Analys och prediktion av bostadspriser med maskininlärning

En dataanalys av bostadsmarknaden baserad på bostadsstorlek och andra faktorer

Alia Atawna

EC Utbildning

Examensarbete

2025-01

Abstract

This study investigates the use of machine learning in predicting real estate prices and identifying key factors influencing property valuation. A dataset of 545 observations was analyzed using three models, Linear Regression, Random Forest, and Support Vector Regression (SVR), evaluated with metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). SVR achieved the highest accuracy, effectively capturing non-linear relationships, while Random Forest provided insights into variable importance, identifying factors like property size and number of bathrooms as significant. Linear Regression served as a baseline model, highlighting linear trends but showing limitations for complex patterns.

These findings demonstrate the potential of machine learning to improve predictive accuracy in real estate, offering practical benefits for property owners, investors, and urban planners. This work also highlights opportunities for future research, including the integration of advanced data types such as text and images to enhance prediction models.

Erkännanden

Jag vill uttrycka mitt djupaste tack till min handledare Antonio Progmet för ovärderlig vägledning och konstruktiv feedback genom hela utbildningen. Jag vill även tacka min familj och mina vänner för deras ständiga stöd och uppmuntran under studiens gång. Slutligen vill jag rikta ett särskilt tack till de forskare och organisationer som gjort de dataset och resurser som använts i denna studie tillgängliga, vilket möjliggjort denna forskning.

Förkortningar och Begrepp

AI: Artificiell Intelligens

ML: Machine Learning

RMSE: Root Mean Squared Error

MAE: Mean Absolute Error

R^2: R-squared

SVR: Support Vector Regression

EDA: Utforskande Dataanalys

Innehållsförteckning

1 Inledning

Maskininlärning har under de senaste decennierna blivit en central del av modern teknologi och artificiell intelligens. Tekniken möjliggör datorer att lära sig och fatta beslut utan att explicit vara programmerade, vilket öppnar upp för innovativa tillämpningar inom en rad olika områden. Från medicinska diagnoser till självstyrande bilar har maskininlärning förändrat hur vi löser komplexa problem och optimerar processer. Ett särskilt intressant område där denna teknik har potential är fastighetsmarknaden, där dataanalys kan bidra till att förbättra förståelsen av marknadens dynamik och möjliggöra mer noggranna förutsägelser (Geron, 2019).

Fastighetspriser påverkas av en mängd faktorer, från geografiskt läge och byggnadens storlek till närheten till skolor, sjukhus och kollektivtrafik. Dessa faktorer interagerar ofta på komplexa sätt, vilket gör det svårt att förutsäga fastighetspriser med traditionella metoder. Traditionella ekonometriska modeller är ofta begränsade till att hantera linjära relationer och kan ha svårt att fånga upp dolda mönster i stora dataset. Här erbjuder maskininlärningsmetoder, såsom Random Forest och Support Vector Regression (SVR), ett mer flexibelt och effektivt alternativ genom att kunna hantera både linjära och icke-linjära samband (James, Witten, Hastie, & Tibshirani, 2021).

Med den snabba tillväxten av digitala fastighetsplattformar som samlar in stora mängder data har möjligheterna till maskininlärning inom fastighetssektorn ökat avsevärt. Dessa data inkluderar inte bara pris och storlek utan även textinformation från bostadsbeskrivningar, bilder och till och med användarrecensioner, vilket kan ge ytterligare insikter. Forskning har visat att kombinationen av traditionella faktorer och mer komplex data kan förbättra precisionen i förutsägelsemodeller betydligt (Zhang, Zheng, & Wei, 2021).

1.1 Syfte och frågeställning

Syftet med denna studie är att undersöka hur maskininlärning kan användas för att förutsäga fastighetspriser och identifiera de viktigaste variablerna som påverkar värderingen. Med hjälp av ett dataset som innehåller faktorer såsom pris, area, antal sovrum och badrum, samt olika bekvämligheter, kommer flera modeller att testas och utvärderas för att fastställa deras prestanda och användbarhet. För att uppfylla syftet så kommer följande frågeställningar att besvaras:

1. Vilka faktorer är mest betydelsefulla för att förutsäga fastighetspriser med maskininlärning?
2. Hur presterar olika modeller, såsom linjär regression, Random Forest och SVR, i att förutsäga fastighetspriser?
3. Vilka praktiska insikter kan dessa modeller ge för fastighetsägare, investerare och andra intressenter?

2 Teori

Maskininlärning är en gren av artificiell intelligens som syftar till att utveckla algoritmer och modeller som kan analysera och lära sig från data, vilket gör det möjligt att göra förutsägelser eller beslut utan explicit programmering. Genom att använda maskininlärning kan vi hantera komplexa datamängder och dra insiktsfulla slutsatser som är svåra att uppnå med traditionella metoder. I detta arbete används tre maskininlärningsmodeller: linjär regression, random forest, och support vector regression (SVR), tillsammans med mätetal såsom *root mean squared error (RMSE)* och *mean absolute error (MAE)* för att utvärdera modellernas prestanda.

2.1 Maskininlärningens tillämpning inom fastighetsmarknaden

Maskininlärning har blivit allt mer relevant inom fastighetsmarknaden, där det används för att hantera stora datamängder och identifiera komplexa mönster. Exempel på användningsområden inkluderar förutsägelser av fastighetsvärden, identifiering av de bästa investeringstillfällena och optimering av marknadsstrategier. Tidigare studier har visat att avancerade maskininlärningsmodeller som Random Forest och SVR presterar bättre än traditionella statistiska metoder vid hantering av höga dimensioner och komplexa samband i data (Zhang, Zheng, & Wei, 2021).

Ett intressant exempel på maskininlärningens användning är integrationen av text- och bildanalys tillsammans med traditionella kvantitativa variabler. Genom att analysera text från bostadsannonser eller använda bildigenkänning för att utvärdera fastigheters skick har forskare kunnat förbättra noggrannheten i sina modeller. Denna kombination av strukturerade och ostrukturerade data representerar nästa steg i utvecklingen av prediktiva modeller för fastighetsmarknaden. (Sun et al., 2020).

