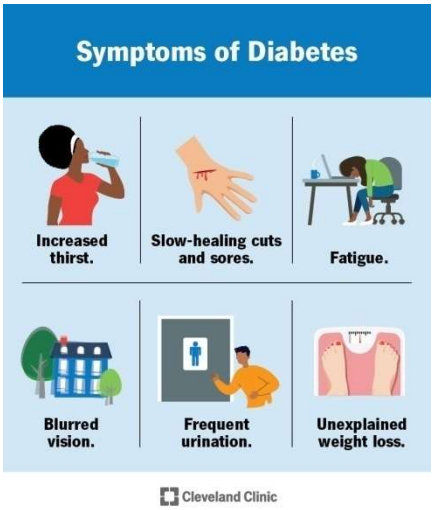
Տարբերում ենք շաքարային դիաբետի երեք հիմանական տեսակ՝



Դիաբետի խորհրդանիշ կապույտ օղը.

- 1-ին տիպի շաքարային դիաբետ, որի պատճառը հանդիսանում է ենթաստամոքսային գեղձի անբավարար քանակի ինսուլինի արտադրումը՝ բետա բջիջների կորստի պատճառով: Այս տեսակը նաև կոչվում է ինսուլին-կախյալ շաքարային դիաբետ:
- 2-րդ տիպի շաքարային դիաբետը սկսվում է ինսուլինի հանդեպ կայունությամբ, որտեղ բջիջները ունակ չեն լինում ինսուլինի հանդեպ համապատասխան ռեակցիա առաջացնել: Այս տեսակը նաև կոչվում է ինսուլինակախյալ շաքարային դիաբետ: Հիմնական պատճառներներն են մարմնի ավելորդ քաշը և ոչ բավարար ֆիզկական աշխատանք:
- Գեստացիոն դիաբետ, որտեղ հղի կանանց մոտ դիտվում է գերշաքարարյունություն, առանց անցյալում ունենալու նման ախտանշաններ:

Կանխարգելման և բուժման համար անհրաժեշտ է պահպանել առողջ սննդակարգ՝ հնարավորինս քիչ ածխաջրերի ընդունմամբ, ցուցաբերել բարձր ֆիզիկական ակտիվություն և խուսափել ծխելուց:



Որո՞նք են շաքարախտի ախտանիշները:

Շաքարախտի ախտանիշները ներառում են.

- Ծարավի ավելացում (պոլիդիպսիա) և բերանի չորացում:
- Հաճախակի միզարձակում.
- Հոգնածություն.
- Մշուշոտ տեսողություն.
- Անբացատրելի քաշի կորուստ.
- Ձեռքերում կամ ոտքերում թմրություն կամ քորոց:
- Դանդաղ բուժվող վերքեր կամ կտրվածքներ:

Pima Indians Diabetes Dataset

Այս տվյալների համակարգը հավաքագրվել է Շաքարախտի և մարսողական համակարգի և երիկամների հիվանդությունների ազգային ինստիտուտից(National Institute of Diabetes and Digestive and Kidney Diseases): Տվյալների հավաքածուի նպատակն է դիագնոստիկորեն կանխատեսել, թե արդյոք հիվանդը ունի շաքարախտ, թե ոչ՝ հիմնվելով տվյալների բազայում ներառված որոշակի ախտորոշիչ չափումների վրա: Մի քանի սահմանափակումներ են դրվել ավելի մեծ տվյալների բազայից այս դեպքերի ընտրության վրա: Մասնավորապես, այստեղ բոլոր հիվանդները Պիմա հնդկացիների ցեղխմբի ժառանգության առնվազն 21 տարեկան կանայք են:

Տվյալների համակարգը բաղկացած է 769 տողից (ներառյալ առանձնահատկությունների վերնագրերը) և 9 սյուներից: Առնաձնահատկություններն են Pregnancies, Glucose(mg/dL), BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age: ԵՎ վերջում ներկայացված է Outcome սյունակը, որը ներկայացնում է պիտակները 0-ների և 1-երի տեսքով:

Առանձնահատկությունների սեղմ նկարագիրը.

Pregnancies - հղիությունների քանակ.