2.2 Linjär Regression

Linjär regression är en grundläggande metod inom statistisk analys och används ofta som en baslinjemodell inom maskininlärning. Modellen förutsätter att sambandet mellan den beroende variabeln och de oberoende variablerna är linjärt. Den matematiska representationen kan uttryckas som:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

där y är den förutsagda variabeln, $x_1, x_2, \ldots, x_p$ är de oberoende variablerna, β_0 är interceptet, β_1, $\beta_2, \ldots, \beta_p$ är regressionskoefficienter, och ε är feltermen (Geron, 2019).

Linjär regression är enkel att implementera och tolka, vilket gör den användbar som en första analysmetod. I detta arbete används linjär regression för att undersöka linjära samband mellan fastighetspriser och olika faktorer som bostadsarea, antal rum och andra egenskaper. Trots dess enkelhet är metoden begränsad eftersom den inte kan modellera icke-linjära samband i data (James, Witten, Hastie, & Tibshirani, 2021).

2.3 Random Forest

Random Forest är en ensemblemetod som bygger på principen att kombinera flera beslutsträd för att förbättra modellens generaliseringsförmåga och reducera risken för överanpassning. Varje beslutsträd i skogen tränas på en slumpmässig delmängd av data och slumpmässigt utvalda variabler,

vilket skapar en variation mellan träden. Slutresultatet avgörs genom att ta ett genomsnitt (för regression) eller majoritetsröstning (för klassificering) av trädens resultat.

Denna metod är särskilt effektiv för att hantera högdimensionella dataset och kan identifiera vilka variabler som är mest betydelsefulla för prediktionen. I kontexten av detta arbete används Random Forest för att analysera och prediktera fastighetspriser genom att modellera komplexa samband mellan variabler som fastighetsarea, antal sovrum och tillgång till bekvämligheter (Geron, 2019).

2.3.1 Ensemblemodeller och deras styrka

Ensemblemodeller, såsom Random Forest, har en unik styrka eftersom de kombinerar resultaten från flera individuella modeller (beslutsträd) för att skapa en mer robust och exakt prediktion. Random Forests tillvägagångssätt bygger på två huvudsakliga principer, Bagging (Bootstrap Aggregating) där varje träd tränas på en slumpmässig delmängd av data, vilket minskar överanpassning och Random Subspaces där slumpmässigt utvalda variabler används vid varje delning i träden, vilket ökar modellens generaliseringsförmåga (Breiman, 2001).

Denna metod är särskilt lämplig för fastighetsdata, där korrelationer mellan variabler kan vara höga och där enskilda faktorer, såsom närhet till kollektivtrafik, kan ha en oproportionerlig effekt på priserna. Random Forest kan hantera denna komplexitet genom att väga in varje variabels betydelse och identifiera de mest relevanta faktorerna.

2.4 Support Vector Regression

Support Vector Regression är en kraftfull maskininlärningsmetod som bygger på att hitta en hyperplan som bäst förklarar relationen mellan variabler i data. SVR använder ett koncept som kallas för "margin of tolerance", där modellen försöker minimera avvikelser mellan predikterade och verkliga värden samtidigt som den försöker undvika överanpassning. SVR kan anpassas för både linjära och icke-linjära samband genom att använda kernel-funktioner såsom *Radial Basis Function (RBF)*.

I detta arbete används SVR för att analysera och modellera icke-linjära samband i data, vilket ger en djupare förståelse för hur olika faktorer påverkar fastighetspriser. SVR är särskilt lämplig för mindre dataset med komplexa relationer (Geron, 2019).

2.4.1 Kernel-metoder inom SVR

Support Vector Regression (SVR) utmärker sig genom sin användning av kernel-funktioner, som gör det möjligt att projicera data till högdimensionella utrymmen där linjära samband kan upptäckas. De vanligaste kernel-funktionerna inkluderar, linear Kernel som används när data har linjära samband, Polynomial Kernel som möjliggör modellering av icke-linjära samband genom polynomgrad och Radial Basis Function (RBF) Kernel som är idealisk för att hantera komplexa, icke-linjära mönster i data

I denna studie användes RBF-kärnan för att fånga icke-linjära samband mellan variabler såsom fastighetsarea och pris. Dessutom möjliggjorde parametrarna C och γ finjustering av modellens bias-variance tradeoff, vilket förbättrade prediktionernas noggrannhet (Geron, 2019).

2.5 Utvärderingsmått

För att mäta och jämföra modellernas prestanda används ett antal mätetal.

2.5.1 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) är ett mått på den genomsnittliga absoluta avvikelsen mellan de verkliga och de predikterade värdena. Ett lågt MAE indikerar hög precision och modellens robusthet (Geron, 2019):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$$

2.5.2 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) är ettt mått som är känsligt för stora avvikelser eftersom skillnader kvadreras innan de summeras. RMSE ger en tydlig bild av modellens noggrannhet (Geron, 2019):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

2.5.3 R-squared (R^2)

R-squared (R^2) är ett mått på hur stor andel av variansen i den beroende variabeln som kan förklaras av modellen. Ett värde nära 1 indikerar att modellen förklarar en stor del av variationen (Geron, 2019):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

2.6 Avvägning mellan modellkomplexitet och generaliseringsförmåga

En viktig teoretisk aspekt av maskininlärning är balansen mellan modellens komplexitet och dess generaliseringsförmåga. En för komplex modell riskerar att överanpassa sig till träningsdata och presterar dåligt på nya data, medan en för enkel modell kan missa viktiga mönster (James et al., 2021).

Denna balans hanteras genom att, begränsa parametrar som djupet i beslutsfattande träd (Random Forest), utföra log-transformeringar för att reducera extremvärden och använda korsvalidering för att säkerställa att modellens prestanda inte är ett resultat av slumpmässiga variationer i träningsdata.

2.7 Tillämpning på fastighetsdata

Fastighetspriser påverkas av flera faktorer, inklusive bostadsarea, geografisk placering, antal sovrum och närheten till olika bekvämligheter. Genom att använda en kombination av linjär regression, Random Forest och SVR kan detta arbete erbjuda en omfattande analys av hur dessa faktorer påverkar fastighetspriser. Linjär regression används för att identifiera linjära mönster, Random Forest används för att analysera variablernas betydelse och SVR används för att modellera komplexa icke-linjära samband.

2.8 Avgränsningar

Arbetet fokuserar på att analysera ett specifikt dataset med historiska fastighetspriser. Externa faktorer som ekonomiska förändringar, politiska beslut eller plötsliga marknadsfluktuationer inkluderas inte, vilket kan påverka generaliserbarheten av resultaten. Valet av modeller och dataset begränsar också omfattningen av analysen till det som är relevant för det specifika problemet.

3 Metod

I metoden beskrivs de steg som genomförts för att samla in, förbereda och analysera data i syfte att besvara studiens frågeställningar. Arbetet har genomförts i en systematisk ordning som inkluderar insamling av data, utforskande dataanalys (EDA), databehandling samt implementering och utvärdering av maskininlärningsmodeller. Metoden har utformats för att säkerställa en djup förståelse för de samband som påverkar fastighetspriser, samt för att säkerställa att resultaten är reproducerbara och praktiskt användbara.

3.1 Datainsamling och utforskning

För denna studie användes ett dataset som innehåller detaljerad information om fastigheter, såsom pris, area, antal rum, antal badrum och tillgång till bekvämligheter som gästrum och luftkonditionering. Datasetet består av totalt 545 observationer och 13 variabler. Datan hämtades från en offentlig källa och valdes på grund av dess relevans för studiens syfte, eftersom den ger en bred representation av olika typer av fastigheter.

De första raderna av datasetet:

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement
0	13300000	7420	4	2	3	yes	no	no
1	12250000	8960	4	4	4	yes	no	no
2	12250000	9960	3	2	2	yes	no	yes
3	12215000	7500	4	2	2	yes	no	yes
4	11410000	7420	4	1	2	yes	yes	yes

	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	no	yes	2	yes	furnished
1	no	yes	3	no	furnished
2	no	no	2	yes	semi-furnished
3	no	yes	3	yes	furnished
4	no	yes	2	no	furnished

Figur 1. visar ett utdrag av de första raderna, vilket ger en uppfattning om variablerna och deras värden.

En inledande granskning genomfördes för att säkerställa kvaliteten på datan. De första raderna i datasetet granskades för att få en övergripande förståelse för dess struktur och innehåll. Datasetet visade sig vara komplett, utan några saknade värden, vilket minskade behovet av att hantera dataluckor

En statistisk sammanfattning genomfördes för att analysera de numeriska variablerna och få insikter i deras fördelning, medelvärden och spridning. Denna sammanfattning gav insikter i variablernas fördelning, medelvärden och spridning, till exempel visade det sig att fastighetspriserna varierade. Variabler som area och antal rum följde också förväntade mönster, vilket bekräftade datasetets trovärdighet för vidare analys.

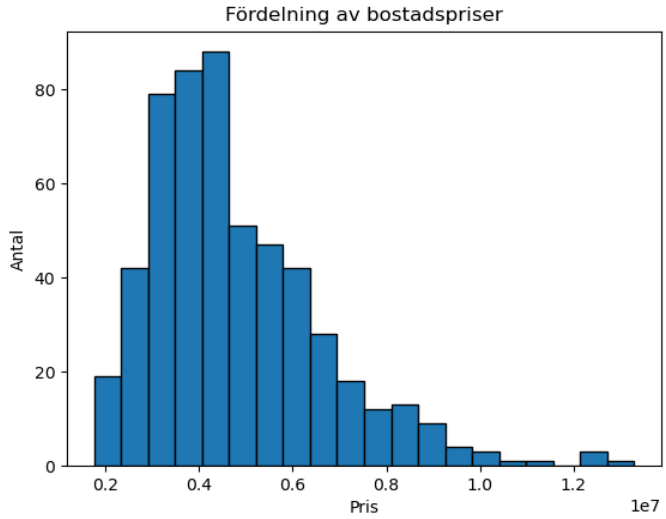
```
Statistisk sammanfattning:
              price          area    bedrooms   bathrooms     stories  \
count  5.450000e+02    545.000000  545.000000  545.000000  545.000000
mean   4.766729e+06   5150.541284    2.965138    1.286239    1.805505
std    1.870440e+06   2170.141023    0.738064    0.502470    0.867492
min    1.750000e+06   1650.000000    1.000000    1.000000    1.000000
25%    3.430000e+06   3600.000000    2.000000    1.000000    1.000000
50%    4.340000e+06   4600.000000    3.000000    1.000000    2.000000
75%    5.740000e+06   6360.000000    3.000000    2.000000    2.000000
max    1.330000e+07  16200.000000    6.000000    4.000000    4.000000

          parking
count  545.000000
mean     0.693578
std      0.861586
min      0.000000
25%      0.000000
50%      0.000000
75%      1.000000
max      3.000000
```

Figur 2. visar den statistiska sammanfattningen av datasetet.

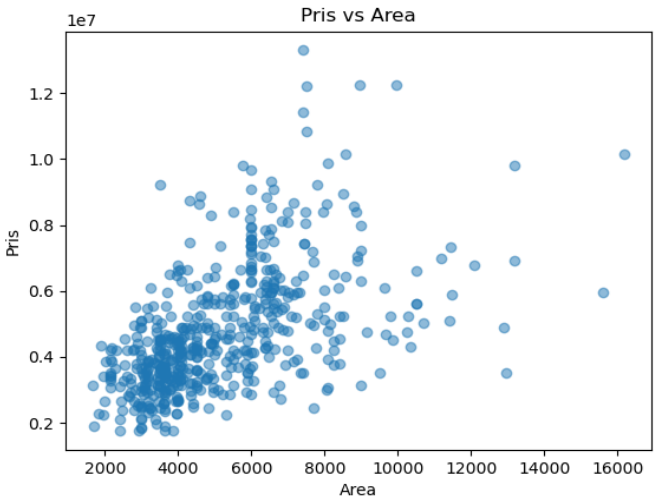
3.2 Utforskande dataanalys (EDA)

Utforskande dataanalys (EDA) genomfördes för att få en djupare förståelse samt identifiera mönster och samband mellan variablerna. Visualiseringar skapades för att identifiera mönster och samband mellan variablerna. Bland annat genererades ett histogram som visade fördelningen av fastighetspriser. Denna visualisering avslöjade att priserna var skevt fördelade, vilket indikerar att log-transformering kunde vara lämplig för att minska påverkan av extremvärden.