Բազմակի հղիությունները կարող են ազդել գլյուկոզայի երկարատև հանդուրժողականության վրա: Հղիության ընթացքում գեստացիոն շաքարային դիաբետը կարող է մեծացնել 2-րդ տիպի շաքարախտի վտանգը ավելի ուշ կյանքում: Կան դեպքեր, երբ կինը հղի չի եղել, սակայն ունեցել է շաքարախտ, հետևաբար այս առանձնահատկությունը կարող է շփոթության մեջ գցել մեր դասակարգիչին:

Glucose(mg/dL) – գլյուկոզա

Գլյուկոզան ներկայացված է ըստ ամերկյան համակարգի՝ միլիգրամներով: Շաքարախտը ուղղակիորեն կապված է արյան շաքարի կարգավորման հետ, ուստի այս հատկանիշը մեծ հարաբերակցություն ունի արդյունքի հետ: Գլյուկոզայի մակարդակը շաքարախտի ամենաուժեղ կանխատեսողներից մեկն է: Հատկանիշը մեծ դեր ունի մեր հետագա մոդելի աշախատանքի համար:

BloodPressure – արյան ճնշում

Ոչ բոլոր դեպքերում է, երբ նորմայից բարձր արյան ճնշում ունեցողները հիվանդ են շաքարային դիաբետով, սակայն այն մադիկ, որոնց մոտ հաստատվել է հիվանդությունը հաճախ ունենում են արյան ճնշման տատանումներ: Այս առանձնահատկությունը այնքան էլ շատ կարևոր չէ, ինչպիսին են գլյուկոզան կամ MBI-ը, և հաշվի առնելով զրոների քանակը տվյալների համակարգում, աշխատանքի մեջ այն չի ներառվել:

SkinThickness - մաշկի հաստությունը

Մաշկի հաստությունը անուղղակի չափանիշ է և կարող է բոլոր դեպքերում խիստ կապ չունենալ շաքարախտի հետ: Սակայն այն կարող է հանդիսանալ շախարախտ ունենալու ախտանիշներից մեկը, երբ մարդը առանց պատճառի սկսում է նիհարել: Այս չափումը վերաբերում է մարմնի ճարպին, բայց դրա ներդրումը սովորաբար ավելի թույլ է:

Insulin – ինսուլին

Շիճուկում ինսուլինի մակարդակը պատկերացում է տալիս ինսուլինի դիմադրության կամ անբավարարության մասին: Ուսումնասիրելով համապատասխան սյունակի տվյալները, տեսնում ենք, որ զրոների՝ բացթողնված տվյալների, քանակը գերազանցում է, ուստի այս առանձնահատկությունը ևս նպատակահարմար չէ աշխատանքի համար:

BMI (Body Mass Index) - մարմնի զանգվածի ինդեքս

BMI-ի բարձր ցուցանիշը հաճախ փոխկապակցված է գիրության հետ, որը 2-րդ տիպի շաքարախտի զգալի ռիսկի գործոն է: Մարմնի ավելորդ քաշը ազդում է ինսուլինի դիմադրության վրա: Առանձնահատկությունը ներառվել է հետագա աշխատանքների մեջ:

DiabetesPedigreeFunction – շաքարային դիաբետի ժառանգման ֆունկցիա

Ցույց է տալիս, թե արդյոք մարդու տոհմում և ծագման մեջ եղել են շաքարախտով հիվանդներ: Որքան առանձնահատկության ցուցանիշները բարձր են, այդքան ավելի հավանական է, որ մարդը ևս հիվանդ է շաքարախտով (ավելի ուժեղ գենետիկ նախատրամադրվածություն շաքարախտի նկատմամբ): Առանձնահատկությունը ներառվել է հետագա աշխատանքների մեջ:

Age – տարիք

Տարիքը կարևոր գործոն է, քանի որ տարիքի հետ ինսուլինի նկատմամբ զգայունությունը ժամանակի ընթացքում նվազում է և դրա հետ զուգընթաց մեծանում է շաքարախտի ռիսկը: Առանձնահատկությունը ներառվել է հետագա աշխատանքների մեջ:

Տվյալների նախնական մշակում

Տվյալների նախնական մշակումը իրենից ներկայացնում է՝

- Տվյալների մաքրում և սխալների ուղղում,
- Չափավորում և արժեքների կարգավորում,
- Ճիշտ անհամապատասպանությունների ուղղում,
- Տվյալների փոխակերպումը իմաստալից հատկանիշների:

Տվյալների համակարգը ուսումնասիրելիս նկատեցի, որ որոշ տվյալներ բացակայում են և իրենց բացակայությունը ուղղակիորեն նշված է որպես 0(չի կարող մարդու արյան ճնշումը լինել 0, ...): Բացակայող տվյալները փոխարինվել են տվյալ սյունակի միջինացված արժեքով(կարճ՝ միջինով):

Feature extraction-ի ժամանակ տվյալներից դուրս են բերվել այն հատկանիշները, որոնց հետ շարունակվելու է հետագա աշխատանքը. դրանք են՝ Glucose, BMI, DiabetesPedigreeFunction, Age: Պիտակները հանդիսանում են մեր տվյալնների Outcome սյունակը, որը նշանակված է 0-ներով(բացակայում է շաքարային դիաբետ) կամ 1-երով (առկա է շաքարային դիաբետ):

Տվյալները նորմալիզացվել և ստանդարտիզացվել են՝ բերվելով -1 - 1 արժեքների միջակայքում ընկած թվերի: Նշեմ, որ տվյալների մեջ գլյուկոզան ներկայացված է ամերիկյան համակարգով՝ միլլիգրամներով, մինչդեռ եվրոպական ստանդարտներում ներկայացվում է մոլերով:

K - Nearest Neighbors

Ի՞նչ է իրենից ներկայացնում K-NN-ը.

K-nearest neighbors (KNN) ալգորիթմը ոչ պարամետրիկ, վերահսկվող ուսուցման դասակարգիչ է, որն օգտագործում է մոտիկությունը՝ անհատական տվյալների կետի խմբավորման վերաբերյալ դասակարգումներ կամ կանխատեսումներ կատարելու համար: Դա դասակարգման և ռեգրեսիայի դասակարգիչների հանրաճանաչ և ամենապարզ դասակարգիչներից մեկն է, որն այսօր օգտագործվում է մեքենայական ուսուցման մեջ:

k-NN ալգորիթմի **k** արժեքը սահմանում է, թե քանի հարևան կստուգվի կոնկրետ հարցման կետի դասակարգումը որոշելու համար: Օրինակ, եթե k=1, օրինակը վերագրվելու է նույն դասին, ինչ իր միակ մոտակա հարևանը:

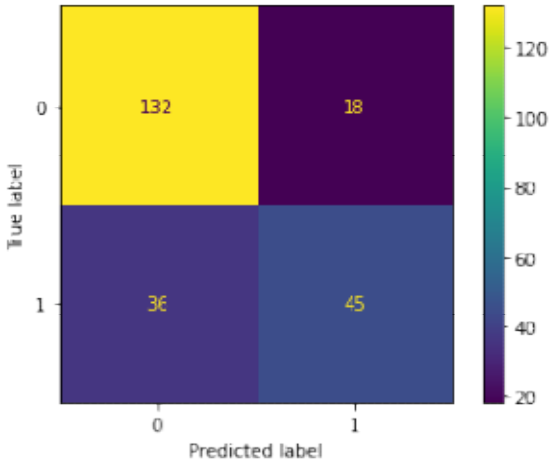
$$d(x,y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

Էվկլիդյան բանաձևը, որով հիմնականում հաշվվում է կետերի միջև հեռավորությունը

Աշխատանքը K-NN-ի հետ.