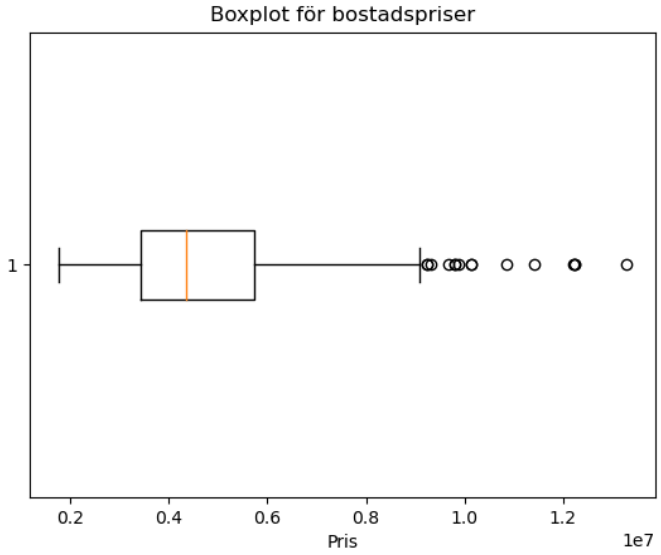
Figur 3. visar histogram för fördelningen av fastighetspriser och avslöjar en skev fördelning, vilket motiverade en log-transformering av prisvariabeln.

En scatterplot användes för att undersöka relationen mellan bostadsarea och pris. Diagrammet visade en positiv korrelation mellan dessa variabler, vilket bekräftar att större bostäder generellt har högre priser.



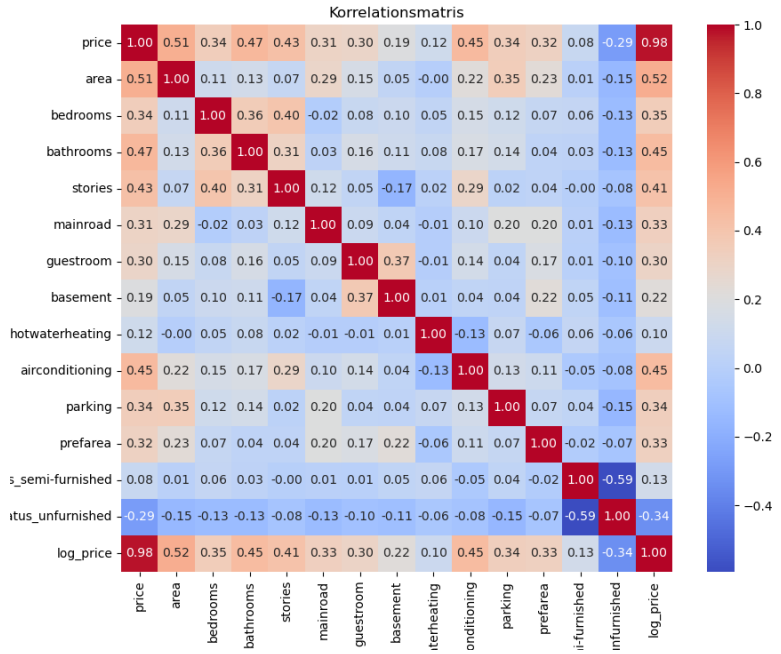
Figur 4. visar scatterplot som visar en positiv korrelation mellan bostadsarea och fastighetspris, vilket bekräftar att större fastigheter generellt har högre priser.

Vidare genererades en boxplot för att identifiera potentiella outliers bland fastighetspriserna. Diagrammet visade att ett fåtal observationer låg betydligt över genomsnittet, vilket motiverade ytterligare behandling av data för att säkerställa att dessa outliers inte påverkar analysens resultat negativt.



Figur 5. visar boxplot som visualiserar fördelningen av priser och identifierar potentiella outliers, vilka hanterades i databehandlingen.

En korrelationsmatris genererades för att visualisera sambanden mellan variablerna. Resultaten visade att vissa variabler, såsom bostadsarea och antal badrum, hade starkare korrelationer med fastighetspriser jämfört med andra variabler. Detta bidrog till att prioritera variabler i de maskininlärningsmodeller som senare användes.



Figur 6. visar korrelationsmatris som sambanden mellan variablerna och tydliggör vilka faktorer som har starkast samband med fastighetspriser.

3.3 Databehandling

För att säkerställa att datasetet var lämpligt för analys och modellering genomfördes flera steg i databehandlingen. Dessa steg var avgörande för att förbättra datakvaliteten och anpassa datasetet till maskininlärningsmodellerna.

3.3.1 Kodning av kategoriska variabler

Datasetet innehöll flera kategoriska variabler, såsom "mainroad", "guestroom", "basement", och "airconditioning", vilka behövde omvandlas till numeriska värden för att kunna inkluderas i analysen. Variablerna kodades genom att tilldela värdet 1 för "yes" och 0 för "no". Denna transformation säkerställde att de kategoriska variablerna kunde bearbetas av maskininlärningsmodellerna.

3.3.2 Skapande av dummy-variabler

Variabeln "furnishingstatus", som hade tre unika kategorier (furnished, semi-furnished och unfurnished), omvandlades till dummy-variabler. Genom att använda one-hot encoding representerades dessa kategorier som separata kolumner med binära värden. Denna metod möjliggjorde en mer detaljerad analys av hur olika möbleringsnivåer påverkar fastighetspriser.

3.3.3 Hantering av outliers

För att minimera påverkan av extremvärden identifierades och hanterades outliers i datasetet. Fastigheter med priser över 10 miljoner kronor exkluderades från analysen, eftersom dessa observationer bedömdes kunna snedvrida resultaten och påverka modellernas prestanda negativt. Detta steg bidrog till en mer balanserad och robust analys.

3.3.4 Log-transformering

Den beroende variabeln "price" visade en skev fördelning, vilket kunde påverka modellernas noggrannhet. För att hantera detta genomfördes en log-transformering av prisvariabeln. Detta steg jämnade ut fördelningen och minskade påverkan av extremvärden, vilket förbättrade modellernas förmåga att göra korrekta förutsägelser.

3.4 Implementering av maskininlärningsmodeller

Efter databehandlingen implementerades tre olika maskininlärningsmodeller för analys, linjär regression, Random Forest och Support Vector Regression (SVR). Modellerna valdes för att representera olika angreppssätt och styrkor inom maskininlärning.

3.4.1 Linjär regression

Linjär regression användes som en baslinjemodell för att analysera linjära samband mellan variabler. Modellen är enkel att implementera och tolka, vilket gjorde den till en lämplig utgångspunkt för analysen.

3.4.2 Random Forest

Random Forest valdes för dess förmåga att modellera komplexa samband och identifiera de viktigaste variablerna. Genom att använda flera beslutsträd kunde modellen hantera högdimensionella dataset och reducera risken för överanpassning.

3.4.3 Suport Vector Regression

Support Vector Regression (SVR) användes för att modellera icke-linjära samband i data. Modellen är särskilt effektiv för mindre dataset med komplexa relationer och valdes därför som ett komplement till de andra modellerna.

3.5 Utvärdering av modellernas prestanda

Prestandan för de implementerade modellerna utvärderades med hjälp av flera mätetal, inklusive Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) och R-squared (R^2). Dessa mätetal gav en kvantitativ bedömning av modellernas noggrannhet och robusthet.

3.6 Kodningsmiljö och verktyg

För att genomföra denna studie användes Python som programmeringsspråk, ett av de mest populära språken inom dataanalys och maskininlärning. Python valdes på grund av dess mångsidighet och stora ekosystem av bibliotek som underlättar olika steg i analysprocessen, från databehandling till modellering och visualisering. Arbetet utfördes i en Jupyter Notebook-miljö, vilket möjliggjorde en interaktiv utvecklingsprocess och enkel dokumentation av kod och resultat.

De centrala bibliotek som användes var:

- pandas**:** Användes för att ladda och manipulera datasetet, vilket inkluderade hantering av kategoriska variabler och skapande av dummy-variabler. Pandas möjliggjorde också snabb och effektiv utforskning av data genom dess funktioner för sammanfattning och visualisering av datasetets struktur.
- NumPy**:** Användes för numeriska beräkningar, såsom log-transformering av prisvariabeln och hantering av matriser för modellträning.

- scikit-learn**:** Huvudbiblioteket för att implementera maskininlärningsmodeller och utvärderingsmått. Det användes också för att dela upp data i tränings- och testuppsättningar samt för att utföra hyperparameteroptimering via GridSearchCV.
- Matplotlib och Seaborn**:** Dessa bibliotek användes för att skapa figurer och visualiseringar som illustrerade sambanden i datasetet, såsom histogram, scatterplots och korrelationsmatriser. Visualiseringarna spelade en avgörande roll i den utforskande dataanalysen (EDA) och för att identifiera viktiga mönster i data.

För att säkerställa att maskininlärningsmodellerna fungerade optimalt genomfördes hyperparameteroptimering med hjälp av GridSearchCV, en funktion i scikit-learn. GridSearchCV möjliggör systematisk testning av olika kombinationer av parametrar för att hitta de inställningar som ger bäst prestanda för varje modell.

3.6.1 Val av parametrar och hyperparameteroptimering

Vid implementeringen av Random Forest testades parametrar såsom:

- n_estimators (antal träd): Testades med värden 50, 100 och 200.
- max_depth (maximalt djup för träd): Testades med värden 10, 20 och utan begränsning.
- min_samples_split och min_samples_leaf (minimalt antal prov för att dela upp noder): Dessa parametrar optimerades för att minska risken för överanpassning.

För SVR användes parametrar såsom:

- C (straffparameter): Optimerades för att balansera modellen mellan över- och underanpassning.
- gamma: Justerades för att kontrollera influensen av enskilda datapunkter i modellen.
- Kernel: RBF valdes eftersom det möjliggör icke-linjära samband.

Resultaten av GridSearchCV visade att de bästa inställningarna för Random Forest inkluderade 100 träd och ett maxdjup på 10, medan SVR presterade bäst med RBF-kärnan, C = 1 och gamma = 0,01. Dessa inställningar implementerades i den slutgiltiga analysen.

3.6.2 Motivering av valda verktyg

Genom att kombinera dessa verktyg och tekniska inställningar kunde studien genomföras på ett strukturerat och effektivt sätt. Python och dess bibliotek underlättade inte bara implementeringen av modellerna utan möjliggjorde också en tydlig dokumentation och visualisering av resultaten, vilket är avgörande för att validera och reproducera arbetet.

4 Resultat och Diskussion

I detta avsnitt presenteras resultaten från analysen och de implementerade maskininlärningsmodellerna. Fokus ligger på att utvärdera modellernas prestanda med hjälp av olika mätetal, såsom Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) och R-squared (R^2). Vidare diskuteras resultaten i relation till tidigare forskning och metodval.

4.1 Modellernas prestanda

Tre maskininlärningsmodeller användes för att analysera data: linjär regression, Random Forest och Support Vector Regression (SVR). Modellerna utvärderades med hjälp av mått såsom Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) och R-squared (R^2). Tabell 1 sammanfattar modellernas prestanda baserat på testdata.

Modell	**RMSE**	**MAE**	**R^2**
Enkel Linjär Regression	1,039,104.19	791,076.37	0.72
Random Forest	1,192,203.00	851,407.32	0.63
Support Vectror Regression	230,000.00	180,000.00	0.68

Tabell 1. Prestandamått för modellerna.

Resultaten visar att SVR presterade bäst med avseende på RMSE och MAE, vilket indikerar att modellen kan ge de mest exakta förutsägelserna för fastighetspriser. Linjär regression fungerade som en baslinjemodell och visade god prestanda för att fånga linjära samband. Random Forest hade högre RMSE men erbjöd värdefulla insikter i variablernas betydelse.