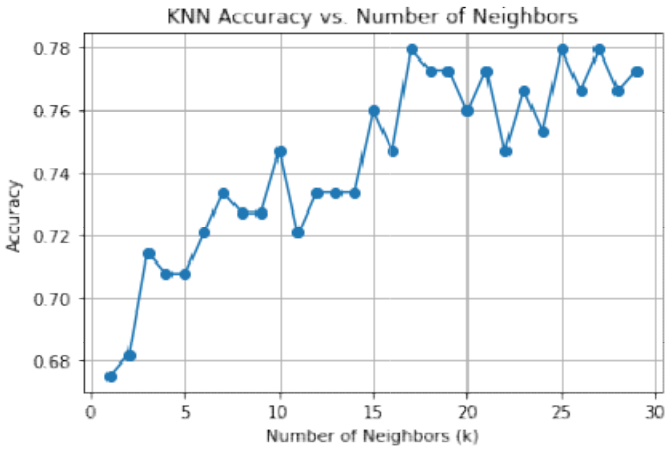
Դասընթացների ընթացքում ուսումնասիրելով մոդելը, այն կիրառվեց նաև վերջնական աշխատանքի մեջ: Training data-ի վրա վարժեցվելով training label-ների համար ցուցաբերեց 79.14% ճշգրտություն, իսկ testing data-վրա՝ 76.62%, երբ k=9:

Աշխատանքում, դեռևս առաջին անգամ առերեսվելով knn-ի հետ, մոդելի գնահատում իրականացվեց նաև confusion matrix-ի միջոցով:



Confusion matrix for KNN

Իրականացվեց cross validation՝ k հիպերպարամետրի համար լավագույն արժեքը ընտելու համար: Արդյունքում՝ k=17-ի դեպում մոդելը համապատասխանաբար ցուցաբերեց 81.43%(training) և 77.92%(testing):



Decision Tree

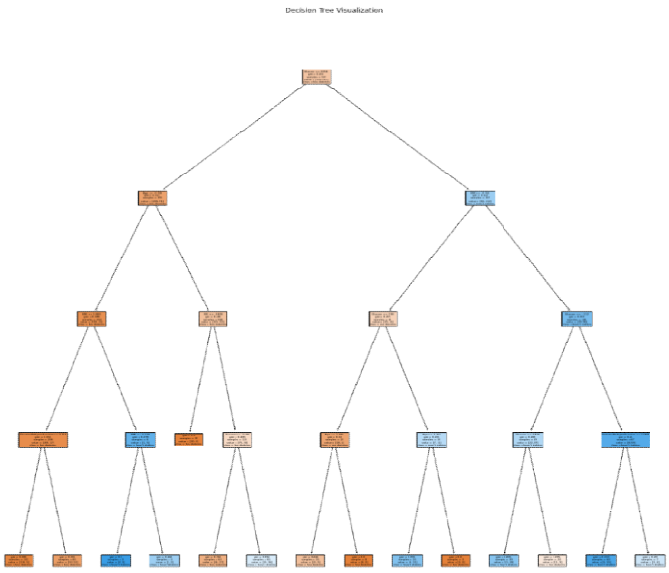
Ի՞նչ է իրենից ներկայացնում Decision Tree-ն

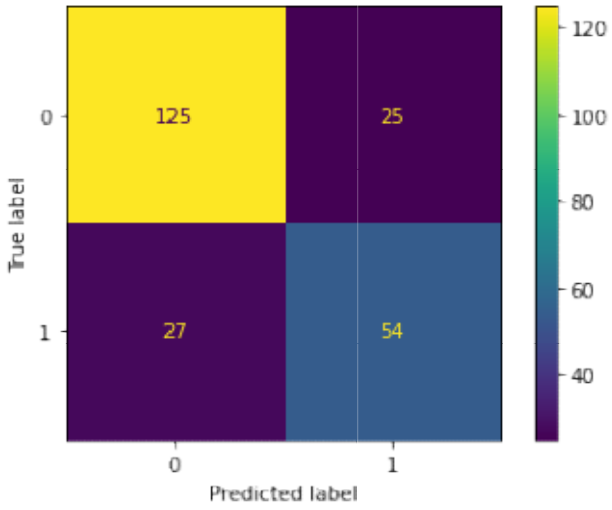
Decision Tree-ն ոչ պարամետրային վերահսկվող ուսուցման ալգորիթմ է, որն օգտագործվում է և՛ դասակարգման, և՛ ռեգրեսիայի առաջադրանքների համար: Այն ունի հիերարխիկ, ծառային կառուցվածք, որը բաղկացած է արմատային հանգույցից(root node), ճյուղերից(branches), ներքին հանգույցներից(internal nodes) և տերևային հանգույցներից(leaf nodes): Decision Tree-ի ուսուցումը կիրառում է բաժանիր և տիրիր ռազմավարությունը՝ իրականացնելով ագահ որոնում՝ ծառի մեջ օպտիմալ պառակտման կետերը բացահայտելու համար: Բաժանման այս գործընթացը այնուհետև կրկնվում է վերևից ներքև, ռեկուրսիվ եղանակով, մինչև բոլոր գրառումները կամ մեծամասնությունը դասակարգվեն հատուկ դասի պիտակների տակ:

Gini Impurity - տվյալների հավաքածուի մեջ պատահական տվյալների կետի սխալ դասակարգման հավանականությունն է, եթե այն պիտակավորվել է տվյալների բազայի դասային բաշխման հիման վրա:

Աշխատանքը Decision Tree-ի հետ.

Սկզբում օգտագործելով դասակարգիչը հետևյալ ձևով՝ *DecisionTreeClassifier(random_state=42, max_depth=4, min_samples_leaf=3),* այն ցուցաբերեց ճշգրտության արժեքներ՝ 81.19% training data-ի վրա և 77.49% testing data-ի վրա: Կիրառվեց նաև Confusion matrix-ը:

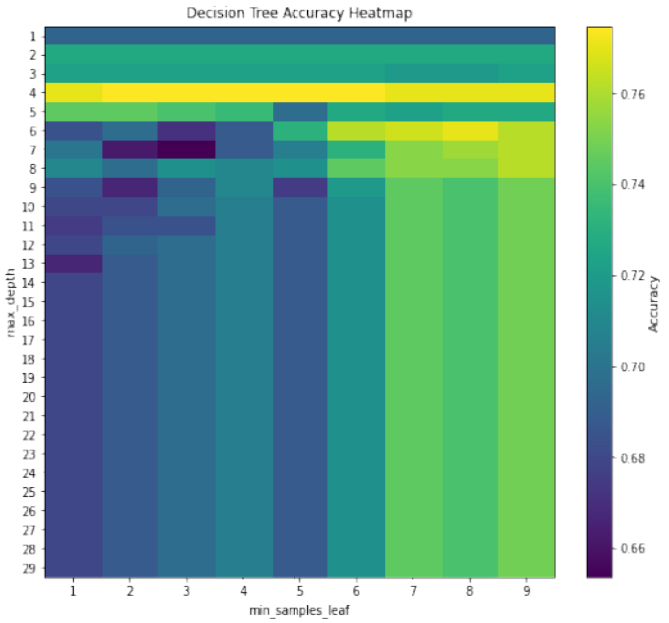




Cross validation-ի արդյունքները ներկայացվել են heatmap-ի միջոցով: Քանի որ որոշման ծառը ունի մի քանի հիպերպարամետր, օգտագործվել են խորության և min_sample_leaf-ի տարբեր արժեքներ: Ի վերջո խորության համար ընտրվել է 4 արժեքը, իսկ min_sample_leaf-ի համար՝ 6:

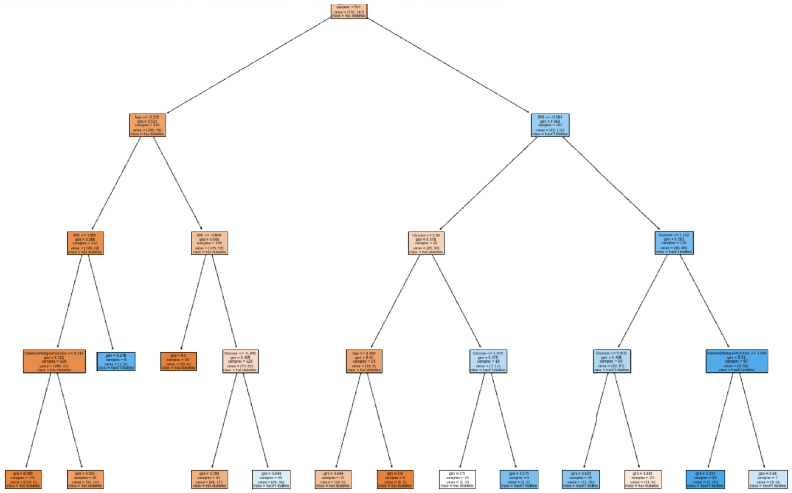
Ճշգրտությունները training data-ի համար ցուցաբերել է 80.63%, իսկ testing data-ի համար՝ 77.49%:

```
depths = list(range(1, 30))  # Tree depths to test
min_leaf_values = list(range(1, 10))  # Minimum samples per leaf to test
results = np.zeros((len(depths), len(min_leaf_values)))  # Store accuracies

# Loop through depths and min_samples_leaf
for i, depth in enumerate(depths):
    for j, min_leaf in enumerate(min_leaf_values):
        # Split the data with the fixed seed
        X_train, X_test, y_train, y_test = train_test_split(
            norm, labels, test_size=0.3, random_state=42, stratify=labels
        )

        tree_classifier = DecisionTreeClassifier(
            max_depth=depth, min_samples_leaf=min_leaf, random_state=42
        )
        tree_classifier.fit(X_train, y_train)

        # Test the model
        y_pred = tree_classifier.predict(X_test)
        acc = accuracy_score(y_test, y_pred)

        # Store accuracy
        results[i, j] = acc

# Visualize the results with a heatmap
plt.figure(figsize=(10, 8))
plt.imshow(results, interpolation='nearest', cmap='viridis', aspect='auto')
plt.colorbar(label='Accuracy')
plt.xticks(ticks=np.arange(len(min_leaf_values)), labels=min_leaf_values)
plt.yticks(ticks=np.arange(len(depths)), labels=depths)
plt.xlabel('min_samples_leaf')
plt.ylabel('max_depth')
plt.title('Decision Tree Accuracy Heatmap')
plt.show()
```




```
X_train, X_test, y_train, y_test = train_test_split(norm, labels, test_size=0.3, random_state=42, stratify=labels)

dt_classifier = DecisionTreeClassifier(random_state=42, max_depth=4, min_samples_leaf=6)
dt_classifier.fit(X_train, y_train)

y_pred_1 = dt_classifier.predict(X_train)
acc_1 = accuracy_score(y_train, y_pred_1) * 100
print(f"Decision Tree train accuracy: {acc_1:.2f}%")

y_pred = dt_classifier.predict(X_test)
acc = accuracy_score(y_true=y_test, y_pred=y_pred)*100
print(f"Decision Tree test accuracy: {acc:.2f}%")

plt.figure(figsize=(20, 15))
plot_tree(dt_classifier, feature_names=column_names,
          class_names=['has diabetes', "hasn't diabtes"],
          filled=True)
plt.title('Decision Tree Visualization')
plt.show()
```

```
Decision Tree train accuracy: 80.63%
Decision Tree test accuracy: 77.49%
```



Դասընթացներում ուսումնասիրված մոդելներից ընտրվեց Decision Tree-ն, իր ցուցաբերած արդյունքների համար, որոնք համեմատաբար ավելի բարձր էին քան K-NN-ի դեպքում:

Ծառի վերջնական տեսքը

Support Vector Machine (SVM)