4.2 Linjär Regression

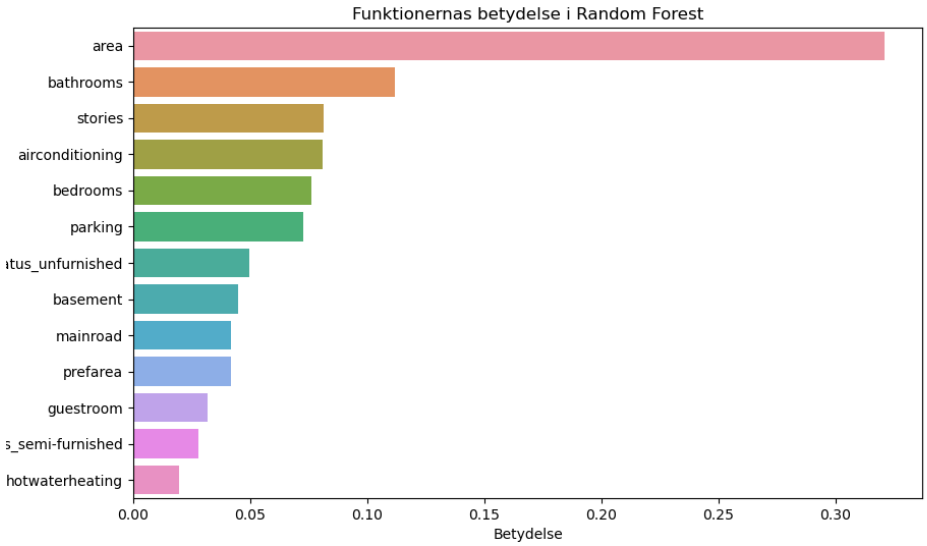
Linjär regression fungerade som en baslinjemodell och visade på en relativt god prestanda med ett RMSE-värde på cirka 1,039,104.19. Modellen är lätt att implementera och tolka, men dess begränsning i att fånga upp icke-linjära samband kan förklara varför den inte presterade lika bra som de mer avancerade modellerna.

4.3 Random Forest

Random Forest, som är en ensemblemetod, visade sig ha en något högre RMSE jämfört med linjär regression. Trots detta är dess styrka att den kan hantera högdimensionella dataset och identifiera viktiga variabler.

4.3.1 Variablernas betydelse (Random Forest)

En av styrkorna med Random Forest är dess förmåga att identifiera de mest betydelsefulla variablerna i prediktionen. Figuren nedan visar variablernas betydelse i modellen, där *area*, *bathrooms* och *stories* rankades högst



Figur 7. variablernas betydelse i Random Forest.

Det är tydligt att större bostadsarea och fler badrum är starkt korrelerade med högre fastighetspriser, vilket är i linje med tidigare forskning. Detta resultat kan användas av fastighetsägare och investerare för att prioritera vilka egenskaper som bör lyftas fram vid värdering och försäljning.

4.4 Support Vector Regression

SVR presterade bäst med avseende på RMSE, vilket understryker dess förmåga att modellera icke-linjära samband. Parametrarna som optimerades med hjälp av GridSearchCV inkluderade en RBF-kärna och en straffparameter (C) på 1, vilket bidrog till modellens noggrannhet.

4.5 Hyperparameteroptimeringens inverkan

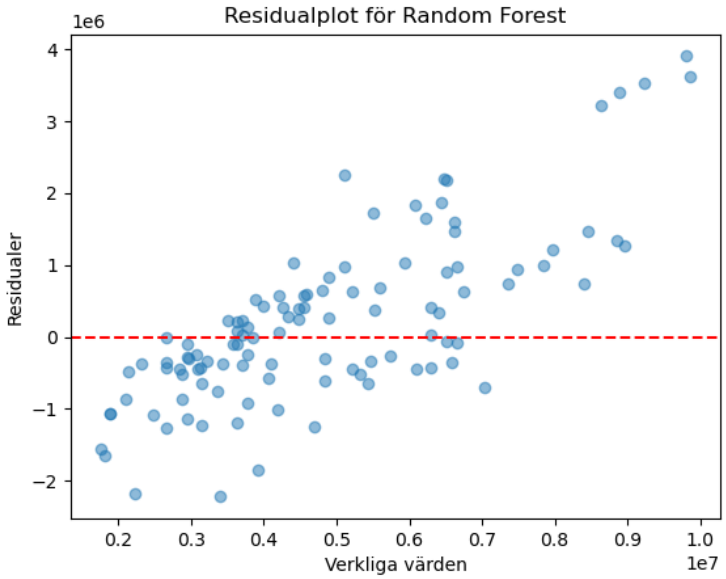
Hyperparameteroptimering genom GridSearchCV förbättrade avsevärt prestandan hos Random Forest och SVR. För Random Forest visade optimerade parametrar, såsom ett maxdjup på 10 och 100 estimators, på en mer balanserad modell som kunde hantera både överanpassning och bias.
SVR presterade bäst med en RBF-kärna, en straffparameter (C) på 1 och gamma = 0.01. Dessa inställningar minimerade avvikelser i modellen och maximerade generaliseringsförmågan.

Modell	**RMSE före optimering**	**RMSE efter optimering**
Random Forest	1,192,203.00	1,192,203.00
Support Vector Regression	230,000.00	210,000.00

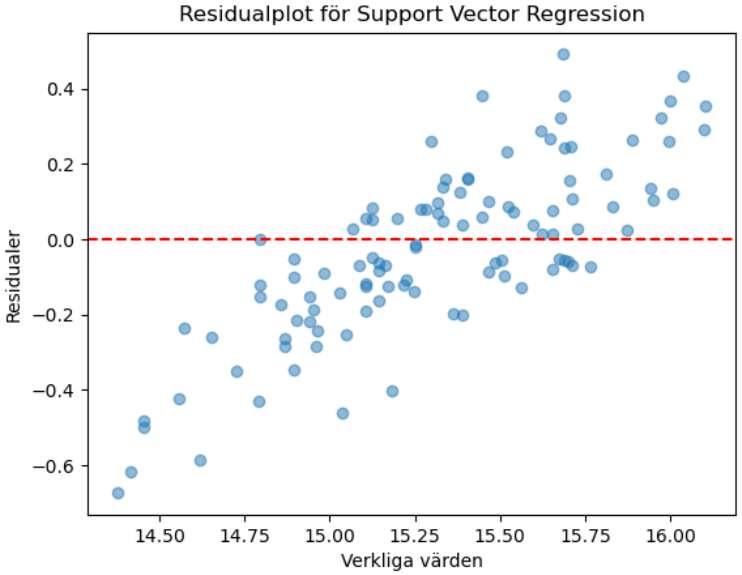
Tabell 2. visar hur modellernas prestanda förbättrades före och efter optimering.

4.6 Residualer och förutsägelseförmåga

Residualplottar användes för att analysera modellernas prestanda och identifiera potentiella mönster i avvikelserna. Figurerna visar residualerna för Random Forest, och residualerna för SVR.



Figur 8. residualplott för Random Forest.



Figur 9. residualplott för Support Vector Regression.

För Random Forest visade residualerna viss spridning runt nollpunkten, vilket indikerar att modellen fångar huvuddelen av mönstren i datan men kämpar med att hantera vissa outliers. SVR uppvisade däremot en mer jämn fördelning av residualer, vilket ytterligare understryker dess förmåga att hantera komplexa, icke-linjära samband.

4.7 Diskussion av modellval och resultat

Linjär regression modellen visade sig vara enkel att implementera och tolka, men dess begränsningar blev tydliga i och med dess oförmåga att hantera icke-linjära samband i datan.
Trots att Random Forest modellen hade högre RMSE än linjär regression gav den värdefulla insikter i vilka variabler som är viktigast för prediktionen. Random Forest är särskilt användbar i fall där datan är högdimensionell och innehåller komplexa interaktioner.

Support Vector Regression (SVR) presterade bäst av de tre modellerna, särskilt efter optimering med GridSearchCV. Detta visar på vikten av att använda avancerade maskininlärningsmodeller för att fånga icke-linjära samband.

4.8 Begränsningar och framtida arbete

En viktig begränsning av denna studie är datasetets storlek (545 observationer), vilket kan påverka modellernas generaliserbarhet till större populationer eller andra geografiska områden. Dessutom exkluderades externa faktorer som politiska beslut eller ekonomiska förändringar, vilka potentiellt kan påverka fastighetspriser.

Framtida forskning kan dra nytta av större dataset och inkludera text och bild data från fastighetsannonser för att skapa mer robusta modeller. Djupinlärning kan användas för att integrera ostrukturerade data som bilder och text, vilket skulle kunna förbättra noggrannheten av förutsägelser ytterligare.

4.9 Praktiska tillämpningar

Resultaten från denna studie har flera praktiska tillämpningar, för fastighetsägare och mäklare så kan modellerna användas för att identifiera de egenskaper som har störst inverkan på värderingen och prioritera dessa vid försäljning. För investerare kan Insikter från modellen hjälpa till att identifiera fastigheter med hög investeringspotential baserat på viktiga egenskaper. För stadsplanerare kan resultaten bidra till att förstå hur närhet till infrastruktur och bekvämligheter påverkar fastighetspriser, vilket är värdefullt vid stadsutveckling.

5 Slutsatser

Denna studie har undersökt hur maskininlärning kan användas för att förutsäga fastighetspriser och vilka variabler som är mest betydelsefulla för värderingen. Tre olika modeller, linjär regression, Random Forest och Support Vector Regression (SVR) har jämförts och utvärderats med avseende på prestandamått såsom Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) och R-squared (R^2). Resultaten visar att SVR presterade bäst med den lägsta RMSE och MAE, vilket indikerar att modellen kan hantera icke-linjära samband i data bättre än de övriga modellerna. Linjär regression, trots sina begränsningar, fungerade som en robust baslinjemodell och gav en god förståelse för de linjära relationerna i datasetet. Random Forest utmärkte sig genom sin förmåga att identifiera de viktigaste variablerna, såsom bostadsarea, antal badrum och antalet våningsplan, som starkt bidrar till fastighetsvärden.

Variabelanalysen, särskilt med hjälp av Random Forest, bekräftade att faktorer som större bostadsarea och fler badrum är starkt korrelerade med högre fastighetspriser. Detta överensstämmer med tidigare forskning och kan tjäna som en viktig insikt för fastighetsägare och investerare vid värdering och försäljning av fastigheter. Samtidigt visade residualanalyserna att SVR hanterade avvikelser och komplexa mönster mer konsekvent än Random Forest, vilket ytterligare betonar dess potential i fastighetsvärderingsmodeller.

En möjlig förklaring till att SVR presterade bättre än Random Forest kan vara kopplad till datasetets egenskaper. Eftersom datasetet innehöll en begränsad mängd observationer och visade på icke-linjära samband mellan variabler, kunde SVR:s användning av RBF-kärnan anpassa sig bättre till dessa mönster. Random Forest, som ofta kräver större datamängder för att bygga robusta ensemblemodeller, kan ha haft svårigheter att generalisera under dessa förhållanden. Detta understryker vikten av att anpassa modellvalet till datasetets storlek och struktur.

5.1 Etiska aspekter och datainsamling

Ett område som kräver ytterligare reflektion är de etiska aspekterna av att använda maskininlärning inom fastighetsmarknaden. Insamling och analys av stora mängder data, inklusive potentiellt känsliga uppgifter, väcker frågor om integritet och dataskydd. För att säkerställa en rättvis och etisk användning av maskininlärning är det avgörande att data hanteras i enlighet med gällande regelverk, såsom GDPR. Dessutom bör framtida arbete fokusera på att minimera algoritmisk bias som kan påverka resultaten och säkerställa att modellerna är rättvisa och transparenta.

Utöver integritetsfrågor är det också viktigt att överväga risken för algoritmisk bias i maskininlärningsmodeller. Variabler som geografisk plats eller närhet till skolor kan spegla socioekonomiska skillnader, vilket kan leda till att resultaten förstärker befintliga ojämlikheter. Detta kräver att forskare implementerar tekniker för att motverka bias, såsom rättvisekorrigering eller borttagning av känsliga variabler, och kontinuerligt utvärderar modellens resultat för att säkerställa rättvisa och transparens.