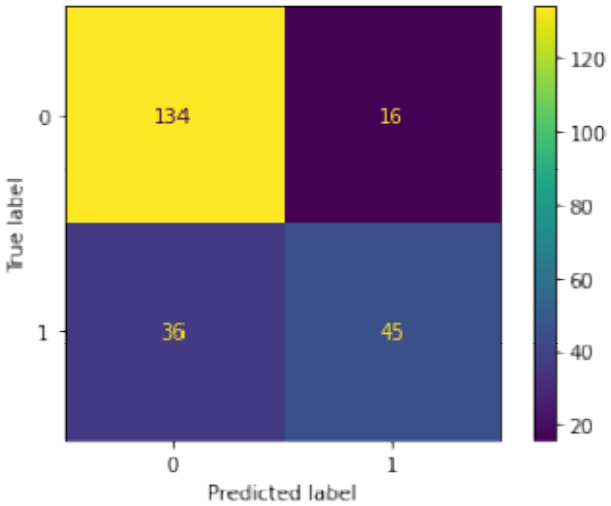
Ի՞նչ է իրենից ներկայացնում SVM-ը

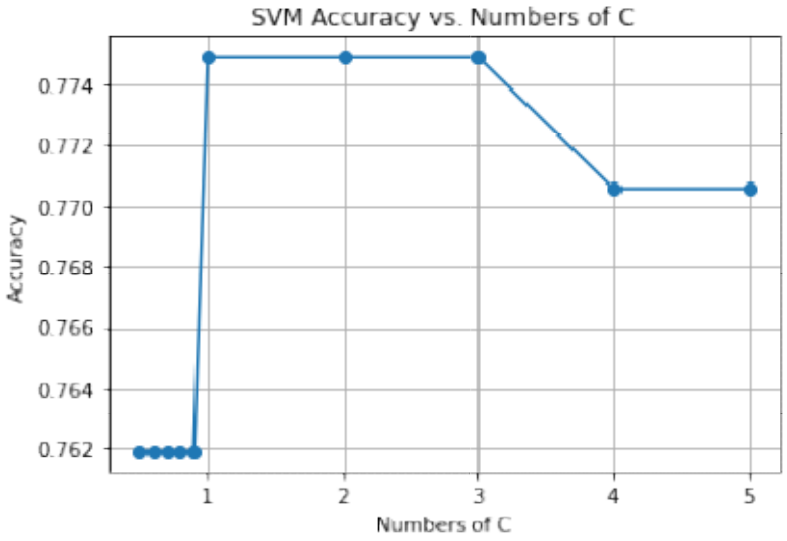
SVM-ը դասակարգման և ռեգրեսիայի կայուն տեխնիկա է, որը առավելագույնի է հասցնում մոդելի կանխատեսման ճշգրտությունը՝ առանց training data-ի overfitting-ի: SVM-ը հատկապես հարմար է շատ մեծ թվով (օրինակ՝ հազարավոր) կանխատեսող դաշտերով տվյալների վերլուծության համար: SVM-ն աշխատում է՝ տվյալները քարտեզագրելով մեծ չափերի հատկանիշի տարածության վրա, որպեսզի տվյալների կետերը կարողանան դասակարգվել, նույնիսկ երբ տվյալները այլ կերպ գծային բաժանելի չեն: Գտնվում է կատեգորիաների միջև բաժանարար, այնուհետև տվյալները փոխակերպվում են այնպես, որ բաժանարարը կարող է գծվել որպես հիպերպլան: Դրանից հետո նոր տվյալների բնութագրերը կարող են օգտագործվել՝ կանխատեսելու այն խումբը, որին պետք է պատկանի նոր գրառումը:

Աշխատանքը SVM-ի հետ.

Որպես նոր դասակարգիչ ընտրվեց SVM-ը, որի աշխատանքը ներկայացվել է կոդի վերջում, նույն քայլերի հաջորդականությամբ, ինչ նախորդ դասակարգիչների դեպքում: Հիպերպարամետրերից դիտարկվել է միայն մեկը՝ C-ն, որը վերահսկում է փոխզիջումը առավելագույնի հասցնելու և վերապատրաստման սխալի ժամկետը նվազագույնի հասցնելու միջև: Նախնական աշխատանքի ժամանակ ցուցաբերեց 81.56%(training) և 77.49%(testing): Կիրառվեց նաև confusion matrix:

Cross validation-ից հետո C-ի համար արժեք ընտրվեց 1-ը, երբ kernel=”rbf”: Այս դեպքում ցուցաբերեց նույն արժեքը ինչ սկզբնական աշխատանքի ժամանակ:





Օգտագործված գրականություն.

About Diabetes -
https://hy.wikipedia.org/wiki/%D5%87%D5%A1%D6%84%D5%A1%D6%80%D5%A1%D5%B5%D5%AB%D5%B6_%D5%A4%D5%AB%D5%A1%D5%A2%D5%A5%D5%BF

About KNN -
https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20(KNN)%20algorithm%20is%20a%20non,used%20in%20machine%20learning%20today.

About Decision Tree – https://www.ibm.com/topics/decision-trees

About SVM - https://www.ibm.com/docs/en/spss-modeler/saas?topic=node-svm-expert-options