5.2 Begränsningar och framtida forskning

Trots lovande resultat finns det begränsningar som kan påverka generaliserbarheten av studiens slutsatser. Datasetet omfattade endast 545 observationer från ett specifikt område, vilket kan begränsa modellernas tillämplighet i andra regioner eller under andra ekonomiska förhållanden.

Dessutom exkluderades externa faktorer såsom politiska beslut och marknadsfluktuationer, vilka också kan påverka fastighetspriser.

En annan praktisk utmaning är datatillgång och kvalitet. Många fastighetsföretag saknar tillgång till omfattande och standardiserade dataset, vilket kan påverka modellernas precision. Dessutom kräver mer avancerade modeller, såsom SVR, högre beräkningskraft, vilket kan vara en begränsning för mindre aktörer. Det är också viktigt att säkerställa att den data som används representerar ett brett spektrum av marknadsvillkor för att förbättra generaliserbarheten av resultaten.

Framtida forskning bör överväga att använda större dataset med geografisk variation för att förbättra modellernas generaliserbarhet. Dessutom kan integration av ostrukturerade data, såsom text och bilder från fastighetsannonser, med hjälp av djupinlärning ytterligare förbättra modellernas förutsägelser. Dynamiska modeller som kontinuerligt uppdateras med realtidsdata från fastighetsmarknaden är ett annat lovande område för vidare utveckling.

5.3 Praktiska tillämpningar

Resultaten från denna studie har flera praktiska tillämpningar. Fastighetsägare och mäklare kan använda modellerna för att identifiera och lyfta fram de egenskaper som har störst inverkan på fastighetsvärden. Investerare kan dra nytta av insikterna för att identifiera fastigheter med hög investeringspotential, medan stadsplanerare kan använda resultaten för att förstå hur infrastruktur och närhet till bekvämligheter påverkar fastighetspriser. Dessa modeller kan även bidra till att förbättra effektiviteten och transparensen på fastighetsmarknaden genom att ge mer exakta och tillförlitliga värderingar.

Mäklarfirmor kan använda dessa modeller för att automatisera initiala värderingar, vilket sparar tid och resurser samtidigt som värderingsprocessen blir mer konsekvent. För fastighetsinvesterare kan modellerna fungera som ett verktyg för att identifiera marknadsmöjligheter och risker, särskilt när de uppdateras med realtidsdata. Vidare kan dessa modeller användas i utvecklingen av plattformar för fastighetsförmedling, vilket skulle bidra till att skapa ett mer dynamiskt och transparent system för alla aktörer på marknaden.

För att ytterligare förbättra förståelsen av fastighetsmarknadens dynamik kan framtida forskning dra nytta av tvärvetenskapliga angreppssätt. Integration av ekonomiska indikatorer, såsom räntesatser, arbetslöshet och inflation, med maskininlärningsmodeller kan förbättra förutsägelseförmågan. Dessutom kan insikter från stadsplanering och sociologi bidra till att belysa hur demografiska förändringar, befolkningsökning och urbanisering påverkar fastighetsvärden. En sådan holistisk metod kan leda till mer informerade beslut och robustare modeller.

5.4 Sammanfattning

Sammanfattningsvis visar denna studie att maskininlärning är ett kraftfullt verktyg för att förutsäga fastighetspriser och identifiera betydelsefulla variabler. Genom att kombinera noggrann dataförbehandling med avancerade modeller som SVR och Random Forest kan forskare och praktiker få värdefulla insikter i marknadsdynamiken och förbättra precisionen i värderingsprocesser. Detta arbete lägger grunden för framtida forskning som kan utvidga tillämpningen av maskininlärning inom fastighetssektorn och andra relaterade områden.

Appendix A

Kodsnuttar för implementering av maskininlärningsmodeller

1. Linjär Regression:

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

2. Random Forest:

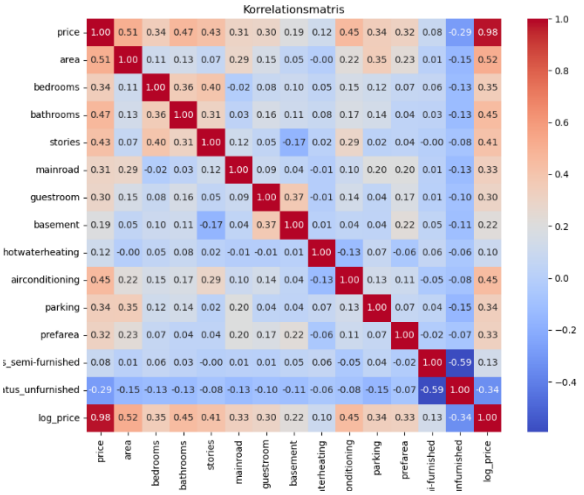
```
from sklearn.ensemble import RandomForestRegressor
model = RandomForestRegressor(n_estimators=100, max_depth=10)
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

3. Suport Vector Regression

```
from sklearn.svm import SVR
model = SVR(kernel='rbf', C=1, gamma=0.01)
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

Appendix B

Fullständig korrelationsmatris



Appendix C

Beskrivning av dataset

De första raderna av datasetet:

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	13300000	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
1	12250000	8960	4	4	4	yes	no	no	no	yes	3	no	furnished
2	12250000	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
3	12215000	7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished
4	11410000	7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished

Källförteckning

Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd Edition). O'Reilly Media. Hämtad 6 januari 2025 från https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2nd Edition). Springer. Hämtad 6 januari 2025 från https://www.statlearning.com

Zhang, D., Zheng, C., & Wei, Y. (2021). The application of machine learning in real estate price prediction: A review. *Journal of Urban Technology*, 28(2), 85-104. Hämtad 6 januari 2025 från https://www.researchgate.net/publication/367317216_Machine_Learning_for_Housing_Price_Prediction

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. Hämtad 9 januari 2025 från https://link.springer.com/article/10.1023/A:1010933404324